A concise introduction to mathematical statistics

Tonatiuh Sánchez-Vizuet
The University of Arizona

Last update: June 1, 2025

Contents

Preface

There is a wealth of textbooks covering the subject of mathematical statistics. Most of them are encyclopedic efforts that could be used to teach three or more semester long courses covering many topics ranging from theory, to computation to application. Due to their encyclopedic nature, these books, while being excellent reference sources, can be quite daunting and intimidating for the student first approaching the subject of mathematical statistics. The present set of notes were prepared for their use in the class *MATH 466 - Theory of Statistics* at the University of Arizona during the Fall 2022 semester. They are meant to be a short, self contained introduction to the subject that satisfies two conditions:

1. they contain enough material to build a *basic-but-solid* theoretical foundation, and
2. they should remain manageable in size and contents so that any student can be reasonably expected to "read them from cover to cover" and work through every problem during the span of one semester.

Due to these constraints the notes are by necessity incomplete; cover a very short ground and leave out many important subjects. However, the student who takes the time to go over the entirety of this selection, working out carefully *all* the problems provided will have built a solid understanding that will allow them to successfully undertake the study of the many more advanced (and complete) texts available.
True to their intent on being used in the classroom, the notes are divided into lectures, rather than into sections or chapters. Each lecture corresponds to the contents of a fifty minute long class and is followed by a short selection of problems that are *essential* for the full assimilation of the concepts in the lecture. It is stressed that students are expected to solve every problem provided for a full grasp of the subject.
It is important to remark that *MATH 466 - Theory of statistics*, the class for which the notes were prepared, **is not a first course in statistics**, but rather a first course in *mathematical statistics*. Students enrolled in the class are expected to have completed introductory courses in both statistics and probability, as well as courses on single and multi variable calculus and linear algebra at a level equivalent to those offered for freshmen and sophomore students in U.S. colleges, and at least one class involving abstract mathematical arguments and proofs. For instance introductory real analysis (at the level of [6, 7]), introduction to proofs, introduction to abstract algebra, etc. Importantly, the student should be already familiar and comfortable with the mechanics and mathematical manipulations involved in computing probabilities, finding confidence intervals, performing hypothesis testing and linear regression, etc. Moreover, the students should have taken, either previously or concurrently, an introductory course on the theory of probability. On the other hand, they are not expected to have been exposed to the theory and justifications behind statistical methods or to rigorous measure theory (which will be kept to the very minimum).
No claim of originality or novelty is made. The content of these lectures is based on the textbook *Introduction to statistical theory* by Paul G. Hoel, Sidney C Port and Charles J. Stone [3] and has been complemented with material from other references, most notably *Introduction to probability theory* [4] by the same authors for the initial review on probability, Wasserman's *All of statistics* [8], and Hodges' and Lehmann's *Basic concepts in probability and statistics* [2]. Some other references that have been used while preparing these notes can be found in the reference list at the end of the document.

Lecture 1

Review of probability theory I

Contents

1.1 Probability spaces

Consider an experiment whose result is determined by chance. We will refer to any particular result of the experiment as an ***outcome*** and will denote it as ω. An ***event*** A is a set constituted by one or more outcomes. A probability model or ***probability space*** for this experiment consists of three essential pieces: a sample space, a sigma algebra of events and a probability function. We define each of the in what follows

- The ***sample space***, that will be denoted as Ω, is the set of all possible outcomes of the the experiment. In mathematical notation, if ω denotes an outcome, A is an event and Ω is the probability space, then

$$A := \{\omega : \omega \text{ satisfies a given condition}\},$$
$$\Omega := \{\omega : \omega \text{ is an outcome}\}.$$

- A $\boldsymbol{\sigma}-$**algebra** (***sigma-algebra***) of events. We will not get into the theoretical details of the following, but in order to account for all possible outcomes and combinations of events, a probability model must consider a broad collection of events, denoted as $\mathcal{F}$, satisfying the following properties:

  1. $\Omega \in \mathcal{F}$.
  2. If the event $A \in \mathcal{F}$ then the event $A^c \in \mathcal{F}$.
  3. If every event from the *countable* family $A_1, A_2, A_3, \ldots$ belongs to Ω. Then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.
  4. If every event from the *countable* family $A_1, A_2, A_3, \ldots$ belongs to Ω. Then $\cap_{n=1}^{\infty} A_n \in \mathcal{F}$.

  A set $\mathcal{F}$ satisfying all the properties above is known in mathematical analysis as a $\sigma-$algebra (read as *"sigma algebra"*).

- The ***probability***, that will be denoted as $P(\cdot)$, is a measurable function that assigns a real number to every event A. This kind of function is sometimes called a *set function* because it can take as argument either a single outcome ω, or a set of them, i.e. and event, which is a subset of the sample space. If we

denote respectively by $|\Omega|$ and $|A|$ the *size*[1] of the sample space and of the event A, we can then write all this rigorously as

$$\begin{aligned} P : \Omega &\longrightarrow \mathbb{R} \\ A &\longmapsto \frac{|A|}{|\Omega|}. \end{aligned}$$

1.2 Properties of the probability function

From this definition, where we have implicitly assumed that $|\Omega| < \infty$, we can see that the probability satisfies the following properties

1. $0 \leq P(A) \leq 1$.

   We will use the *measure-theoretic* facts that $A \subseteq B$ implies that $|A| \leq |B|$ (this property is called monotonicity) and that $|\varnothing| = 0$. Hence, since

   $$\varnothing \subseteq A \subseteq \Omega,$$

   we have that

   $$0 = |\varnothing|/|\Omega| \leq |A|/|\Omega| \leq |\Omega|/|\Omega| = 1,$$

   and applying the definition of probability it follows that $0 \leq P(A) \leq 1$ as desired.

2. If the events A and B are such that $A \cap B = \varnothing$ (i.e. they are ***disjoint***) then

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

   We will use again a result from measure theory that states that if two sets are disjoint, then the measure of their union is the sum of their individual measures (this follows from a property of measures known as *subadditivity*). Using this property, the result follows from the fact that we can decompose $A \cup B$ into

   $$A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B)$$

   Since all of the sets on the right are disjoint it then follows that the size of the left hand side is the sum of each of the sizes on the right. Therefore, if $A \cap B = \varnothing$ it follows that

   $$P(A \cup B) = \frac{|A \cup B|}{|\Omega|} = \frac{|A \cap B^c| + |B \cap A^c| + |A \cap B|}{|\Omega|} = \frac{|A| + |B| + 0}{|\Omega|} = P(A) + P(B).$$

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

Using the three properties above, it is then possible to prove (this will be your first exercise) the following further properties

4. If all the events $A_1, A_2, \ldots, A_{n-1}, A_n$ are ***mutually disjoint*** or ***mutually exclusive*** (i.e. $A_i \cap A_j = \varnothing$ for every $i \neq j$), then

   $$P(A_1 \cup A_2 \cup \ldots A_{n-1} \cup A_n) = P(A_1) + P(A_2) + \ldots P(A_{n-1}) + P(A_n).$$

[1] The precise meaning of "size" is the subject of a course on measure theory, however for the purpose of this course we can use an intuitive meaning. If the sample space is discrete and finite, then $|A|$ is simply the *cardinality* of the set A, i.e. the number of outcomes that belong to the event A. If the sample space is continuous then $|A|$ is the *measure* of the set A. You can think of the measure of A as its length in one dimension, its surface in two dimensions, its volume in three, etc.

5. If $A \subseteq B$ then $P(A) \leq P(B)$.
6. ***Inclusion-exclusion***. For any two events A and B

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

7. $P(A^c) = 1 - P(A)$.
8. If $B \subseteq A$, then $P(A \cap B^c) = P(A) - P(B)$.

We will require that property 4 holds for an infinite but countable union of pairwise disjoints sets (this will enable us to use ideas from calculus).

9. If $A_i \cap A_j = \varnothing$ for every $i \neq j$, then

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

Any function that takes events and arguments and satisfies properties 1,2 and 9 will be called a ***probability***.

Definition 1.1. Consider a countable family of nested sets $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$ We will say that the family *increases* to some set A if it holds that $\lim_{n\to\infty} \cup_{i=1}^{n} A_i = A$.

Definition 1.2. Consider a countable family of nested sets $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$ We will say that the family *decreases* to some set A if it holds that $\lim_{n\to\infty} \cap_{i=1}^{n} A_i = A$.

Theorem 1.1. Continuity of probability. *If a nested family of events $\{A_n\}$ increases or decreases to some event A, then*

$$P(A) = \lim_{n\to\infty} P(A_n).$$

Proof. Consider that the sequence $\{A_n\}$ increases to A. We define $B_n := A_n \cap (A_{n-1})^c$ and note that (1) B_n and B_m are disjoint for every n and m, that (2) $A_n = \cup_{i=1}^{n} B_i$, and that (3) $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n = A$. Then, since the family $\{B_n\}$ is disjoint, we obtain

$$P(A) = P(\cup_{i=1}^{\infty} A_i) = P(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \lim_{n\to\infty} P(\cup_{i=1}^{n} B_i) = \lim_{n\to\infty} P(A_n).$$

We now consider the case where the sequence $\{A_n\}$ decreases to A. The key is to observe that if $\{A_n\}$ decreases to A then $\{(A_n)^c\}$ increases to A^c; thus

$$P(A) = 1 - P(A^c) = 1 - \lim_{n\to\infty} P((A_n)^c) = \lim_{n\to\infty} (1 - P((A_n)^c)) = \lim_{n\to\infty} P(A_n).$$

□

1.3 Exercises

1. Prove that a probability function satisfies properties 4, 5, 6, 7, and 8 from section 1.2.

Lecture 2

Review of probability theory II

Contents

2.1 Independence and conditional probability

Definition 2.1. The ***conditional probability*** of B given A is defined as

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

Note that in general $P(B|A) \neq P(A|B)$. The definition above is often used in the form

$$P(A \cap B) = P(A) \cdot P(B|A),$$

which is commonly referred to as the ***product rule***.
The probability of an event B is roughly speaking the fraction of the sample space that is occupied by B. When we consider the conditional probability of B given A we are no longer considering the entire sample space and instead we restrict ourselves to the smaller subset A. Hence, the conditional probability of B given A refers to the fraction of the event A that is occupied by the outcomes also contained in the event B.

Definition 2.2. The events A and B will be called ***independent events*** if

$$P(A \cap B) = P(A) \cdot P(B).$$

It follows then that if A and B are independent

$$P(B|A) = P(B).$$

From the definition of conditional probability it is possible to derive ***Baye's formula***

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)}. \tag{2.1}$$

2.2 Random variables and distribution functions

Definition 2.3. Consider a probability space Ω. A ***random variable*** X is a ***measurable function***[1] whose domain is the the probability space. The range of X is known as the ***state space***. In other words, a random variable is a function that takes elements of Ω as arguments and such that the pre-image of any measurable subset of the state space is itself measurable.

Definition 2.4. Consider a random variable $X : \Omega \to \mathbb{R}$ taking values over the real numbers. Then, the function $F_X(x)$ associating a real number x to the probability that the value of the random variable X is equal to or less than x

$$F_X(x) := P(X \leq x)$$

is known as the ***cumulative distribution function***. It is common to use the acronym **CDF** as shorthand for "cumulative distribution function". The cumulative distribution function F_X captures and encodes the behavior of the random variable X. In a sense, the CDF of X contains all the information that you need in order to do mathematics involving X.

Remark 2.1. In probability and statistics, it is customary to use capital letters X, Y, Z to denote random variables while lower case letters x, y, and z are reserved for real numbers. Hence, it is common to find expressions like $X \leq x$ (the value of the random variable X is less than or equal to the real number x) which can be confusing at the beginning—specially on the blackboard. Eventually you will get used to the notation and immediately realize from the capitalization when a text is talking about a random variable or about a real number. In a similar fashion, capital letters like F_X, G_X, and H_X are used to denote cumulative distributions, while the lower case letters f_X, g_X, and h_X are typically reserved to the associated probability density functions (we will discuss this concept a little later). In this context, the subscript "X" is meant to stress the fact that these are no ordinary functions, but are associated to a random variable instead (note the capitalization of the subscript).

Theorem 2.1. *Any cumulative distribution function F_X has the following properties*

1. *F_X is a* **non-decreasing function**, *meaning that if $x \leq y$ then $F_X(x) \leq F_X(y)$.*
2. *$\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$.*
3. *F_X is* **right continuous**, *by which we mean that*

$$\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$$

4. *If we define*

$$F_X(x-) := \lim_{p\to x^-} F_X(p),$$

then

$$F_X(x-) = P(X < x).$$

5. *$P(X = x) = F_X(x) - F_X(x-)$.*
6. *$P(a \leq X \leq b) = F_X(b) - F_X(a)$,*

Proof. We will prove each of these properties

[1] This concept again falls within the realm of measure and integration theory and we shall not delve too much on it here. Without entering into technical details, a measurable function is a function such that its pre-image of a measurable set is itself measurable.

1. Let $x \leq y$, then it follows that the sets $\{\omega \in \Omega : X(\omega) \leq x\} \subseteq \{\omega \in \Omega : X(\omega) \leq y\}$ and thus, using property 5 from the previous lecture it follows that
$$F_X(x) = P(X \leq x) \leq P(X \leq y) = F_X(y).$$

2. We define now the family of sets $A_n \subseteq \Omega$ as
$$A_n := \{\omega \in \Omega : X(\omega) \leq n\}.$$
We see that with this definition the following facts hold
   (a) $\ldots \subset A_{-2} \subset A_{-1} \subset A_0 \subset A_1 \subset A_2 \subset \ldots$
   (b) $\cap_{n=0}^{\infty} A_{-n} = \varnothing$ (The family $\{A_{-n}\}$ decreases to $\varnothing$).
   (c) $\cup_{n=0}^{\infty} A_n = \Omega$ (The family $\{A_n\}$ increases to Ω).

   Hence, using the continuity of probability we have
$$\lim_{x\to-\infty} F_X(x) = \lim_{x\to-\infty} P(X \leq x) = \lim_{n\to\infty} P(A_{-n}) = P(\varnothing) = 0.$$
   Analogously
$$\lim_{x\to\infty} F_X(x) = \lim_{x\to-\infty} P(X \leq x) = \lim_{n\to\infty} P(A_n) = P(\Omega) = 1.$$

3. Let $x_0 \in \mathbb{R}$ and define that the sets
$$A_n := \{\omega \in \Omega : X(\omega) \leq x_0 + \tfrac{1}{n}\}.$$
Clearly $A_1 \supseteq A_2 \supseteq A_2 \supseteq \ldots$ and also $\cap_{n=1}^{\infty} A_n = \{\omega \in \Omega : X(\omega) \leq x_0\}$. Hence
$$\lim_{x\to x_0^+} F_X(x) = \lim_{x\to x_0^+} P(X \leq x) = \lim_{n\to\infty} P(A_n) = P(X \leq x_0) = F_X(x_0).$$

4. We will prove the statement using analogous arguments to those used in the two previous points. Define
$$A_n := \{\omega \in \Omega : X(\omega) \leq x - \tfrac{1}{n}\},$$
and note that $\cup_{n=1}^{\infty} A_n = \{\omega \in \Omega : X(\omega) < x\}$ as well as $A_1 \subset A_2 \subset A_3 \subset \ldots$ hence
$$F_X(x-) := \lim_{p\to x^-} F_X(p) = \lim_{n\to\infty} F_X(x - \tfrac{1}{n}) = \lim_{n\to\infty} P(A_n) = P(X < x).$$

5. To prove this we note that
$$\{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : X(\omega) < x\} \cup \{\omega \in \Omega : X(\omega) = x\},$$
where the sets on the right hands side are disjoint. Therefore, using property 2 of the previous lecture we have that
$$F_X(x) = P(X \leq x) = P(X < x) + P(X = x) = F_X(x-) + P(X = x),$$
where we have used point 4. Statement 5 follows readily.

6. We first note that
$$\{\omega \in \Omega : a \leq X(\omega) \leq b\} = \{\omega \in \Omega : X(\omega) \leq a\}^c \cap \{\omega \in \Omega : X(\omega) \leq b\},$$
and that
$$\{\omega \in \Omega : X(\omega) \leq a\} \subseteq \{\omega \in \Omega : X(\omega) \leq b\}.$$
Therefore, we can use property 8 from Section 1.2 to conclude that
$$\begin{aligned} P(\{\omega \in \Omega : a \leq X(\omega) \leq b\}) &= P(\{\omega \in \Omega : X(\omega) \leq b\}) - P(\{\omega \in \Omega : X(\omega) \leq a\}) \\ &= F_X(b) - F_X(a). \end{aligned}$$

□

2.3 Exercises

1. Assume that the sample space Ω is divided into n mutually disjoint sets $\{E_i\}_{i=1}^n$ such that

$$\Omega = \cup_{i=1}^n E_i.$$

   Such a family of sets is known as a ***partition of the sample space***. Prove that for any arbitrary event $A \subset \Omega$ we can write

$$A = \cup_{i=1}^n (A \cap E_i).$$

2. Consider a partition of the sample space $\{E_i\}_{i=1}^n$. Show that for an arbitrary event $A \subset \Omega$ it holds that

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i).$$

   This relation is known as the ***law of total probability***.

3. Let E and F be mutually exclusive. Prove that

$$P(F|E^c) = \frac{P(F)}{1 - P(E)}.$$

   Is this fact also true if E and F are not exclusive? Prove or provide a counterexample.

4. If $P(A) = .4$, $P(B) = .5$, and $P(A \cup B) = .7$, are events A and B independent?

5. Use the product rule for conditional probability to derive Baye's formula (2.1).

6. Show that, for any events E and F, it holds that

$$P(E|F) + P(E^c|F) = 1.$$

7. Show that if E and F are independent , then E and F^c are also independent. [Hint: use the fact that $P(E) = P(E \cap F) + P(E \cap F^c)$]

8. Show that if E and F are independent, then E^c and F^c are independent too. [Hint: use the previous problem twice]

9. Show that if E and F are mutually exclusive, then they can not be independent unless either $P(E) = 0$ or $P(F) = 0$.

Lecture 3

Review of probability theory III

Contents

3.1 Discrete and continuous random variables

Definition 3.1. Consider a probability space Ω and a random variable $X : \Omega \to \mathbb{R}$. We will say that X is a

- ***Discrete random variable*** if there exists countable set D such that $P(X \in D) = 1$. In other words, X is a discrete random variable if it takes at most countably many values.
- ***Continuous random variable*** if its cumulative distribution function F_X is continuous for every x.

Remark 3.1. There are other alternate definitions for a continuous random variable. A common alternative is to say that X is a continuous random variable if $P(X = x) = 0$ for any $\infty < x < \infty$. Intuitively you can interpret this statement as the fact that if you were to randomly chose *any* real number between say, 0 and 1, due to the fact that there are uncountable many choices, the probability of choosing any particular number is in fact zero.

For discrete random variables we define the ***probability mass function*** as

$$p_X(x) := P(X = x).$$

A probability mass function must be such that

$$p_X(x) \geq 0 \quad \text{for all } x \in \mathbb{R} \tag{3.1}$$

$$\sum_{s \in D} p_X(s) = 1. \tag{3.2}$$

It should be easy to see that, for a discrete random variable, the cumulative distribution function and the probability mass function are related by

$$F_X(x) = P(X = x) = \sum_{D \ni s \leq x} p_X(s) \tag{3.3}$$

which is piecewise constant with jumps at every $x \in D$. Discrete random variables are completely determined by specifying their mass function.
For simplicity of exposition we will make the additional assumption that the distribution F_X of a continuous random variable X has a derivative[1] that we will call the ***probability density function*** and will denote as

$$f_X(x) := \tfrac{d}{dx} F_X(x).$$

Whenever possible, we will use the same letter as the distribution, but in lower case. It is customary to use the acronym **PDF** to refer to the probability density function and to use the notation

$$X \sim f_X$$

(that should be read "X is distributed as f_X") to say that the behavior of X is described by f_X.
Note that the definition along with the fundamental theorem of calculus and property 2 from Theorem 2.1 imply that

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds. \tag{3.4}$$

$$\int_{-\infty}^{\infty} f_X(s)ds = 1. \tag{3.5}$$

In general, any non negative function $f(x) \geq 0$ satisfying the property (3.5) defines a probability density function with corresponding cumulative distribution given by (3.4). Equation (3.4), which is sometimes used to define the cumulative distribution function, has the following consequence: for an event $A \subset \mathbb{R}$ and a random variable X it holds that

$$P(X \in A) = \int_A f_X(s)ds. \tag{3.6}$$

This relation, which holds as well for multidimensional random variables, is commonly referred to the ***law*** of the random variable X, or the ***distribution*** of X. We say that two random variables are ***equal in law*** or ***equal in distribution*** and write

$$X \overset{d}{=} Y$$

if $F_X(x) = F_Y(x)$ for every x. Note that this *does not mean* that $X = Y$ but rather than all statements involving the probabilities of X and Y are equivalent.

3.2 Transforming random variables

Often times new random variables arise from operating over existing random variables. The CDF and PDF functions of the new variables arising from these manipulations are related to those from the original one. If X is the original random variable, f_X is its associated PDF, and we define a new random variable as $Y = r(X)$, the new PDF f_Y can be determined in general as follows.

1. For each y, find the set $A_y := \{x : r(x) \leq y\}$.
2. Find the CDF for the transformed variable

$$F_Y(y) = P(Y \leq y) = P(r(X) \leq y) = P(x \in A_y) = \int_{A_y} f_X(x)\, dx.$$

[1] In this context we shall understand the term "differentiable" in the sense that the derivative can be computed from the right and from the left of any given point x, but that these two left and right derivatives might be different. We use this definition for convenience, but strictly speaking we would have to generalize the concept of derivative for functions with corners and discontinuities, which would bring us to the world of distributional derivatives.

3. The PDF is then given by $f_Y(y) = F'_Y(y)$.

Example. Consider a random variable X with density function

$$f_X(x) = \begin{cases} 0 & x \in (-\infty, 0) \\ e^{-x} & x \in [0, \infty) \end{cases}$$

and define $Y := \log(X)$. Hence, to find $f_Y(y)$ we proceed as follows. We first identify the set

$$A_y := \{r(X) \leq y\} = \{\log X \leq y\} = \{X \leq e^y\}.$$

Hence

$$F_Y(y) = P(X \leq e^y) = F_X(e^y).$$

We then find

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds = \int_0^x e^{-s}ds = 1 - e^{-x}.$$

Therefore

$$F_Y(y) = F_X(e^y) = 1 - e^{-e^y}.$$

Finally, we determine the PDF by differentiation

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(1 - e^{-e^y}\right) = e^y e^{-e^y}.$$

Speaking loosely, the CDF of a random variable takes as input a real number and provides the probability that the random variable takes on values smaller than the input. Thinking about statistical sampling, the CDF takes certain value of a quantity of interest as input and gives back the proportion of the population that falls below such value. However, sometimes we want to provide a target fraction of the population and determine what value of the quantity of interest is such that the provided fraction of the population falls below this point. In terms of probability, we would like to provide a certain value of the probability p, and be able to determine the real number x such that the probability of $P(X \leq x) = F_X(x) = p$. In other words, we would like to *invert* the distribution function. This is accomplished by the quartile function.

Definition 3.2. Let X be a continuous random variable with CDF F_X. The ***inverse cumulative distribution function*** or ***quantile function*** is defined by

$$F_X^{-1}(q) := \inf\{x : F_X(x) > q\} \qquad \text{for } q \in [0, 1].$$

If F_X is strictly increasing, then $F_X^{-1}(q) = x$ is the unique real number x for which $F(x) = q$.

We will refer to $F_X^{-1}(1/4)$ as the *first quartile*, $F_X^{-1}(1/2)$ as the *median* or *second quartile*, and $F_X^{-1}(3/4)$ as the *third quartile*.

3.3 Exercises

1. Verify that, for all the discrete random variables in Section A.1, it holds that

$$\sum_{x \in D} p(x) = 1.$$

2. Prove that the definition of a continuous random variable given in Definition 2.4 is equivalent to the ones given in Remark 3.1. Namely, prove that for any function F_X satisfying the properties in Theorem 2.1, the following statements are equivalent

(a) $F_X(x)$ is continuous.

(b) $P(X = x) = 0$ for all x.

3. We say that X is a ***symmetric random variable*** if X and $-X$ have the same distribution function. Prove that a probability density function f_X is symmetric (i.e. $f_X(x) = f_X(-x)$) if and only if X is symmetric.

4. Let X be a continuous random variable with density function f_X and let $g : \mathbb{R} \to \mathbb{R}$ be a strictly monotonic differentiable function. Prove that the random variable defined by $Y := g(X)$ has density

$$f_Y(y) = f_X(g^{-1}(y)) \left| \tfrac{d}{dy} g^{-1}(y) \right|.$$

5. Use the fundamental theorem of calculus and property 2 from Theorem 2.1 to prove the equalities (3.4) and (3.5).

6. Let $X \sim Exp(\theta)$. Find the CDF F_X and the quantile function $F_X^{-1}(q)$.

7. Let X have CDF F_X. Find the CDF of the random variables

$$(a)\ X^+ = \max\{0, X\} \quad \text{and} \quad (b)\ X^- = \min\{0, X\}.$$

[Hint: Both answers will be piecewise defined functions.]

8. Let X be a continuous random variable with density function f_X. Let A be a subset of the real line and the indicator function of A as

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Find the cumulative distribution function of the random variable $Y := I_A(X)$. [Hint: First realize that Y is actually a *discrete* random variable and start by finding its probability mass function].

9. Consider a random variable X having a continuous, strictly increasing CDF, F_X. Define $Y = F_X(x)$, the PDF of Y is known as the ***probability transform***.

(a) Find the PDF for Y.

(b) Consider a uniformly distributed variable $U \sim U(0, 1)$ and let $X = F_X^{-1}(U)$. Show that $X \sim F$.

This process can be used to generate random samples with distribution F_X using random samples generated uniformly over the interval $(0, 1)$.

10. Following the process described in the previous problem, find the probability transform and write an R code that generates random variables distributed as the exponential distribution $Exp(\theta)$.

Lecture 4

Review of probability theory IV

Contents

4.1 Bivariate random variables and joint distributions

For simplicity, in what follows we will focus on functions depending on two random variables, but all the results and concepts that will be discussed can be easily extended to functions depending on more than two. A function $f(x, y) : \mathbb{R}^2 \to \mathbb{R}$ is called a ***joint probability density function*** (PDF) for the continuous random variables X and Y if

1. $f_{X,Y}(x, y) \geq 0$ for all (x, y).
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1$.
3. For any set $A \subset \mathbb{R}^2$ it holds that $P((X, Y) \in A) = \int \int_A f_{X,Y}(x, y)\, dx\, dy$.

The joint CDF function is related to the joint PDF by

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y) = \begin{cases} \displaystyle\sum_{D_y \ni n \leq y} \sum_{D_x \ni m \leq x} p_{X,Y}(m, n) & \text{(for discrete RVs)} \\ \displaystyle\int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv & \text{(for continuous RVs).} \end{cases} \tag{4.1}$$

where the sets D_x and D_y denote the (domain) of the discrete RVs x and y respectively, and the notation $P(X \leq x, Y \leq y)$ is shorthand for $P(X \leq x \text{ and } Y \leq y)$.

4.2 Marginal distributions and densities

The joint CDF $F_{X,Y}$ for two random variables X and Y encodes the *simultaneous* probabilistic behavior of the random variables. The one dimensional probability of a single RV (insulated from the behavior of the other one) gives rise to the ***marginal distribution function*** which, for continuous RVs, is given by

$$F_X(x) := P(X \leq x) = \lim_{y\to\infty} F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx,$$
$$F_Y(y) := P(Y \leq y) = \lim_{x\to\infty} F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy.$$

The corresponding ***marginal probability density function*** is given by

$$f_X(x) := \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy,$$
$$f_Y(y) := \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

The corresponding marginal CDF for discrete RVs are given by

$$F_X(x) := P(X = x) = \sum_{D_x \ni m \leq x} \sum_{n\in D_y} p_{X,Y}(m,n)$$
$$F_Y(y) := P(Y = y) = \sum_{D_y \ni n \leq y} \sum_{m\in D_x} p_{X,Y}(m,n),$$

while the PDFs are

$$f_X(x) := \sum_{n\in D_y} p_{X,Y}(m,n) \qquad \text{and} \qquad f_Y(y) := \sum_{m\in D_x} p_{X,Y}(m,n).$$

4.3 Conditional densities

If X and Y are discrete, we can then compute the conditional distribution of X given that we have observed certain value of $Y = y$. As we have seen before

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

This motivates the definition of the ***conditional probability mass function***

$$p_{X|Y}(x|y) := \frac{p_{X,Y}(x,y)}{p_Y(y)} \qquad (\text{if } p_Y(y) > 0). \tag{4.2}$$

Note that if X and Y are continuous, then strictly speaking the meaning of the expression $P(X \in A|Y = y)$ has to be handled with care, since the event $\{Y = y\}$ has probability 0. This wrinkle can however be sorted out (we will not delve into the details) and the definition above can be generalized for continuous random variables as follows:
For continuous random variables X and Y, the ***conditional probability density function*** is defined as

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)} \qquad (\text{if } f_Y(y) > 0), \tag{4.3}$$

and

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\,dx.$$

The definition of the conditional density and mass function lead to the following useful formula (familiar from Baye's formula (2.1) in Lecture 2)

$$f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x).$$

4.4 Independent random variables

We say that X and Y are ***independent random variables*** if for every pair of events A and B it holds that

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B),$$

otherwise we say that X and Y are ***dependent random variables***.
The joint PDF for independent RVs can be expressed in terms of their marginal densities.

Theorem 4.1. *If X and Y have joint PDF given by $f_{X,Y}$, then they are independent if and only if up to a set of measure zero it holds that*

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

Note that, as we shall soon see, since the PDF of a random variable is only ever used under integral sign, the fact that the equality above fails to hold for a set of measure zero is not an issue.
A slightly different result that is often useful is the following.

Theorem 4.2. *Consider that X and Y are defined over a possibly infinite rectangle. If there exist two functions $g(x)$ and $h(y)$* ***not necessarily one-dimensional PDFs themselves*** *such that*

$$f_{X,Y}(x, y) = g(x)h(y)$$

then X and Y are independent.

Note that in the preceding theorem g and h need not be the marginal densities, nor should they integrate to 1 over the real line.

4.5 Sums, products and quotients of continuous random variables

It is also common and useful to combine more than one random variable to give rise to new ones. In these cases, we would like to be able to obtain the CDF and PDF of the new variable from those of the original ones. Consider that the random variables X and Y are known, and denote the new random variable $Z := Z(X, Y)$. The general technique for finding F_Z and f_Z is as follows. We first find the set

$$A_Z : \{(x, y) : Z(X, Y) \leq z\}.$$

Then, from the definition of F_Z and Property 3 in 4.1, it follows that

$$F_Z(z) = P(Z \in A_Z) = \int\int_{A_Z} f_{X,Y}(x, y)\,dx\,dy.$$

If we can manipulate the integral above into an expression of the form

$$F_Z(z) = \int_{-\infty}^{z} h(s)ds,$$

then, by the property (3.4) it will necessarily follow that $f_Z(z) = h(z)$.
The following theorem gives the PDF of three common and useful combinations of RVs, namely the sum, product and quotient or two known random variables. We will illustrate the process described above by deriving the formula for one of these combinations and will leave the remaining two as practice problems.

Theorem 4.3. *Consider that the continuous random variables X and Y are known and have joint PDF given by $f_{X,Y}(x,y)$. Then*

1. *The PDF of the random variable $Z := X + Y$ is given by*

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)dx.$$

2. *The PDF of the random variable $Z := XY$ is given by*

$$f_Z(z) := \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}(x, z/x)dx.$$

3. *The PDF of the random variable $Z := Y/X$ is given by*

$$f_Z(z) := \int_{-\infty}^{\infty} |x| f_{X,Y}(x, zx)dx.$$

Proof. We will prove point 2 and will leave the other two as practice problems. We start by identifying the set

$$A_Z := \{(x,y) : Z(X,Y) \leq z\} = \{(x,y) : xy \leq z\}.$$

We notice that if $x < 0$ then $xy < z \Leftrightarrow y > z/x$, while if $x \geq 0$ then $xy \leq z \Leftrightarrow y \leq z/x$ [1]. Therefore

$$A_Z = \{(x,y) : x < 0 \cap y > z/x\} \cup \{(x,y) : x \geq 0 \cap y \leq z/x\},$$

hence we have that

$$\begin{aligned} F_Z(z) &:= \int\int_{A_Z} f_{X,Y}(x,y)\, dx\, dy \\ &= \int_{-\infty}^{0}\int_{z/x}^{\infty} f_{X,Y}(x,y)\, dy\, dx + \int_{0}^{\infty}\int_{-\infty}^{z/x} f_{X,Y}(x,y)\, dy\, dx \end{aligned}$$

Recalling that we would like to have "z only" in the integration limits, we perform the change of variables $s = xy$ in both inner integrals. With this change of variables the integrals become

$$\begin{aligned} &= \int_{-\infty}^{0}\int_{z}^{-\infty} \frac{1}{x} f_{X,Y}(x, s/x)\, ds\, dx + \int_{0}^{\infty}\int_{-\infty}^{z} \frac{1}{x} f_{X,Y}(x, s/x)\, dy\, dx \\ &= \int_{-\infty}^{0}\int_{-\infty}^{z} \left(-\frac{1}{x}\right) f_{X,Y}(x, s/x)\, ds\, dx + \int_{0}^{\infty}\int_{-\infty}^{z} \frac{1}{x} f_{X,Y}(x, s/x)\, dy\, dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z} \frac{1}{|x|} f_{X,Y}(x, s/x)\, ds\, dx \end{aligned}$$

changing the order of the integrals above yields

$$= \int_{-\infty}^{z}\int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}(x, s/x)\, dx\, ds = \int_{-\infty}^{z} f_Z(s) ds,$$

[1] The limiting case $x = 0$ is covered by $y < \lim_{x\to\infty} z/x = \infty$.

where

$$f_Z(z) := \int_{-\infty}^{\infty} \frac{1}{|x|} f_{X,Y}(x, z/x)\, dx.$$

□

Corollary 4.4. *If X and Y are **independent** continuous random variables with PDFs f_X and f_Y, then*

1. *The PDF of the random variable $Z := X + Y$ is given by*

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$

2. *The PDF of the random variable $Z := XY$ is given by*

$$f_Z(z) := \int_{-\infty}^{\infty} \frac{1}{|x|} f_X(x) f_Y(z/x) dx.$$

3. *The PDF of the random variable $Z := Y/X$ is given by*

$$f_Z(z) := \int_{-\infty}^{\infty} |x| f_X(x) f_Y(zx) dx.$$

Remark 4.1. Given real-valued functions g and h, the ***convolution*** of g and h is given by

$$(g * h)(x) = \int_{-\infty}^{\infty} g(s) h(x - s) ds.$$

The result above then tells us that the mass function of the sum of two continuous independent random variables is given by the convolution of the two mass functions.

4.6 Multidimensional random variables and joint distributions

All the concepts introduced in the previous section for two random variables can be extended naturally to multiple variables.
We can define an $n-$dimensional ***vector-valued random variable $\boldsymbol{X}$*** as the vector of component-wise random variables

$$\boldsymbol{X} := (X_1, X_2, \ldots, X_{n-1}, X_n),$$

where each of the entries X_i is a scalar-valued random variable. For such a random vector, we can define the ***joint distribution function*** as

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) := P\left((X_1 \leq x_1) \cap \ldots \cap (X_n \leq x_n)\right).$$

This encodes the probability that *all* the inequalities are satisfied *simultaneously*. The notation on the right hand side above is often shortened as

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) := P(X_1 \leq x_1, \ldots, X_n \leq x_n).$$

This joint distribution function has properties analogous to the ones summarized on Theorem 2.1 for scalar random variables.
Multivariate random variables can also be distinguished between discrete and continuous and they also have associated joint probability mass functions

$$p_{\boldsymbol{X}}(x_1, \ldots, x_n) = P(X_1 = x_1 \ldots, X_n = x_n)$$

for discrete variables.
We will say that the random variables $X_1, X_2, \ldots, X_n$ are ***independent identically distributed*** (which will be shortened as ***IID***) if they are all pairwise independent and they all share the same PDF $f_X(x)$. In that case, the joint pdf for the vector valued RV $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ can be expressed as the product of the individual PDFs

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

4.7 Exercises

1. **Addition of continuous random variables.** Consider that X and Y are continuous random variables with joint density function $f_{X,Y}(x, y)$.

   (a) Prove that the density function of the random variable defined as $Z := X + Y$ is given by

   $$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x)dx.$$

   (b) How does the final formula change if X and Y are independent? Justify your answer.

2. Consider two independent random variables X and Y both having an exponential distribution with parameter λ (See section A.2 for the PDF). Using the result proven in the previous problem, find the PDF for the random variable $Z := X + Y$.

3. **Addition of independent discrete random variables.** Consider that X and Y are independent integer-valued random variables with mass functions p_X and p_Y. Prove that the random variable $Z := X + Y$ has the mass function

   $$p_Z(x) = \sum_{n \in \mathbb{Z}} p_X(n)p_Y(x - n).$$

   Remark. Given two integer-valued functions f and g the ***discrete convolution*** of f and g is given by

   $$(f * g)(x) := \sum_{n \in \mathbb{Z}} f(n)g(x - n).$$

   The result above tells you that the mass function of the sum of two discrete independent random variables is given by the discrete convolution of the two mass functions.

4. Let X and Y be continuous random variables with joint density $f_{X,Y}$.

   (a) Find the **joint** CDF and PDF of the random variables

   $$W := a + bX. \qquad \text{and} \qquad Z := c + dY,$$

   where $b > 0$ and $d > 0$.

   (b) Show that if X and Y are independent, then W and Z are independent too.

5. Let X and Y be continuous random variables with joint distribution $F_{X,Y}$ and density $f_{X,Y}$.

   (a) Find the joint CDF and joint PDF of the random variables

   $$W := X^2 \qquad \text{and} \qquad Z := Y^2.$$

   (b) Show that if X and Y are independent, W and Z are independent as well.

6. Let $X_1, X_2, \ldots, X_n \sim Exp(\lambda)$ be IID, and define $Y = \max\{X_1, X_2, \ldots, X_n\}$. Find the PDF of Y. [Hint: Notice that $Y \leq y$ if and only if $X_i \leq y$ for every $i \in \{1, \ldots, n\}$.]

Lecture 5

Review of probability theory V

Contents

5.1 Expected value

Discrete random variables. Consider a discrete random variable $X : \Omega = \{\omega_1, \omega_2, \ldots\} \to \mathbb{R}$ and a real valued function $g(X)$ taking arguments on the state space. We define the ***expected value*** or ***expectation*** as

$$E[g(X)] := \sum_{\omega_i \in \Omega} g(X(\omega_i)) P\{\omega_i\},$$

whenever the sum converges absolutely[1]. This definition, while intuitive (the expected value considers the value of the function g when evaluated at every event in the sample space appropriately weighted by the probability of the corresponding event), does not provide a practical formula for computing the expected value of g in terms of its probability mass function p_X.

In order to obtain such a formula, we first define for every real number x in the state space S (i.e. all the possible values of X) the set of events ω_i such that $X(\omega_i) = x$ as

$$R(x) := \{\omega \in \Omega : X(\omega) = x\}.$$

[1]Recall that a sum $\sum_{n=1}^{\infty} x_i$ is said to converge absolutely if the sum $\sum_{n=1}^{\infty} |x_i|$ converges.

We then proceed from the definition as follows

$$
\begin{aligned}
E[g(X)] &:= \sum_{\omega_i \in \Omega} g(X(\omega_i))P\{\omega_i\} && \text{(Definition)} \\
&= \sum_{x \in S} \sum_{\omega \in R(x)} g(X(\omega))P\{\omega\} && \text{(Group all events yielding the same value of } x \text{ and sum over all } x \in S) \\
&= \sum_{x \in S} \sum_{\omega \in R(x)} g(x)P\{\omega\} = \sum_{x \in S} g(x) \sum_{\omega \in R(x)} P\{\omega\} && \text{(For all } \omega \in R(x) \text{ we have: 1) } X(\omega) = x \text{ and 2) } g(x) \text{ is constant)} \\
&= \sum_{x \in S} g(x)P(\omega \in R(x)) && \text{(Since all } w \in R(x) \text{ are disjoint)} \\
&= \sum_{x \in S} g(x)P(X = x) && \text{(Since all } w \in R(x) \text{ implies that } X = x) \\
&= \sum_{x \in S} g(x)p_X(x) && \text{(Definition of the probability mass function } p_X).
\end{aligned}
$$

Hence, we have shown that the expected value of the function g can be computed from the probability mass function by

$$E[g(X)] = \sum_{x \in S} g(x)p_X(x) \tag{5.1}$$

whenever the sum converges absolutely.

Continuous random variables. We will use this result to derive a similar formula for a continuous random variable. First, we consider a small value $\Delta x > 0$ and, given a continuous random variable X, we will define the discrete variable

$$\widetilde{X} := \lfloor X \rfloor_{\Delta x},$$

where the function $\lfloor \cdot \rfloor_{\Delta x}$ rounds down the argument to the nearest integer multiple of Δx. The discrete variable $\widetilde{X}$ has a discrete state space that we will denote by $S_{\Delta x}$. Moreover, we observe that

$$\lim_{\Delta x \to 0} |\widetilde{X} - X| = 0 \qquad \text{and} \qquad \lim_{\Delta x \to 0} S_{\Delta x} = S.$$

We then use (5.1) to compute

$$
\begin{aligned}
E[g(\widetilde{X})] &= \sum_{\widetilde{x} \in S_{\Delta x}} g(\widetilde{x})p(\widetilde{x}) \\
&= \sum_{\widetilde{x} \in S_{\Delta x}} g(\widetilde{x})P(\widetilde{X} = \widetilde{x}) \\
&= \sum_{\widetilde{x} \in S_{\Delta x}} g(\widetilde{x})P(\widetilde{x} \leq X < \widetilde{x} + \Delta x) \qquad (\text{since } \widetilde{X} = \widetilde{x} \Leftrightarrow \widetilde{x} \leq X < \widetilde{x} + \Delta x) \\
&= \sum_{\widetilde{x} \in S_{\Delta x}} g(\widetilde{x}) \left(F_{\widetilde{X}}(\widetilde{x} + \Delta x) - F_{\widetilde{X}}(\widetilde{x}) \right).
\end{aligned}
$$

Then, recalling that we are assuming that the CDF of a continuous random variable is differentiable we can use a Taylor approximation to see that

$$F_X(\widetilde{x} + \Delta x) = F_X(\widetilde{x}) + \Delta x f_X(\widetilde{x}) + \mathcal{O}\left(\Delta x^2\right),$$

therefore

$$E[g(\widetilde{X})] = \sum_{\widetilde{x} \in S_{\Delta x}} g(\widetilde{x}) \left(f_X(\widetilde{x})\Delta x \ + \mathcal{O}\left(\Delta x^2\right) \right).$$

We observe that the first term above is actually a Riemann sum, while the second one will vanish as $\Delta x \to 0$. Hence, if the sum above converges absolutely for any Δx we see that

$$E[g(X)] = \lim_{\Delta x \to 0} E[g(\widetilde{X})] = \lim_{\Delta x \to 0} \left(\sum_{\widetilde{x} \in S_{\Delta x}} g(\widetilde{x}) \left(f_X(\widetilde{x}) \Delta x + \mathcal{O}\left(\Delta x^2\right) \right) \right) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx.$$

Therefore, whenever the integral below converges absolutely, the expected value of the function g with respect to the random variable X will be given by

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{5.2}$$

Remark 5.1. The expected value of a quantity is also referred to as the ***mean value***. The following alternate notations for the expected value of the function g are also very common in the literature

$$E[g(X)] \equiv \overline{g(X)} \equiv \langle g(X) \rangle.$$

Remark 5.2. Notice that if the function g is the identity, then by substituting $g(x) = x$ in either (5.1) or (5.2) we obtain formulas for the mean value of the random variable itself

$$E[X] = \sum_{x \in S} x p_X(x) \quad \text{for discrete RVs, or} \qquad E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \quad \text{for continuous RVs.}$$

The mean or expected value of X is often denoted by μ_X or simply μ.

Theorem 5.1. *If $X_1, \ldots, X_n$ are all random variables with well-defined means, and $c_1, \ldots, c_n$ are real numbers, then*

$$E\left[\sum_{i=1}^{n} c_i X_i \right] = \sum_{i=1}^{n} c_i E\left[X_i\right] \tag{5.3}$$

5.2 Some special cases

Some special choices of the function g are particularly important. Here we provide a summary of the most common

1. If $g(x) = x^k$ for $k = 1, 2, 3, \ldots$ then $E[g(X)]$ is called the **k-th *moment***. Note that the first moment corresponds to the mean value of X.
2. If $g(x) = (x)_k$ where $(x)_k := (x)(x-1)(x-2)\cdots(x-k+1)$, then $E[g(X)]$ is called the **k-th *factorial moment***.
3. If we denote the mean value of X by $\mu := E[X]$ and define the function $g(x) = (x-\mu)^k$, then $E[g(x)]$ is called the **k-th *central moment***
4. If $\boldsymbol{X}$ is vector valued in $\mathbb{R}^d$, and we denote the ***Euclidean inner product*** between $\boldsymbol{X} := (X_1, \ldots, X_d)$ and $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_d)$ as
$$\langle \boldsymbol{\theta}, \boldsymbol{X} \rangle := \sum_{i=1}^{d} \theta_i\, X_i$$
then the function
$$\phi(\boldsymbol{\theta}) = E\left[e^{i \langle \boldsymbol{\theta}, \boldsymbol{X} \rangle} \right]$$
is called the ***characteristic function*** or the ***Fourier transform*** of $f_{\boldsymbol{X}}$.

5. If $\boldsymbol{X}$ is vector valued in $\mathbb{R}^d$, then the function

$$m(\boldsymbol{\theta}) := E\left[e^{\langle \boldsymbol{\theta}, \boldsymbol{X}\rangle}\right]$$

is called the ***moment generating function*** or the ***Laplace transform*** of $f_{\boldsymbol{X}}$.

6. If X takes values in the positive integers $\mathbb{Z}^+$ and $g(x) = z^x$, then the function

$$\rho(z) := E[g(x)] = \sum_{n=0}^{\infty} P\{X = x\} z^x$$

is called the ***probability generating function***.

5.3 Exercises

1. Let c be a fixed real number, X be a discrete random variable and Y be a continuous random variable. Prove that
   (a) The expected value of c with respect to X is $E[c] = c$.
   (b) The expected value of c with respect to Y is $E[c] = c$.

2. We will use induction to prove a slightly more general version of the identity (5.3) in three steps.
   (a) First let X_1 and X_2 be random variables over a countable sample space Ω having a common state space S. Let g_1 and g_2 be two integrable real-valued functions taking arguments on S, and let c_1 and c_2 be real numbers. Show that

$$E[c_1 g_1(X_1) + c_2 g_2(X_2)] = c_1 E[g_1(X_1)] + c_2 E[g_2(X_2)].$$

   Note that **equation (5.1) does not apply in this case** since we have *two* random variables involved. Instead you will have to study the steps that were used to derive (5.1) and adapt them to this case.
   (b) Then use the induction hypothesis, (i.e. that the result holds for $n = k$) and show that this implies that the result holds for $n = k + 1$.
   (c) Finally, choose g_i appropriately and conclude that (5.3).

3. (a) Consider two discrete random variables X and Y with well-defined means and such that $X \leq Y$. Show that

$$E[X] \leq E[Y].$$

   Hint: write $p_X(x) = p_Y(x) + d^+(x) - d^-(x)$ where

$$d^+(x) := \max\{p_X(x) - p_Y(x), 0\} \qquad \text{and} \qquad d^-(x) := |\min\{p_X(x) - p_Y(x), 0\}|.$$

   (b) Modify your proof above to consider the case when X and Y are continuous RVs.

4. Let X and Y be independent continuous random variables. Use Corollary 4.4 to Prove that $E[XY] = E[X]E[Y]$. [Hint: handle carefully the limits of integration.]

5. Consider a real valued continuous random variable X and a real number μ. We say that X is symmetric about μ, if

$$f_X(\mu + x) = f_X(\mu - x).$$

   Consider that X is symmetric about μ and prove that

(a) The random variables $Z_1 := X - \mu$ and $Z_2 := \mu - X$ have the same PDF.

(b) $E[X] = \mu$. [Hint use part (a) to conclude that $E[Z_1] = E[Z_2]$ and derive the result (b) from this equality].

6. Let X be a non-negative continuous random variable with PDF f_X and CDF F_X.

(a) Show that X has finite expectation if and only if

$$\int_0^\infty (1 - F_X(x))dx < \infty$$

(b) Show that

$$E[X] = \int_0^\infty (1 - F_X(x))dx;$$

7. Let $Z_1, \ldots, Z_n$ be random variables such that $E[Z_i] = \zeta$ for $i = 1, \ldots, n$. We define the ***sample mean*** of size n as the random variable given by

$$\overline{Z}_n := \frac{1}{n}\left(Z_1 + \ldots + Z_n\right).$$

Prove that $E[\overline{Z}_n] = \zeta$.

Lecture 6

Review of probability theory VI

Contents

6.1 Variance, covariance and correlation

The second central moment is of particular importance. It is called the ***variance*** of X, denoted as σ^2 and is defined as

$$\mathrm{Var}(X) \equiv \sigma^2 := E\left[(X-\mu_X)^2\right],$$

where μ is the expected value of X. The definition above can be used to compute the variance, but the following formula (you will have to prove it as an exercise) provides an easier tool for computing the variance

$$\sigma^2 = E\left[X^2\right] - (E[X])^2. \tag{6.1}$$

Note that, since the variance is defined as the expected value of a non-negative random variable, then it has a well defined square root, which is called the ***standard deviation*** and it is denoted by σ. This quantity is commonly used as a measure of the spread of a distribution.
The following useful properties of the variance can be proven from the definition and the properties of expectation.

Theorem 6.1. *If X is a random variable with well defined mean μ, and c is a real number then:*

1. *Var$[c] = 0$,*
2. *Var$[X + c]$ = Var$[X]$,*
3. *Var$[cX] = c^2$Var$[X]$.*

Proof. Exercise 2. □

Two related quantities are the ***covariance***

$$\mathrm{Cov}(X,Y) := E\left[(X-\mu_X)(Y-\mu_Y)\right],$$

and the ***correlation***

$$\mathrm{Corr}(X,Y) \equiv \rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y},$$

which provide a measure of the linear connection between the random variables X and Y.
The following properties of the covariance and the correlation are easy to prove

Theorem 6.2. *If X and Y are variables with joint PDF $f_{X,Y}(x,y)$, well defined means μ_X and μ_Y and with positive variances σ_X and σ_Y then*

$$|Corr(X,Y)| \leq 1. \tag{6.2}$$

Moreover, if X and Y are independent then

$$Cov(X,Y) = 0. \tag{6.3}$$

When the equality above holds true, we say that X and Y are **uncorrelated**.

Proof. We will prove the first property for continuous random variables (the proof for the discrete case is completely analogous) and will leave the proof of the second property as Exercise 3.
We start from the definition of expectation and will use the fact that a PDF is non negative and the Cauchy-Schwarz inequality.

$$\begin{aligned}
|\mathrm{Cov}(X,Y)| &= |E\left[(X-\mu_X)\,(Y-\mu_Y)\right]| \\
&= \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f_{X,Y}(x,y)\,dxdy\right| \\
&= \left|\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)\left(\sqrt{f_{X,Y}(x,y)}\right)\left((y-\mu_y)\sqrt{f_{X,Y}(x,y)}\right)\,dxdy\right| \\
&\leq \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)^2 f_{X,Y}(x,y)\,dxdy\right)^{1/2}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y-\mu_y)^2 f_{X,Y}(x,y)\,dxdy\right)^{1/2} \\
&= \sigma_X\,\sigma_Y
\end{aligned}$$

Hence

$$|\mathrm{Corr}(X,Y)| = \frac{|\mathrm{Cov}(X,Y)|}{\sigma_X\sigma_Y} \leq 1.$$

□

If we have two random variables X and Y with well defined means μ_X and μ_Y, we can compute the variance of the random variable $Z := X+Y$. We first note that if we denote the expected value of Z as μ_Z we have

$$\mu_Z = E[X+Y] = E[X] + E[Y] = \mu_X + \mu_Y.$$

We then proceed as follows

$$\begin{aligned}
\mathrm{Var}[Z] &= E\left[(Z-\mu_Z)^2\right] \\
&= E\left[Z^2 - 2Z\mu_Z + \mu_Z^2\right] \\
&= E\left[X^2 + 2XY + Y^2 - 2(X+Y)(\mu_X+\mu_Y) + \mu_X^2 + \mu_Y^2 + 2\mu_X\mu_Y\right] \\
&= E\left[X^2 - 2X\mu_X + \mu_X^2 + Y^2 - 2Y\mu_Y + \mu_Y^2 + 2\left(XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y\right)\right] \\
&= E\left[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)\right] \\
&= E\left[(X-\mu_X)^2\right] + E\left[(Y-\mu_Y)^2\right] + 2E\left[(X-\mu_X)(Y-\mu_Y)\right] \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2E\left[(X-\mu_X)(Y-\mu_Y)\right] \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Corr}(X,Y).
\end{aligned}$$

We have thus proven that

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Corr}(X,Y). \tag{6.4}$$

Moreover, if X and Y are uncorrelated

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]. \tag{6.5}$$

6.2 Exercises

1. Using the definition of variance and the properties of the expected value prove the identity (6.1)

$$\sigma^2 = E\left[X^2\right] - (E[X])^2 .$$

2. Prove the three properties in Theorem 6.1.

3. Here we will prove property (6.3) from Theorem 6.2. We will proceed in two steps

   (a) Using the properties of the expectation and the definition of covariance, show that for any RVs X and Y

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

   (b) Show that if X and Y are independent, the expression above implies that $\text{Cov}(X, Y) = 0$. [Hint: See problem 4 from Lecture 5].

4. Let X and Y be random variables with mean 0, variance 1, and correlation ρ.

   (a) Show that Y and $X - \rho Y$ are uncorrelated.

   (b) Show that $X - \rho Y$ has mean 0 and variance $1 - \rho^2$.

5. Here we explore more properties of the ***sample mean***. Show that, if $Z_1, \ldots, Z_n$ are uncorrelated RVs with well defined means and having common variance σ^2, then the variance of the sample mean $\overline{Z}_n$ is given by

$$\text{Var}\left(\overline{Z}_n\right) = \frac{\sigma^2}{n}.$$

Lecture 7

Review of probability theory VII

Contents

7.1 Sample mean

Let $X_1, \ldots, X_n$ be independent identically distributed random variables with finite mean μ and non zero variance σ^2. The ***sample mean*** of size n, $\overline{X}_n$, often also called the ***empirical average*** is defined as

$$\overline{X}_n := \frac{1}{n}\left(X_1 + \ldots + X_n\right).$$

The cumulative distribution function for the sample mean is called the ***empirical distribution function*** and is defined as

$$\begin{aligned}\overline{F}_n(x) &:= P(\overline{X}_n \leq x) \\ &= \frac{1}{n}\,(\#\text{ of observations of } X_i \text{ that fall below } x) \\ &= \sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(x_i),\end{aligned}$$

where the function $\mathbb{I}_{(-\infty,x]}(x_i)$ is the indicator function of the interval $(-\infty, x]$ given by

$$\mathbb{I}_{(-\infty,x]}(x) = \begin{cases} 1 & \text{if } x \in (-\infty, x] \\ 0 & \text{if otherwise} \end{cases}$$

The sample mean is very important in statistics. You can think of the problem of estimating the mean value of certain population parameter that is inaccessible to you due to the impossibility of polling the entire population. Instead, you set out to poll a smaller subset of the population with size n, and compute the average of the response. In this situation, each random variable X_i represents the act of asking a randomly selected person, since all the people on your sample come from the same population they will all follow the same common distribution. Moreover, if you select your sample randomly (and you design your experiment appropriately)

then the answer provided by each individual will bare no direct relation with the answers of the rest, i.e. your sample will consist of independent identically distributed individuals. Since a different poller conducting the same experiment will end up interviewing a different set of individuals, $\overline{X}_n$ is indeed a random variable. Your empirical average will correspond to one realization or observation of the possible values of $\overline{X}_n$. You might then ask how well will your empirical average approximate the value of the population mean μ. We will address this question in what follows.

7.2 Law of large numbers

In exercises 7 from Lecture 5 and 5 from Lecture 6 we have already explored two important properties of this random variable, which we will recall here.

1. $E\left[\overline{X}_n\right] = \mu$,
2. $\text{Var}\left[\overline{X}_n\right] = \dfrac{\sigma^2}{n}$.

We can combine these two properties into the following mathematical statement

$$\int_S (x-\mu)^2 f_{\overline{X}_n}(x)\, dx = \frac{\sigma^2}{n},$$

where the integral is computed over all possible values of $\overline{x}_n$ (the state space S). From the expression above, it follows that

$$\lim_{n\to\infty} \int_S (x-\mu)^2 f_{\overline{X}_n}(x)\, dx = 0.$$

Since all of the terms inside the integral above are non negative, the only way in which the limit can be zero is if somehow the possible values of $\overline{x}_n$ get closer and closer to the distribution mean μ as more and more samples are gathered. Note that, due to the presence of the PDF and the fact that the entire expression is being integrated, the statement above **does not imply** that $\lim_{n\to\infty} \overline{x}_n = \mu$, but rather than, as the number of samples increases it becomes more and more unlikely to find sample mean values that differ much from the distribution mean μ. We will make this statement more precise with the aid of the following inequality.

Theorem 7.1 (***Chebyshev's Inequality***). *Let X be a real valued random variable with finite mean μ and variance σ^2, probability density function f_X, and let k be a positive constant. Then, the probability that the distance between a realization of X and the mean μ is larger than or equal to k is bounded as*

$$P(|X-\mu| \geq k) \leq \frac{Var(X)}{k^2}. \tag{7.1}$$

Proof. We start by defining the set where the distance between X and μ is at least k

$$A_k := \{x : |X-\mu| \geq k\},$$

and noting that for every $x \in A_k$ it holds that $1 \leq (x-\mu)^2/k^2$. We want to compute the probability that a

sample of X belongs to A_k, so we proceed from the definition

$$\begin{aligned} P(|X-\mu| \geq k) &= \int_{A_k} f_X(x)\,dx \\ &\leq \int_{A_k} \frac{(x-\mu)^2}{k^2} f_X(x)\,dx \\ &\leq \frac{1}{k^2}\int_{A_k} (x-\mu)^2 f_X(x)\,dx + \frac{1}{k^2}\int_{A_k^c} (x-\mu)^2 f_X(x)\,dx \\ &= \frac{1}{k^2}\int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\,dx \\ &= \frac{\mathrm{Var}(X)}{k^2}. \end{aligned}$$

□

We can then combine this result with the particular properties of the sample mean $\overline{X}_n$ to state that

$$P(|\overline{X}_n - \mu| \geq k) \leq \left(\frac{\sigma}{k}\right)^2 \frac{1}{n} \tag{7.2}$$

This statement is known as the ***law of large numbers*** and it is sometimes expressed by saying that that the sample mean **converges with probability 1** to the distribution mean. The law of large numbers tells us that

- It is safe to estimate the expectation of a random variable by repeatedly sampling and computing the empirical average.
- If we agree use the standard deviation as a measure of the distance between our approximation and the true value, then the "size" of the error ϵ is given by

  $$\epsilon = \sqrt{\mathrm{Var}\left[\overline{X}_n\right]} = \frac{\sigma}{\sqrt{n}},$$

  where σ is the standard deviation of the *unknown* distribution. This fact might seem like an insurmountable issue. How are we supposed to know the variance if we do not know the expectation? As we shall see later, this is not such a big problem, as for some distributions it is possible to compute the variance theoretically, while for some others approximating σ by the variance of a sample provides a remarkably good approximation. This will lead to what is known as the T distribution.

7.3 Central limit theorem

The central limit theorem also pertains to the behavior of the sample mean as the number of samples increases, but it is a much stronger result than the law of large numbers. It is in fact one of the deepest results in applied mathematics, and its proof is not easy, which is why we will not attempt it here (it belongs in a course of advanced probability).
Consider a family of random variables X_n with corresponding CDFs F_{X_n}. We say that they ***converge in distribution*** to the random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for every point x where $F_X(x)$ is continuous.

We recall here the CDF of a normal random variable with mean μ and standard deviation σ, which is given by

$$\Phi(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

We can now state the ***central limit theorem***.

Theorem 7.2. *Central limit theorem Let $X_1, \ldots, X_n$ be independent identically distributed random variables with mean μ and standard deviation σ, and let $\overline{X}_n$ be their empirical average or sample mean. Then*

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{(\sigma/\sqrt{n})} \leq x\right) = \Phi(x).$$

Intuitively, what the theorem tells us is that, as we mix more and more independent identically distributed random variables through the empirical average, the only features that remain relevant in the CDF of the resulting random variable are their common mean μ and standard deviation σ. The remarkable and surprising fact is that this result remains true *regardless of the of the particular form of the underlying PDFs.*
One of the useful consequences of this result for statistics is that, regardless of what is the PDF of the variable we are studying we can approximate probabilistic statements about it by using with a normal distribution with the same mean and standard deviation, as long as our sample size is big enough.

7.4 Exercises

1. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent identically distributed random vectors with mean (μ_x, μ_y) and variance (σ_x^2, σ_y^2). Let

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \overline{Y}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

and define $Z_n := \overline{X}_n/\overline{Y}_n$. Find $\lim_{n\to\infty} F_{Z_n}(z)$.

Lecture 8

Statistical decision theory I

Contents

8.1 Decisions, loss and risk

The basic idea of ***inferential statistics*** is to make a decision based on information obtained from data observations $(x_1, \ldots, x_n)$. The observations are modeled as realizations of a random variable X that has a (fully or partially) unknown distribution F_X. [1] Moreover, in ***parametric statistics***, the distribution function is assumed to depend on one or more parameters that we will denote as $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_m)$. To emphasize the dependence of the probabilistic results on the parameter θ, we will denote the expectation as E_θ, the variance as Var_θ and so forth. The assumed *true* value of these parameters is sometimes referred to as the ***state of nature*** and is chosen from a set of admissible values known as the ***parameter space*** that we will denote as Θ.

In a well designed experiment, the outcome of one measurement should not depend on the previous ones (i.e. all measurements should be independent from each other) which is why a set of n measurements is considered—in the language of probability—to be a set of independent, identically distributed random variables $X_1, \ldots, X_n$. In the language of statistics a particular realization of IID random variables is known as a ***simple random sample***.

Given data in the state space $(x_1, \ldots, x_n) \in S_n$, we want to make a decision, which is a function of the data. The set of all possible decisions (i.e. functions) is called the ***action space*** A. This leads us to introduce the ***decision function*** or ***decision rule***

$$d : S_n \to A.$$

Decisions have consequences, and we would like to be able to quantify the negative effect of a decision made (and to minimize the negative impact). This leads to the introduction of a ***loss function*** which takes as arguments the values of the parameters and the decision made and gives back a real number that will be interpreted as ta measure of the consequences

$$\mathcal{L} : \Theta \times A \to \mathbb{R}.$$

[1] To *infer* means to deduce something hence, statistical inference is the process of we *learning* something new about our random variable by analyzing statistical information. This not-so-new idea of using a computer (a machine) to process statistical data in order to obtain information about certain process now goes by the much cooler name of *machine learning*.

Hence, it the state of nature is given by the parameter θ then $\mathcal{L}(\theta, a)$ quantifies the loss incurred upon taking the action a.

Note that, since the action is a function of the random variable $a = d(X)$, the loss is itself a random variable. Every time that an experiment, poll, measurement, etc. is conducted, a different sample $(x_1, \ldots, x_n)$ will be obtained and this will in turn lead to a different action a. We would like to measure the consequence of the decision in a way that is not random, and a very natural way is to consider what would be the expected loss incurred by our decision. This leads to the introduction of the ***risk function*** which takes a parameter and a decision function and gives back the expected loss incurred by it

$$\begin{aligned} \mathcal{R} : \Theta \times \mathcal{D} &\longrightarrow \mathbb{R} \\ (\theta, d) &\longmapsto \mathcal{R}(\theta, d) := E_\theta[\mathcal{L}(\theta, d(X))]. \end{aligned} \tag{8.1}$$

The particular way of measuring the loss is up to the user and may change from application to application. Different choices can lead to different optimal actions and decisions, as we shall see in the following example where we explore here three natural choices for the problem of parameter estimation.

8.2 Different loss functions lead to different optimal decisions

Consider the problem of approximating the value of the parameter θ by a certain number x determined by our data. The following three choices of loss functions seem all natural ways of quantifying the error in the approximation

1. $\mathcal{L}_1(\theta, a) = |\theta - a|$. (***Absolute error loss***)

2. $\mathcal{L}_2(\theta, a) = (\theta - a)^2$. (***Squared error loss***)

3. $\mathcal{L}_3(\theta, a) = \begin{cases} 0 & \text{if } a = \theta, \\ 1 & \text{if } a \neq \theta. \end{cases}$ (***Zero-one loss***)

We will consider that our data is drawn from a discrete random variable with probability mass function $p_X(x)$ which, for simplicity, will be considered to be independent of θ in the following calculations. The decision function takes the measured data and gives us back a number that will be used as an approximation of θ. This can be written as $d = d(X_1, \ldots, X_n)$. We would like to determine the choice for d that minimizes the risk with respect to each of the choices of loss functions.

In the following example we will consider the decision function $d(x) = x$ and will explore the different optimal answers that stem from each of the choices of loss function listed above.

1. We start with $\mathcal{L}_1(\theta, a)$ and we compute the risk associated to the decision $d(x) = x$.

$$\begin{aligned}
\mathcal{R}_1(\theta, d) &= E_\theta[\mathcal{L}_1(\theta, d(X_1, \ldots, X_n)] \\
&= E_\theta[|\theta - x|] \\
&= \sum_{x \in S_n} |\theta - x| p_X(x) \\
&= \sum_{x \leq \theta} |\theta - x| p_X(x) + \sum_{x > \theta} |\theta - x| p_X(x) \\
&= \sum_{x \leq \theta} (x - \theta) p_X(x|\theta) + \sum_{x > \theta} (\theta - x) p_X(x) \\
&= \theta \left(\sum_{x > \theta} p_X(x) - \sum_{x \leq \theta} p_X(x|\theta) \right) + \sum_{x < \theta} x p_X(x) - \sum_{x > \theta} x p_X(x) \\
&= \theta \big(P(X > \theta) - P(X \leq \theta) \big) + \sum_{x < \theta} x p_X(x) - \sum_{x \geq \theta} x p_X(x) \\
&= \theta \big(1 - P(X \leq \theta) - P(X \leq \theta) \big) + \sum_{x < \theta} x p_X(x) - \sum_{x \geq \theta} x p_X(x) \\
&= \theta \big(1 - 2P(X \leq \theta) \big) + \sum_{x < \theta} x p_X(x) - \sum_{x \geq \theta} x p_X(x).
\end{aligned}$$

   We then recall that, at a minimum value, it must hold that

$$\frac{\partial}{\partial \theta} \mathcal{R}_1(\theta, d) = \big(1 - 2P(X \leq \theta) \big) = 0.$$

   This implies that the optimal value of X must satisfy

$$P(X \leq \theta) = 1/2,$$

   which is the definition of the median. Hence, the risk would be minimized by taking x to be the median of the data.

2. For $\mathcal{L}_2(\theta, d)$ and we proceed similarly and start by computing the risk associated to the decision $d(x) = x$

$$\mathcal{R}_2(\theta, d) = E_\theta \left[(\theta - x)^2 \right] = \sum_x (\theta - x)^2 p_X(x).$$

   Hence, computing the derivative with respect to θ and imposing the optimality condition of vanishing derivative we obtain

$$\frac{\partial}{\partial \theta} \mathcal{R}_2(\theta, d) = 2 \sum_x (\theta - x) p_X(x) = 2(\theta - E_\theta[X]) = 0.$$

   Hence, the value of X that minimizes the risk it the one satisfying

$$E_\theta[X] = \theta,$$

   namely, the mean of the data. When the square loss is used, the risk function is also commonly known as the ***mean squared error*** (or ***MSE*** for short), and the square root of the risk is referred to as ***root mean squared error*** (or ***RMSE*** for short).

3. For $\mathcal{L}_3(\theta, d)$ and $d(x) = x$ we have that

$$\mathcal{R}_3(\theta, d) = 0 \cdot p_X(X = \theta) + 1 \cdot p_X(X \neq \theta) = p_X(X \neq \theta)$$

Hence the risk will be minimized if the probability of $X \neq \theta$ is as small as possible. This is equivalent to making $p_X(X = \theta)$ as large as possible which in turn implies that the best choice would be to make x equal to the mode of the data.

8.3 Exercises

1. Consider the PDF

$$f_X(x, \theta) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)(x-\theta)^2},$$

and the loss function $\mathcal{L}(\theta, a) = (a - \theta)^2$. Suppose that a single observation of X can be measured. Compute the value of $\mathcal{R}(\theta, d)$ if d is chosen to be the function $d(x) = cx$.

2. Consider a discrete random variable X binomially distributed with

$$f_X(x, \theta) = \binom{2}{x} \theta^x (1 - \theta)^{2-x} \quad (\text{ where } x = 0, 1, 2 \quad \text{and } 0 < \theta < 1),$$

and a loss function $\mathcal{L}(\theta, a) = (a - \theta)^2$.

   (a) Calculate $\mathcal{R}(\theta, d)$ for $d(x) = x/2$.

   (b) Calculate $\mathcal{R}(\theta, d)$ for $d(x) = (x + 1)/4$.

Lecture 9

Statistical decision theory II

Contents

9.1 Comparing decision functions

Given a loss function, our goal is to use statistics to find a "good" decision function that minimizes the risk. The fact that the probability density functions involved will, in general, depend on parameter values implies that the risk associated to a particular decision is itself a function of the parameter. Ideally, we would compare two different decisions d_1 and d_2 by analyzing the plots of their respective risks and choosing the one that always stays below (as in the left panel of Figure 9.1). However, the scenario where the risk associated to d_1 is smaller than that of d_2 for all values of θ rarely happens. Instead, it is common for the graphs to cross over as the parameter value changes (as depicted in the right panel of Figure 9.1). This may lead to situations where two different decisions d_1 and d_2 can be made, and two different parameter possible parameter values θ_1 and θ_2 can be chosen for which

$$\mathcal{R}(\theta_1, d_1) < \mathcal{R}(\theta_1, d_2) \qquad \text{but} \qquad \mathcal{R}(\theta_2, d_1) > \mathcal{R}(\theta_1, d_2).$$

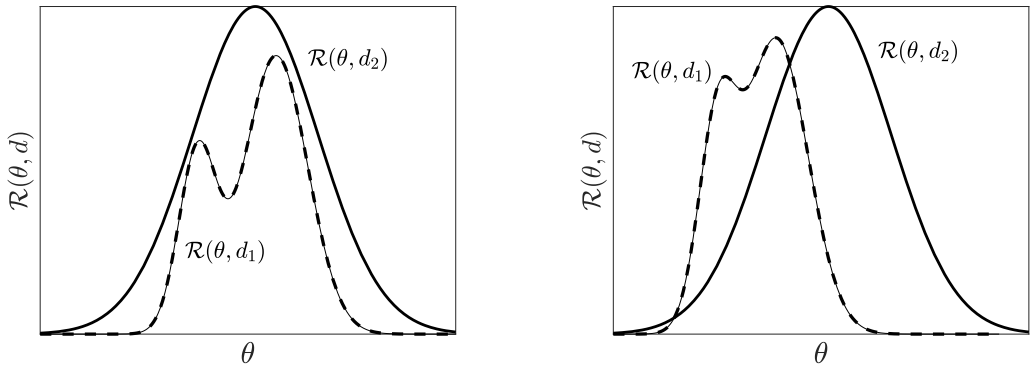

Figure 9.1: Left: the risk associated to the decision d_1 remains below that for decision d_2 for all values of θ, making the decision d_1 the better option. Right: Depending on the value of θ the risk associated with d_1 is larger or smaller than that associated with d_2, making a choice between the two decisions unclear.

Since in practical applications the "true" value of the parameters involved is not known and one must rely in an approximation $\widehat{\theta}$, we would like to be able to compare different decision functions and chose the one that is "better" even if our approximation $\widehat{\theta}$ is not optimal or accurate.

9.2 Minimax decision rules

One rule that can guide our choice is to consider the worst case scenario for two different decisions, and then chose the one for which the worst possible outcome is less severe. Since in our case the worst case scenario involves a very large risk, we will first compute the maximum risk (as a function of the parameter θ) for each of the decision functions under consideration, and then chose the function with the smallest maximum risk. When we chose a decision function based on this criterion we say that d is the best decision function in the ***minimax sense***. We can make this rigorous through the following definition

Definition 9.1. The function d_0 is called a ***minimax decision function*** in the class D if it satisfies

$$\max_{\theta\in\Theta} \mathcal{R}(\theta, d_0) = \min_{d\in D} \max_{\theta\in\Theta} \mathcal{R}(\theta, d),$$

where Θ is the set of all possible parameter values and D is the set of decision functions under consideration.

Example. Consider a Poisson distribution with density

$$f_X(x,\theta) = \frac{e^{-\theta}\theta^x}{x!} \qquad (x \in \mathbb{Z}^+),$$

the family of decision functions $d(x) = cx$ with $c > 0$, and the loss function $\mathcal{L}(\theta, d) = (d-\theta)^2/\theta$. We want to study the risk arising from this family of decision functions, hence we want to compute $E[\mathcal{L}(\theta, cx)]$. To simplify the computation, we will use the fact that X is a Poisson random variable and therefore

$$E[X] = \theta \qquad \text{and} \qquad \text{Var}[X] = \theta.$$

Then we will rewrite the loss function in the following way

$$\mathcal{L}(\theta, cx) = \frac{(cX-\theta)^2}{\theta} = \frac{c^2}{\theta}\left((X-\theta)^2 + 2\theta(1-1/c)(X-\theta) + \theta^2(1-1/c)^2\right).$$

Hence

$$\mathcal{R}(\theta, cx) = E[\mathcal{L}(\theta, cx)] = c^2 + \theta(c-1)^2 = (1+\theta)c^2 - 2\theta c + \theta.$$

This expression, as a function of c, is a positive parabola with vertex at $(1, 1)$. Therefore, the smallest possible risk will be equal to one and will happen for $c = 1$. Hence, the minimax decision function among the class $D := \{d(x) = cx : c > 0\}$ is $d(x) = x$.

By choosing a minimax decision function we are guarding against the worst possible outcome, but this may come at the cost of being "too pessimistic" or of incurring on higher risks than necessary for non-critical parameter values. In the right panel of Figure 9.1, decision d_1 would be the best decision in the minimax sense, since its maximum value is smaller than that of the curve generated by d_2, but for small values of θ (i.e. on the left of the graph), the risk associated with d_1 is in fact larger than that stemming from d_2. It would then seem reasonable then to consider how likely it is that we would find ourselves with parameter values that realize the worst case scenarios. If such an outcome is highly unlikely, then a different decision function may be chosen. If we chose this approach, we must consider then Θ as a random variable, which is the hallmark of ***Bayesian statistics***.

9.3 Exercises

1. Consider the PDF
$$f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)(x-\theta)^2},$$
and the loss function $\mathcal{L}(\theta, a) = (a-\theta)^2$. Determine whether there is a minimax decision function in the class of functions $d(x) = cx$.

2. Consider the discrete PDF
$$f(x|\theta) = \binom{2}{x}\theta^x(1-\theta)^{2-x} \quad (\text{ where } x = 0, 1, 2 \quad \text{and } 0 < \theta < 1),$$
and the loss function $\mathcal{L}(\theta, a) = (a-\theta)^2$. In exercise 2 from Lecture 8 you computed $\mathcal{R}(\theta, d)$ for $d_1(x) = x/2$ and $d_2(x) = (x+1)/4$. Which of the two functions is superior according to the minimax rule?

3. A coin is known to be biased with either $p = 1/4$ or $p = 3/4$, where p is the probability of heads. Which of the two probabilities is accurate is not known and a decision is to be made between these two values on the basis of the outcome of two tosses of the coin. Consider the loss function to be $\mathcal{L}(p, a) = (a-p)^2$

   (a) Calculate the value of $\mathcal{R}(p, d)$ for the three different decision functions listed below, where X denotes the number of heads obtained in two tosses
$$d_1(X) = \begin{cases} 1/4 & \text{if } X = 0 \\ 1/4 & \text{if } X = 1 \\ 1/4 & \text{if } X = 2 \end{cases} \quad d_2(X) = \begin{cases} 3/4 & \text{if } X = 0 \\ 3/4 & \text{if } X = 1 \\ 3/4 & \text{if } X = 2 \end{cases} \quad d_3(X) = \begin{cases} 3/4 & \text{if } X = 0 \\ 1/4 & \text{if } X = 1 \\ 1/4 & \text{if } X = 2 \end{cases}$$

   (b) Which is the minimax decision function with respect to these three options?

4. Let θ be a parameter taking only the values 0 or 1, X be a discrete random variable taking only non-negative integer values (X = 0,1,2,...). We define the probability density function
$$f_X(x, \theta) = \begin{cases} f_X(x, 0) = 2^{-x} & \text{if } \theta = 0 \\ f_X(x, 1) = 2^{-(x+1)} & \text{if } \theta = 1 \end{cases}$$
and a loss function $\mathcal{L}(\theta, d(x))$ such that
$$\mathcal{L}(0, 0) = \mathcal{L}(1, 1) = 0 \qquad \text{and} \qquad \mathcal{L}(1, 0) = \mathcal{L}(0, 1) = 1.$$
Consider the two decision functions
$$d_1(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{if } x \neq 0 \end{cases} \qquad \text{and} \qquad d_2(x) = \begin{cases} 0 & \text{if } x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}.$$
Calculate the risk $\mathcal{R}(\theta, d)$ for these two functions and determine which one is minimax with respect to the two of them.

Lecture 10

Statistical decision theory III

Contents

10.1 Likelihood, prior and posterior

The starting point of ***Bayesian statistics*** is to consider that the parameters appearing in the probability density functions of data measurements *need not be fixed numbers*. Instead, the Bayesian paradigm is to consider the parameter θ as a realization of the random variable Θ with density $\pi(\theta)$ and cumulative distribution $\Pi(\theta)$. In this case, the probability density function $\pi(\theta)$ is referred to as the ***prior density***. The name stems from the fact that we are assumed to have some previous (prior) knowledge about the general behavior of the parameter, which translates into the knowledge of the functional form of $\pi(\theta)$.
In this context, the density function $f_X(x,\theta)$ is referred to as the ***likelihood function*** and should be understood as a ***conditional probability density function***[1] which justifies the notation

$$f_X(x,\theta) \equiv f_{X|\Theta}(x|\theta).$$

The expression on the left side of the equivalence is preferred in the probability literature, while the notation on the right is widespread in statistics. Since our subject matter is statistics we shall stick to the latter, but the reader should keep in mind that the two different notations refer to the exact same concept.
Recalling the definitions (4.2) and (4.3), the joint density function for X and Θ, $f_{X,\Theta}(x,\theta)$, is then given in terms of the prior distribution and the likelihood function by

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)\pi(\theta). \tag{10.1}$$

We can use Baye's formula (2.1) in the expression above to rewrite $f_{X|\Theta}(x|\theta)$, leading to

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)\pi(\theta) = \left(f_{\Theta|X}(\theta|x)\frac{f_X(x)}{\pi(\theta)}\right)\pi(\theta) = f_{\Theta|X}(\theta|x)f_X(x). \tag{10.2}$$

The PDF $f_X(x)$ appearing in the right hand side of the expression above is the marginal distribution for X, while the term $f_{\Theta|X}(\theta|x)$ is the conditional probability for Θ *given the data measurements* x and is known

[1] In this context we will abuse nomenclature and will use the term density for both continuous and discrete random variables, although in the latter case we still mean "mass".

as the ***posterior distribution***. The name refers to the fact that $f_{\Theta|X}(\theta|x)$ provides information about the behavior of Θ *after* incorporating the knowledge of a particular set of measurements, x, of the variable X. The posterior distribution gives information about the question "*How likely it is that the value of the random variable Θ is equal to θ given that we have measured a particular value x of the random variable X?*".

10.2 Bayesian decision rules

Since we are now considering Θ to be a random variable, the risk function that was defined in (8.1) is no longer a real number, but a random variable itself. Hence, it can no longer be used to compare two different decisions. We will now consider how risky a decision is *on average* as the parameter value changes. This leads to the definition of the quantity

$$r(\pi, d) = E[\mathcal{R}(\theta, d)] = \begin{cases} \int \mathcal{R}(\theta, d)\, \pi(\theta)\, d\theta & \text{if } \Theta \text{ is continuous} \\ \sum \mathcal{R}(\theta, d) p_\Theta(\theta) & \text{if } \Theta \text{ is discrete} \end{cases} \tag{10.3}$$

which is known as the ***mean risk*** or ***Bayes risk***. The sum and the integral in the definition are carried over all possible values of θ. As defined above, the mean risk is a real number that can be then used as a criterion to compare different decisions.

Definition 10.1. A decision function d_0 is called a ***Bayes decision function*** with respect to the prior $\pi(\theta)$ and the class D of decision functions if its mean risk as a function of the parameter θ following the distribution $\pi(\theta)$ is the smallest amongst all functions in the class D. Put in mathematical terms, if it satisfies

$$r(\pi, d_0) = \min_{d \in D} r(\pi, d).$$

Example. We will go back to the example in the previous section, where we considered a Poisson random variable with distribution

$$f_X(x, \theta) = \frac{e^{-\theta}\theta^x}{x!} \qquad (x \in \mathbb{Z}^+),$$

and the family of decision functions $d(x) = cx$ with $c > 0$, and the loss function $\mathcal{L}(\theta, d) = (d - \theta)^2/\theta$. There we showed that for this particular loss function and this class of decision functions, the risk is given by

$$\mathcal{R}(\theta, cx) = E[\mathcal{L}(\theta, cx)] = c^2 + \theta(c - 1)^2.$$

Since in the Bayesian approach θ is taken to be a random variable will will consider that it is distributed according to the prior density

$$\pi(\theta) = e^{-\theta} \quad \text{for } \theta > 0.$$

Hence, the mean risk or Bayes risk is given by

$$r(\pi, cx) = E[\mathcal{R}(\theta, cx)] = \int_0^\infty (c^2 + \theta(c - 1)^2)\, e^{-\theta}\, d\theta = c^2 + (c - 1)^2.$$

We must now find the minimum mean risk amongst the class $c > 0$. The optimality condition is then

$$\frac{d}{dc} r(\pi, cx) = 4c - 2 = 0$$

which will hold for $c = 1/2$. Therefore the Baye's decision function over this class will be $d(x) = \frac{x}{2}$ with respect to the prior $\pi(\theta) = e^{-\theta}$. Contrast this with the minimax decision function of the previous section, where $c = 1$.

10.3 Exercises

1. Consider the likelihood function

$$f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)(x-\theta)^2},$$

the loss function $\mathcal{L}(\theta, a) = (a-\theta)^2$, and the class of decision functions $d(x) = cx$ for a constant value c. Assume that the parameter θ is known to be distributed according to the prior density $\pi(\theta) = 1/2$ for $-1 < \theta < 1$ (i.e. uniformly distributed between -1 and 1).

   (a) Calculate the mean risk

   (b) Find the value of c that produces the Bayes solution with respect to this prior distribution.

2. Consider a likelihood function following a binomial distribution

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad \text{with } x = 0, 1, 2, 3 \ldots, n \text{ and } \theta \in (0, 1),$$

and the square loss function $\mathcal{L}(\theta, a) = (a-\theta)^2$.

   (a) Calculate the risk function $\mathcal{R}(\theta, d)$ for the class of decision functions $d(x) = x/n$.

   (b) Given the prior $\pi(\theta) = 1$ for $\theta \in (0, 1)$ calculate the mean risk $r(\pi, d)$ for $d(x) = x/n$.

   (c) What is the Bayes decision over this class of functions? How does it compare to the minimax decision?

3. Repeat the previous problem using the loss function $\mathcal{L}(\theta, a) = \dfrac{(a-\theta)^2}{\theta(1-\theta)}$.

Lecture 11

Estimation I

Contents

11.1 Bias

Here and in the sequel, we shall use the term *density function* to refer to both the density function, if X is continuous, or to the mass function, if X is discrete. We will use interchangeably the symbols

$$f_X(x, \theta) \equiv f(x|\theta)$$

to denote the density function. Recall that the term ***likelihood function*** is often used in connection with the symbol $f(x|\theta)$ while the term probability density function is often used in connection with $f_X(x, \theta)$ however, both symbol and names refer to the same concept.
We begin with a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of independent identically distributed random variables following a family of probabilities P_θ depending on a parameter (or parameters) that will be denoted by θ. The goal of estimation is to determine which particular value of θ is the source of the measured data $\boldsymbol{X} = (X_1, \ldots, X_n)$, and how to use the data measurements to come up with an approximation of the unknown parameter.
We can frame this in the context of the decision theory we explored in the previous sections. In this case the decision function d is called an ***estimator*** and takes data values $(X_1, \ldots, X_n)$ to produce an approximation, also called an ***estimate***, of the unknown parameter value $\widehat{\theta}$ [1]

$$d(X_1, \ldots, X_n) = \widehat{\theta}.$$

Definition 11.1. A ***statistic*** is a function of the random variable that does not depend on any unknown parameter.

Definition 11.2. Consider a function of the unknown parameter $g(\theta)$ and an estimator to this quantity $d(\boldsymbol{X})$. We define the ***bias*** $b_d(\theta)$ is defined as

$$b_d(\theta) := E[d(\boldsymbol{X})] - g(\theta),$$

where the expectation is taken with respect to the random variable $\boldsymbol{X}$. Whenever the bias associated to an estimator is zero, we say that the estimator is ***unbiased***, otherwise we say that the estimator is ***biased***.

[1] It is customary in statistics to use the symbol $\widehat{\ }$ on top of an approximation to a quantity. Hence, θ denotes the true value of the parameter, and $\widehat{\theta}$ should be understood as an approximation to that value.

Example. (*sample mean*) If $X_1, \ldots, X_n$ are IID with mean μ, then the sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator for μ. Indeed, the expected value of the estimator is given by

$$E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}n\mu = \mu,$$

and therefore

$$b_{\overline{X}_n}(\theta) = E\left[\frac{1}{n}\sum_{i=1}^n X_i\right] - \mu = \mu - \mu = 0.$$

Example. (*sample variance*) Consider a sample $X_1, \ldots, X_n$ of IID random variables with mean μ and variance σ^2. In elementary statistics classes, it is usually taught that if one wants to approximate the variance σ^2 based on n measurements then one should use the quantity

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. This quantity is usually referred to as ***sample variance***. Given that the definition of the *true* variance is $E[(X-\mu)^2]$ one would think the most natural way of approximating the variance based on n measurements would be

$$\sigma^2 \approx \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

Therefore, the factor $\frac{1}{n-1}$ appearing in the definition of the sample variance seems awkward. As we shall see now, the natural choice of $\frac{1}{n}$ would lead to a biased estimator, while the slightly bizarre $\frac{1}{n-1}$ leads to an unbiased estimator. Lets consider a slightly more general estimator of the form

$$d(x) = \frac{1}{c}\sum_{i=1}^n (X_i - \overline{X}_n)^2,$$

where c is a constant to be determined, and determine the bias associated to it.
We start by rewriting the sum above as

$$\begin{aligned}
\sum_{i=1}^n (X_i - \overline{X}_n)^2 &= \sum_{i=1}^n \left((X_i - \mu) + (\mu - \overline{X}_n)\right)^2 \\
&= \sum_{i=1}^n \left((X_i - \mu)^2 - 2(X_i - \mu)(\overline{X}_n - \mu) + (\overline{X}_n - \mu)^2\right) \\
&= \sum_{i=1}^n (X_i - \mu)^2 - 2(\overline{X}_n - \mu)\sum_{i=1}^n (X_i - \mu) + \sum_{i=1}^n (\overline{X}_n - \mu)^2
\end{aligned}$$

But $\sum_{i=1}^n X_i = n\overline{X}_n$ and $\sum_{i=1}^n \mu = n\mu$, and the summand in the third term is independent of i yielding

$$\begin{aligned}
&= \sum_{i=1}^n (X_i - \mu)^2 - 2n(\overline{X}_n - \mu)^2 + n(\overline{X}_n - \mu)^2 \\
&= \sum_{i=1}^n (X_i - \mu)^2 - n(\overline{X}_n - \mu)^2.
\end{aligned}$$

Therefore

$$\begin{aligned} E\left[\frac{1}{c}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right] &= \frac{1}{c}E\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}_n-\mu)^2\right] \\ &= \frac{1}{c}\sum_{i=1}^{n}E\left[(X_i-\mu)^2\right] - nE\left[(\overline{X}_n-\mu)^2\right] \\ &= \frac{1}{c}\sum_{i=1}^{n}\text{Var}\left[X_i\right] - \frac{n}{c}\text{Var}\left[\overline{X}_n\right] \\ &= \frac{1}{c}n\sigma^2 - \frac{n}{c}\frac{\sigma^2}{n} \\ &= \sigma^2\,\frac{n-1}{c}. \end{aligned}$$

Recalling the definition of bias, for the estimator $\frac{1}{c}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2$ to be unbiased, its expected value must be equal to σ^2 (the quantity that we are trying to estimate). Hence, it is clear that by choosing $c=n-1$ we obtain the desired unbiased estimator, leading to the known form of the sample variance. If we instead choose $c=n$ to stick to the *natural choice*, we would obtain a biased estimator with bias given by

$$b(\theta) = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X}_n)^2\right] - \sigma^2 = \sigma^2\,\frac{n-1}{n} - \sigma^2 = -\frac{\sigma^2}{n}.$$

Since n and σ^2 are always positive, the estimator above will produce estimates that consistently underestimate variance of the distribution. However, it is also clear from the expression of the bias, that as the sample size n gets lager, the magnitude of the bias will decrease. This leads to the following concept.

Definition 11.3. We say that an estimator $d(X_1,\ldots,X_n)$ for the quantity $g(\theta)$ is a ***consistent estimator*** if

$$\lim_{n\to\infty} b_d\left(g(\theta)\right) = 0.$$

Hence, a biased estimator need not be a hopeless tool. As long as the estimator is consistent, it is possible to reduce the impact of the bias by considering samples that are large enough.

11.2 Exercises

1. Consider a discrete random variable X following a Bernoulli distribution with density

$$f(x|\theta) = \theta^x(1-\theta)^{1-x},$$

where θ represents the probability of a success and x is either 0 or 1 depending whether the attempt failed or succeeded. Consider an experiment where n attempts are carried out and s represents the number of successes obtained.

(a) Prove that the success ratio, defined as

$$d(X_1,\ldots,X_n) = \frac{s}{n},$$

where X_i is equal to either 0 or 1 depending on whether the $i-$th attempt is a success or a failure, is an unbiased estimator for θ. [Hint: Note that $s=\sum_{i=1}^{n}X_i$.]

(b) In problem 2 from Lecture 10 you computed the minimax estimator for the parameter θ in a **binomial** experiment with n attempts. Your computations there should have yielded

$$\widehat{\theta} = \frac{s}{4n},$$

where s is the number of successes obtained in n attempts. Imagine that we use this estimator to approximate the value of the parameter θ for n **Bernoulli** experiments. Show that this estimator is biased. Is it consistent?

2. Given the likelihood function

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

and a random sample $X_1, \ldots, X_n$ from this distribution. Show that the estimator

$$d(X) = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

is an unbiased estimator for θ.

3. Given $f(x|\theta) = 1/\theta$ for $x \in [0, \theta]$, determine the value of a constant c such that $d(x) = cx$ is an unbiased estimator of θ.

4. Given a Gamma random variable following the distribution

$$f(x|\theta) = \frac{x^{\theta-1}e^{-x}}{\Gamma(\theta)} \quad \text{for} \quad x > 0 \quad \text{and } \theta > 0.$$

(a) Find a value of c such that $d(x) = cx$ is an unbiased estimator for θ.

(b) Determine whether it is possible to find a value of c such that $d(x) = cx^2$ is an unbiased estimator of θ.

5. Let $X_1, \ldots, X_n$ be IID with mean $\mu := E[X]$ and consider constants $a_1, \ldots, a_n$. Determine the restrictions that are needed on $a_1, \ldots, a_n$ so that

$$d(X) = \sum_{i=1}^{n} a_i X_i$$

is an unbiased estimator of μ.

Lecture 12

Estimation II

Contents

12.1 Bias and mean squared error

Note that the fact that an estimator is unbiased *does not imply that there is no error associated to the estimator.* Instead, it implies that when using the estimator over a large number of experiments, the aggregated average of the error will be zero. An unbiased estimator will, *on average*, provide the correct answer. When making estimations based on measurements, there will be always some error associated to the estimate. Part of the error stems from the fact that data measurements are realizations of random variables and therefore there is always some degree of variability in the prediction, but another part can potentially come from a flaw in the estimator. The bias refers precisely to a systematic deviation of an estimate that is independent from the random variable: notice that the expectation in the definition of bias *cancels out* the influence of the random variable.

The splitting of the contributions coming from the data variability and the bias of an estimator to the error can be seen cleanly when we consider the ***mean squared error*** (i.e. the risk function when using square loss). Consider that $d(X)$ is an estimator for a certain function of the parameter $g(\theta)$. In this case we have that

$$\begin{aligned}
\text{MSE} &= \mathcal{R}(g(\theta), d) \\
&= E\left[\mathcal{L}(d, \theta)\right] \\
&= E\left[(d(X) - g(\theta))^2\right] \\
&= E\left[\left((d(X) - E[d(X)]) + (E[d(X)] - g(\theta)\right)^2\right] \\
&= E\left[(d(X) - E[d(X)])^2\right] \; + \; 2E\big[\big(d(X) - E[d(X)]\big)\big(E[d(X)] - g(\theta)\big)\big] \; + \; E\left[(E[d(X)] - g(\theta))^2\right]
\end{aligned}$$

However $E\left[(d(X) - E[d(X)])^2\right] = \text{Var}[d(X)]$ and $E[d(X)] - g(\theta) = b_d(g(\theta))$ is independent of X, hence

$$= \text{Var}[d(X)] + 2b_d(\theta) \underbrace{E\left[\big(d(X) - E[d(X)]\big)\right]}_{=0} + b_d(\theta)^2.$$

Therefore it follows that

$$\text{MSE} = \text{Var}[d(X)] + b_d(\theta)^2. \tag{12.1}$$

Thus, the mean square error is comprised of a contribution due to the variability of the data measurements, $\text{Var}[d(X)]$, and a contribution due to the bias of the estimator.
Since the only source of error for an unbiased estimator is the variance $\text{Var}[d(X)]$, it is natural to ask if there is a limit to how small the variance on the left hand side of the inequality can be. Given that $d(X)$ is a function of the data after all, the variance is tied to the data itself; one would expect that the particular properties of the random variables X and Θ should somehow play a role in determining how small can the variance be. We will explore this question in the following section.

12.2 The *information inequality*

In this section we will show how the variance of an unbiased estimator for a function $g(\theta)$ is controlled from below by the sensitivity of g to changes in θ, as well as (inversely) by the sensitivity of the likelihood function to changes in θ. In order to show this important result we first introduce some notation. We remind the reader that $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a vector of data observations sample from IID random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$.

Definition 12.1. Let $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$ be a likelihood function that is differentiable with respect to θ and that is different from zero except possibly at a set of measure zero[1]. We will define the following three related functions

$$\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) \qquad \textbf{\textit{Log-likelihood function}} \tag{12.2}$$

$$\partial_\theta \log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) \qquad \textbf{\textit{Score function}} \tag{12.3}$$

$$I_n(\theta) := \text{Var}\left[\partial_\theta \log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right] \qquad \textbf{\textit{Fisher information}} \tag{12.4}$$

The logarithms appearing above are natural logarithms, while the expected value in the definition of Fisher information is taken with respect to $\boldsymbol{X}$. The subscript n in the definition of the Fisher information emphasizes the fact that the the vector $\boldsymbol{X}$ contains the information of n independent observations.
Let us start by proving the following property

Proposition 12.1. *The score function has mean zero.*

Proof. Since the likelihood function is a PDF, the following equality holds

$$1 = \int_{\mathbb{R}^n} f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x}.$$

Hence, differentiating both sides with respect to θ we have

$$\begin{aligned} 0 &= \int_{\mathbb{R}^n} \partial_\theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \frac{\partial_\theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)}{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)} f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right) f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x} \\ &= E\left[\partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right)\right]. \end{aligned}$$

[1] If you are unfamiliar with the notion of a set of measure zero you can picture it as a set consisting of only isolated points in $\mathbb{R}^1$ (i.e. a set of length zero), or isolated points and isolated lines in $\mathbb{R}^2$ (i.e. a set or area zero), or a set consisting only of isolated points, isolated lines and isolated surfaces in $\mathbb{R}^3$ (a set of volume zero), etc.

Above we have used the fact that the likelihood function was assumed to be different from zero except possibly in a set of measure zero. □

We will use the result above in combination with the following fact about random variables with mean zero.

Proposition 12.2. *Let X and Y be random variables with mean μ_X and μ_Y respectively, and let $\mu_Y = 0$. Then*

$$Cov(X, Y) = E[XY]$$

Proof. The proof follows from the definition of covariance

$$\text{Cov}(X, Y) = E[(X - \mu_x))(Y - \mu_Y)] = E[(X - \mu_x))Y] = E[XY - \mu_X Y] = E[XY] - \mu_x E[Y] = E[XY].$$

□

We now have all the tools that we need to prove the following result known as the ***Cramér-Rao bound*** or or the ***information inequality*** regarding how the interaction between the parameter θ the quantity the we are trying to estimate $g(\theta)$ and the random variable X determine how small can Var $[d(X)]$ be, and therefore how good can we hope for an unbiased estimator to be.

Theorem 12.1. *Let X and Θ be random variables, and let the likelihood function $f_{X|\Theta}(x, \theta)$ be differentiable with respect to θ and non zero except possible for a set of measure zero. Consider a differentiable function $g(\theta)$ and an unbiased estimator for this function $d_g(\boldsymbol{X})$. Then*

$$\frac{(\partial_\theta g(\theta))^2}{I_n(\theta)} \leq Var[d_g(\boldsymbol{X})]. \tag{12.5}$$

Proof. Since $d_g(\boldsymbol{X})$ is an unbiased estimator for $g(\theta)$ we have that

$$g(\theta) = E[d_g(\boldsymbol{X})] = \int_{\mathbb{R}^n} d_g(\boldsymbol{X}) f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x}.$$

Hence, we can differentiate both sides with respect to θ to obtain

$$\begin{aligned}
\partial_\theta g(\theta) &= \int_{\mathbb{R}^n} d_g(\boldsymbol{X}) \partial_\theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} d_g(\boldsymbol{X}) \frac{\partial_\theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)}{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)} f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} d_g(\boldsymbol{X}) \partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right) f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta) d\boldsymbol{x} \\
&= E\left[d_g(\boldsymbol{X}) \partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right)\right].
\end{aligned}$$

However in Proposition 12.1 we showed that the expected value of the score function $\partial_\theta \left(\log f_{X|\Theta}(x, \theta)\right)$ is zero, hence we can apply Proposition 12.2 to the expected value above leading to

$$\partial_\theta g(\theta) = \text{Cov}\left(d_g(\boldsymbol{X}), \partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right)\right). \tag{12.6}$$

We now recall that the correation is bounded above by one, hence

$$\text{Corr}^2(X, Y) = \frac{\text{Cov}^2(X, Y)}{\text{Var}[X]\text{Var}[Y]} \leq 1.$$

Combining this result with the equation (12.6) above, we obtain

$$(\partial_\theta g(\theta))^2 = \text{Cov}^2\left(d_g(\boldsymbol{X}), \partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right)\right) \leq \text{Var}[d_g(\boldsymbol{X})] \,\text{Var}\left[\partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\right)\right]$$

From which we arrive at

$$\frac{(\partial_\theta g(\theta))^2}{\text{Var}\left[\partial_\theta \left(\log f_{\boldsymbol{X}|\Theta}(\boldsymbol{x},\theta)\right)\right]} = \frac{(\partial_\theta g(\theta))^2}{I_n(\theta)} \leq \text{Var}\left[d_g(\boldsymbol{X})\right].$$

The inequality above is known as the ***information inequality*** or the ***Cramér-Rao bound***. □

From the observation that the mean squared error of an estimator has a component due to its bias and one due to its variance, we could conclude that an ideal estimator would be unbiased and have a very small variability—zero if possible. This leads to the notion of an ideal estimator: one that is unbiased and whose variance is smaller than that of any other.

Definition 12.2. We say that an estimator $\widehat{d}_g(X)$ for the function $g(\theta)$ is a ***uniformly minimum variance unbiased estimator*** or ***UMVUE*** if, for any other unbiased estimator $d_g(X)$ it holds that

$$\text{Var}\left[\widehat{d}_g(X)\right] \leq \text{Var}\left[d_g(X)\right].$$

An UMVUE can be regarded as "the gold standard" against which all other estimators are measured. Naturally, the smallest possible variance of an estimator is given by the Crámer- Rao bound derived above. Hence, an UMVUE will have variance equal to the left hand side of (12.5). If $\widehat{d}_g(X)$ is an UMVUE, the ***efficiency of an estimator*** $d_g(X)$ is given by

$$e(d_g(X)) = \frac{\text{Var}\left[\widehat{d}_g(X)\right]}{\text{Var}\left[d_g(X)\right]}.$$

The efficiency takes values between 0 and 1. We say that a particular estimator $d_g(X)$ is an ***efficient estimator*** whenever the variance $\text{Var}[d_g(\boldsymbol{X})]$ is equal to the lower bound given by the information inequality (12.5) and therefore has efficiency equal to 1.

12.3 Examples

Example 1: Measurements from n independent identically distributed random variables. Consider n IID measurements $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a random variable with likelihood function $f_{X|\theta}(X|\theta)$. Since the measurements are independent, the likelihood function for $\boldsymbol{X}$ given θ can be expressed as a product of the marginal likelihood functions

$$f_{\boldsymbol{X}|\theta}(\boldsymbol{X}|\theta) = f_{\boldsymbol{X}|\theta}(X_1, \ldots, X_n|\theta) = f_{X_1|\theta}(X_1|\theta) \times \cdots \times f_{X_1|\theta}(X_n|\theta) = \prod_{i=1}^{n} f_{X_i|\theta}(X_i|\theta).$$

Therefore we can write the log-likelihood function as

$$\log\left(f_{\boldsymbol{X}|\theta}(\boldsymbol{X}|\theta)\right) = \log\left(\prod_{i=1}^{n} f_{X_i|\theta}(X_i|\theta)\right) = \log\left(\sum_{i=1}^{n} f_{X_i|\theta}(X_i|\theta)\right),$$

and the score function as

$$\partial_\theta \log\left(f_{\boldsymbol{X}|\theta}(\boldsymbol{X}|\theta)\right) = \sum_{i=1}^{n} \partial_\theta \log f_{X_i|\theta}(X_i|\theta).$$

Hence the information from n IID measurements is given by

$$I_n(\theta) = \text{Var}\left[\partial_\theta \log\left(f_{\boldsymbol{X}|\theta}(\boldsymbol{X}|\theta)\right)\right] = \text{Var}\left[\sum_{i=1}^{n} \partial_\theta \log f_{X_i|\theta}(X_i|\theta)\right] = \sum_{i=1}^{n} \text{Var}\left[\partial_\theta \log f_{X_i|\theta}(X_i|\theta)\right] = nI(\theta), \quad (12.7)$$

where we have used the fact that the samples are all independent and have common variance; $I(\theta)$ refers to the information carried *by a single measurement.*

Example 2: Independent Bernoulli trials. We consider a Bernoulii random variable with pdf given by

$$f_{X|\Theta}(x|\theta) = \theta^x(1-\theta)^{1-x},$$

where $x = 1$ represents a success and $x = 0$ a failure, and the probability of success given by θ. If we are trying to estimate the value of θ based on n observations we have that

$$\partial_\theta f_{X|\Theta}(x|\theta) = \partial_\theta \log\left(\theta^x(1-\theta)^{1-x}\right) = \partial_\theta\left(x\log\theta + (1-x)\log(1-\theta)\right) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)},$$

and therefore, recalling that for a Bernoulli random variable $\text{Var}[X] = \theta(1-\theta)$ we conclude that

$$I(\theta) = \text{Var}\left[\partial_\theta \log\left(\theta^X(1-\theta)^{1-X}\right)\right] = \text{Var}\left[\frac{X-\theta}{\theta(1-\theta)}\right] = \frac{\text{Var}[X]}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)} = \frac{1}{\text{Var}[X]}.$$

Hence, for n independent measurements, equation (12.7) leads to

$$I_n(\theta) = \frac{n}{\theta(1-\theta)} = \frac{n}{\text{Var}[X]}.$$

Then, from the information inequality we conclude that if we are trying to estimate θ using an unbiased estimator $d_g(X_1,\ldots,X_n)$ we have that the lower bound established by the information inequality is

$$\text{Var}[d_g(X1,\ldots,X_n)] \geq \frac{1}{I_n(\theta)} = \frac{\text{Var}[X]}{n},$$

where we have used that for the estimation of θ the function g appearing in (12.7) is the identity function $g(\theta) = \theta$. If we now use the sample mean as the estimator for θ we have

$$d_g(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and then

$$\text{Var}[d_g(X1,\ldots,X_n)] = \frac{\text{Var}[X]}{n}.$$

Hence, the sample mean has in fact the minimum variance allowed by the information inequality and is therefore a uniformly minimum variance unbiased estimator.

12.4 Exercises

1. Show that the sample mean $d(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is both unbiased and efficient as an estimator for the parameter θ for the Poisson density

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!} \quad x = 0, 1, 2, 3, \ldots$$

2. Consider Problem 5 from Lecture 11. In addition to the constraints that you determined in the aforementioned problem, determine

(a) What is the best set of coefficients a_i if $d(\boldsymbol{X}) = \sum_{i=1}^{n} a_i X_i$ is to be an unbiased estimator of $\mu := E[X]$ with minimum variance?

(b) How should the a_i's be chosen if the measurements X_i are independent variables with the same unknown mean μ but different known variances σ_i^2?

3. Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a random variable X with finite mean $E[X]$.

(a) Show that the estimator

$$d(\boldsymbol{X}) = \frac{X_1 + X_2}{2}$$

is an unbiased estimator for $E[X]$.

(b) Calculate the efficiency of $d(\boldsymbol{X})$ as defined in the previous point if X possesses the density

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}.$$

4. Find the lower bound (as given by the information inequality) for the variance for unbiased estimators of the parameter θ for a Cauchy random variable with density

$$f(x|\theta) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)}.$$

5. The information inequality in Theorem 12.1 *does not apply* to the uniform random variable $f(x|\theta) = 1/\theta$ for $x \in [0, \theta]$. Can you point the hypotheses of the theorem that are not satisfied by $f(x|\theta)$?

Lecture 13

Estimation III

Contents

13.1 Method of Moments

Consider a random variable X with a probability density function $f_X(x, \boldsymbol{\theta})$ depending on one (or multiple) parameters $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)$. Recall the definition of the k-th ***moment*** of the random variable X

$$m_k(\boldsymbol{\theta}) := E\left[X^k\right] = \int_{-\infty}^{\infty} x^k f_X(x, \boldsymbol{\theta}) dx. \tag{13.1}$$

The explicit dependence of m_k on the parameters $\boldsymbol{\theta}$ is a result of the fact that the PDF is a function of the parameters $\boldsymbol{\theta}$ and the integration is carried over with respect to x only.
Now consider a random, independent sample $\boldsymbol{X} := (X_1, \ldots, X_n)$ from the same distribution $f_X(x, \boldsymbol{\theta})$. By analogy, we define the k-th ***sample moment*** as

$$\overline{m}_k := \frac{1}{n} \sum_{i=1}^{n} X_i^k. \tag{13.2}$$

The (strong) law of large numbers says that as the sample size grows, the k-th sample moment converges to the k-th moment of a distribution

$$\lim_{n \to \infty} \overline{m}_k = m_k(\boldsymbol{\theta}).$$

The ***method of moments*** takes advantage of this fact to estimate the values of the parameters appearing in the PDF. The idea is that, for a particular sample of the random variables $X_1, \ldots, X_n$ the function $m_k(\boldsymbol{\theta})$ should be approximately equal to the number $\overline{m}_k$. Thus, the method of moments estimate for $\boldsymbol{\theta}$, denoted as $\widehat{\boldsymbol{\theta}}$, is determined by the equation

$$m_k(\widehat{\boldsymbol{\theta}}) = \overline{m}_k.$$

Hence, the estimator is given by

$$\widehat{\boldsymbol{\theta}} = m_k^{-1}(\overline{m}_k),$$

whenever the inverse function m_k^{-1} exists. Note that the law of large numbers guarantees that the estimator produced by the method of moments is consistent, since

$$\lim_{n\to\infty} \widehat{\boldsymbol{\theta}} = \lim_{n\to\infty} m_k^{-1}(\overline{m}_k) = m_k^{-1}(\lim_{n\to\infty} \overline{m}_k) = m_k^{-1}(m_k(\boldsymbol{\theta})) = \boldsymbol{\theta}.$$

In the sequence of equalities above, we assumed that the moments are continuous with respect to the parameters. We can condense the method into an algorithm as follows.

Step 1. If the PDF has d parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ that we want to estimate, we then use equation (13.1) to compute the first d moments of the distribution

$$m_1(\theta_1, \ldots, \theta_d), \quad m_2(\theta_1, \ldots, \theta_d), \quad \ldots \quad , m_d(\theta_1, \ldots, \theta_d).$$

This yields d expressions for the first d moments as functions of the d parameters $\theta_1, \ldots, \theta_d$.
Step 2. We then use the data observations and equation (13.2) to compute the first d ***sample moments***

$$\overline{m}_1, \quad \overline{m}_2, \quad \ldots \quad , \overline{m}_d.$$

These are d random variables whose particular value will change if we change the sample, but for any fixed sample they are d fixed numbers.
Step 3. The law of large numbers states that for a finite sample size n we have that

$$m_1(\theta_1, \ldots, \theta_d) \approx \overline{m}_1, \quad m_2(\theta_1, \ldots, \theta_d) \approx \overline{m}_2, \quad \ldots \quad , m_d(\theta_1, \ldots, \theta_d) \approx \overline{m}_d.$$

Hence, we will define our estimates for $(\theta_1, \ldots, \theta_d)$ to be the numbers $(\widehat{\theta_1}, \ldots, \widehat{\theta_d})$ that satisfy the system of d equations

$$m_1(\widehat{\theta_1}, \ldots, \widehat{\theta_d}) = \overline{m}_1, \quad m_2(\widehat{\theta_1}, \ldots, \widehat{\theta_d}) = \overline{m}_2, \quad \ldots \quad , m_d(\widehat{\theta_1}, \ldots, \widehat{\theta_d}) = \overline{m}_d.$$

Remarks:

1. The equations arising from the method of moments may be non-linear or may not even be solvable.
2. We are by no means obliged to always use the moments in order. For example in a distribution with a single parameter we do not necessarily have to use the equation for the first moment, we could use the second, third, etc. Sometimes the equations arising from higher order moments may be easier to solve or give rise to estimators with less variability.

13.2 Examples

Example 1. Consider a simple random sample $X_1, \ldots, X_n$ from a Pareto random variable with PDF

$$f_X(x, \beta) = \frac{\beta}{x^{\beta+1}}, \qquad \text{for } x \in (1, \infty).$$

If we wish to estimate the parameter β using the method of moments, we need to compute the first moment from the distribution, which for this particular type or random variable can be shown to be

$$m_1(\beta) = E[X] = \frac{\beta}{\beta - 1}.$$

To obtain the method of moments estimator $\widehat{\beta}$ we then set this moment equal to the first sample moment

$$\frac{\widehat{\beta}}{\widehat{\beta} - 1} = \overline{m}_1.$$

Solving this equation for $\widehat{\beta}$ yields

$$\widehat{\beta} = \frac{\overline{m}_1}{\overline{m}_1 - 1}.$$

Equation (13.2) can be used to compute the first sample moment $\overline{m}_1$ from the data which, upon substitution in the relation above, will yield the desired estimate.

Example 2. Consider an exponential random variable with PDF

$$f_X(x, \lambda) = \lambda e^{-\lambda x} \quad \text{for } x \in [0, \infty) \quad \text{and } \lambda > 0.$$

To obtain the method of moments estimator for λ, we compute the first moment of the distribution

$$m_1(\lambda) = E[X] = \frac{1}{\lambda}.$$

The method of moments estimator $\widehat{\lambda}$ will then be the number for which the relation above holds true when we substitute $m_1(\lambda)$ for the sample moment $\overline{m}_1$, namely

$$\overline{m}_1 = \frac{1}{\widehat{\lambda}}.$$

Solving for $\widehat{\lambda}$ above leads to the estimate

$$\widehat{\lambda} = \frac{1}{\overline{m}_1}.$$

Example 3. The PDF for a Gamma random variable depends on two parameters and is given by

$$f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \alpha^{-\alpha-1} e^{-\beta/x} \quad \text{for } x \in [0, \infty),\ \alpha > 0, \text{ and } \beta > 0,$$

where $\Gamma(\cdot)$ is the Gamma function. If we want to estimate the values of α and β based on n independent measurements $X_1, \ldots, X_n$ drawn from Gamma we will need to use the first two moments of the distribution. For a Gamma random variable it can be proven that

$$E[X] = \alpha/\beta \qquad \text{and} \qquad \text{Var}[X] = \alpha/\beta^2.$$

Since for any random variable $\text{Var}[X] = E\left[X^2\right] - (E[X])^2$ we can then use the equations above to write

$$E[X] = \alpha/\beta \qquad \text{and} \qquad E\left[X^2\right] = \alpha/\beta^2 + (\alpha/\beta)^2.$$

Then, letting $\overline{m}_1$ and $\overline{m}_2$ be the first and second sample moments, the method of moments estimators $\widehat{\alpha}$ and $\widehat{\beta}$ will be the solutions to the equations

$$\overline{m}_1 = \widehat{\alpha}/\widehat{\beta} \qquad \text{and} \qquad \overline{m}_2 = \widehat{\alpha}/\widehat{\beta}^2 + (\widehat{\alpha}/\widehat{\beta})^2.$$

Solving this system in terms of $\widehat{\alpha}$ and $\widehat{\beta}$ leads to

$$\widehat{\alpha} = \frac{\overline{m}_1^2}{\overline{m}_2 - \overline{m}_1^2} \qquad \text{and} \qquad \widehat{\beta} = \frac{\overline{m}_1}{\overline{m}_2 - \overline{m}_1^2}.$$

Hence, once the measurements are made, it is enought to compute the first two sample moments and substitute the values in the expressions above to obtain the method of moments estimators.

13.3 Exercises

1. Show that the k-th ***sample moment*** based on an IID sample of size n of a random variable X as defined in equation (13.2) is an unbiased estimator of the k-th theoretical moment of X.

2. Consider a uniform density function given by

$$f_X(x, \theta) = 1/\theta \quad \text{for } x \in [0, \theta],$$

   (a) Use the method of moments to estimate the value of the parameter θ based on a random sample of size n.

   (b) Use a random number generator (you can use R, matlab, numpy, etc.) To generate a sample of 20 numbers uniformly distributed between 0 and 1/2. Use this sample in combination with the result from the previous point to obtain an estimate of θ. [Note that since you are sampling uniformly between 0 and 1/2 the exact value for theta in this case is $\theta = 1/2$].

3. Consider a simple random sample $X_1, \ldots, X_n$ from a random variable X uniformly distributed in the interval $[a, b]$ where **both a and b are unknown parameters**. Build the method of moments estimator for a and b.

4. Sometimes the calculations using a higher moment are easier to handle than those arising from a lower moment (we say that the moment m^k is higher than the moment $m^{\widetilde{k}}$ if $k > \widetilde{k}$). Consider the random variable X with PDF given by

$$f_X(x, \theta) = \frac{x}{\theta} e^{-\frac{x^2}{2\theta}} \quad \text{for} \quad x > 0,\ \theta > 0.$$

   (a) Derive the method of moments estimator of θ based on a sample of size n **using the second moment of the distribution**.

   (b) Now *try* to derive the method of moments estimator of θ based on a sample of size n using the first moment of the distribution. Is it possible to arrive at a closed-form expression for this moment?

5. The method of moments fails to produce an estimator for the parameter θ of a Cauchy random variable with PDF given by

$$f_X(x, \theta) = \frac{1}{\pi\left(1 + (x - \theta)^2\right)}.$$

   Try to follow the steps to build such an estimator and explain what point of the process breaks down for this particular random variable.

Lecture 14

Estimation IV

Contents

We recall here that the ***likelihood function*** is simply the probability density function $f_X(x, \theta)$ when θ is considered to be a particular realization from a random variable Θ. In this context the notation

$$L(\theta|x) \equiv f_X(x, \theta)$$

is often preferred, and we will use it for this lecture. We remark that even though $f_X(x, \theta)$ is a probability distribution *for the random variable* X the likelihood function is not necessarily a PDF for the random variable Θ. In fact, *when integrated with respect to* θ the PDF may not even integrate to one over the entire parameter space.

14.1 Maximum likelihood estimation

For a parameter-dependent density function $f_X(x, \theta) = L(\theta|x)$ with unknown θ, the principle of ***maximum likelihood*** states that, given a sample $X_1, \ldots, X_n$ drawn from $f_X(x, \theta)$, the value of θ should be approximated by the number $\widehat{\theta}$ that would make our particular sample the most likely to be measured. This is in accordance with the common-sense notion that a collection of random observations of X will most likely contain values of X that happen very commonly. After all, how probable it is to observe a rare occurrence? Guided by this reasoning, the ***maximum likelihood estimator*** for θ is defined as the number $\widehat{\theta}$ such that

$$L(\widehat{\theta}|x) > L(\theta|x) \qquad \text{for all } \theta \in \Theta.$$

We can write the statement above in the alternate form

$$\widehat{\theta}(x) = \arg\max_{\theta \in \Theta} L(\theta|x). \tag{14.1}$$

This definition implicitly assumes that the likely function has a unique global maximum. However, not every function has a global maximum, and even if it does, locating it may present significant computational challenges. Recalling that the natural logarithm is a strictly monotonically increasing function, maximizing

$L(\theta|x)$ is equivalent to maximizing the log-likelihood function $\log[L(\theta|x)]$ which is in most instances much simpler to deal with.
The method is pretty straightforward and can be summarized as follows.

Step 1. Starting from the likelihood function $L(\theta|x)$, compute the log-likelihood function $\log[L(\theta|x)]$.

Step 2. Find the critical points $\theta_{c1}, \theta_{c2}, \ldots$, etc. (there could possibly be more than one) of the log-likelihood function by solving for θ in the equation

$$\partial_\theta \log[L(\theta|x)] = 0.$$

Remember that if θ is defined in an interval with at least a finite endpoint, for instance if $\theta \in [a, b]$ with both a and b finite, or $\theta \in [0, \infty)$, then the endpoints of the interval are also critical points.

Step 3. Verify which of the critical points is a maximizer of the log likelihood function. You can do this through the first or second derivative tests to determine local maxima, and through evaluation and direct comparison of $L(\theta_{ci}|x)$ if there is more than one local maximum.

Step 4. Set the maximum likelihood estimator $\widehat{\theta}$ to be the global maximizeer that you determined in the previous step.

14.2 Examples

Example 1. The PDF for a Bernoulli random variable is

$$f_X(x, \theta) = \theta^x(1-\theta)^{1-x} = L(\theta|\boldsymbol{X}),$$

where $x = 0$ denotes a failure and $x = 1$ denotes a success, and θ is the probability of success. If the parameter θ is unknown and we wish to estimate it based on n independent measurements, the joint PDF for $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

$$L(\theta|\boldsymbol{X}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i^n x_i}(1-\theta)^{\sum_i^n (1-x_i)} = \theta^s(1-\theta)^{n-s},$$

where we have used the fact that the trials are independent to express the joint PDF as a product of the marginals, and $s := \sum_i^n x_i$ is the number of successes obtained after n attempts[1]. We can then compute the log-likelihood function

$$\ln[L(\theta|\boldsymbol{X})] = s\ln(\theta) + (n-s)\ln(1-\theta),$$

which is defined for $\theta \neq 0$ and $\theta \neq 1$ (these two cases would yield trivial cases when x is not a random variable, since success or failure would be certain). From here we can compute the derivative with respect to θ

$$\partial_\theta \ln[L(\theta|\boldsymbol{X})] = \frac{s}{\theta} - \frac{n-s}{1-\theta} = \frac{s-n\theta}{\theta(1-\theta)}$$

Setting the relation above equal to zero and solving for θ leads to the critical point

$$\theta_c = \frac{s}{n},$$

[1]Note that since $x_i = 0$ if the $i-$th attempt failed and $x_i = 1$ if the $i-$th attempt succeeded, the sum $\sum_i^n x_i$ is in fact simply counting the number of successes, s.

which is known as the *success rate*. To verify if this critical point corresponds to a global maximum, we note that the denominator $\theta(1-\theta)$ is always positive. On the other hand, if $\theta < s/n$ then the numerator is positive, while if $\theta > s/n$ the denominator is negative. This implies that the function has a maximum at the value $\theta_c = s/n$. Hence the estimator

$$\widehat{\theta} = \frac{s}{n} \tag{14.2}$$

is the maximum likelihood estimator for the Bernoulli parameter θ.

Example 2. We will now consider data $\boldsymbol{X} = (X_1, \ldots, X_n)$ sampled independently from a normal distribution with mean μ and variance σ^2. Since the samples are taken independently, we can write the joint likelihood function as a product of the marginal distributions:

$$L(\mu, \sigma^2|\boldsymbol{X}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}.$$

If we want to compute the maximum likelihood estimators for the mean and variance of the distribution, it is easier to work with the log-likelihood function

$$\ln\left[L(\mu, \sigma^2|\boldsymbol{X})\right] = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2.$$

Computing the partial derivatives of this function with respect to μ and σ^2 yields[2]

$$\partial_\mu \ln\left[L(\mu, \sigma^2|\boldsymbol{X})\right] = \frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu) \qquad \text{and} \qquad \partial_{\sigma^2} \ln\left[L(\mu, \sigma^2|\boldsymbol{X})\right] = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (x_i - \mu)^2.$$

The critical points $\widehat{\mu}$ and $\widehat{\sigma}^2$ are located by setting the two derivatives equal to zero and solving the corresponding system of equations, which yields

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i = \overline{x}_n \qquad \text{and} \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x}_n)^2. \tag{14.3}$$

We then need to verify if these critical points are indeed maximizers of the likelihood. This can be achieved through the second derivative test, so we compute the corresponding derivatives and evaluate them at $\widehat{\mu}$ and $\widehat{\sigma}^2$, leading to

$$\partial_\mu^2 \ln\left[L(\widehat{\mu}, \widehat{\sigma}^2|\boldsymbol{X})\right] = -\frac{n}{\widehat{\sigma}^2} < 0$$

and

$$\partial_{\sigma^2}^2 \ln\left[L(\widehat{\mu}, \widehat{\sigma}^2|\boldsymbol{X})\right] = \frac{n}{2\widehat{\sigma}^4} - \frac{1}{\widehat{\sigma}^6}\sum_{i=1}^n (x_i - \overline{x}_n)^2 = \frac{n}{2\widehat{\sigma}^4} - \frac{n\widehat{\sigma}^2}{\widehat{\sigma}^6} = -\frac{n}{2\widehat{\sigma}^6} < 0.$$

Hence, the critical points in equation (14.3) correspond to a maximum of the likelihood function and are the maximum likelihood estimators for the mean and variance. Note that the maximum likelihood estimator for the variance is actually a biased estimator.

Example 3. ***Linear regression.*** We now consider two sets of data observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ of pairs of data points that we consider to be related. If a scatter plot of the data suggests a linear relation between the two (as depicted in Figure 14.1), a reasonable model would be to consider them to be linearly related and attribute the deviations from a clean straight line to a random source of error. In

[2]Note that we are using the variance σ^2 as the differentiation variable and **not** the standard deviation σ.

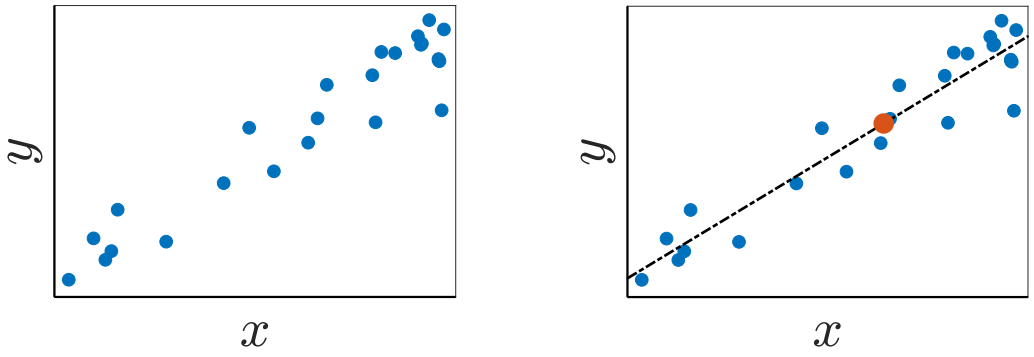

Figure 14.1: Left: The scatter plot of the data suggests a linear relation between the variables X and Y. The deviations out linear behavior are assumed to be due to normally distributed noise $\sim N(0, \sigma^2)$ with unknown variance σ^2. Right: Under the assumption of normally distributed noise, the maximum likelihood coefficients for the linear regression minimize the sum of the squared vertical distances of the data to the theoretical line (dotted). The point with coordinates given by the sample means $(\overline{x}_n, \overline{y}_n)$ (indicated by the red dot in the plot) is located on top of the regression line.

absence of any information about the possible origins of the deviations, the error term can be considered to be a random variable normally distributed with mean zero and unknown standard deviation σ^2. Mathematically, this leads to the relation

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

where α and β are parameters that need to be determined and ϵ_i is an independent realization of the random error $\epsilon \sim N(0, \sigma^2)$.
Our model then includes three unknown parameters that we will estimate using the maximum likelihood strategy. We have assumed the error $\boldsymbol{\epsilon} = (\epsilon_n, \ldots, \epsilon_n)$ to be normally distributed and independent, which leads to the likelihood function

$$L(\sigma^2|\boldsymbol{\epsilon}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon_1^2}{2\sigma^2}} \times \cdots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon_n^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2},$$

where in the final step we used our model to write $\epsilon_i = Y_i - \alpha - \beta X_i$. Since the last expression features the random variables X and Y along with the parameters α, β and σ^2, we will then denote the function above as

$$L(\alpha, \beta, \sigma^2|\boldsymbol{X}, \boldsymbol{Y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}.$$

As before, it is easier to work with the log likelihood function which in this case is

$$\ln\left[L(\alpha, \beta, \sigma^2|\boldsymbol{X}, \boldsymbol{Y})\right] = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

One first observation is that if we were to maximize the function above with respect to α and β alone (in which case the first term above would be constant), the problem would be equivalent to that of minimizing the term

$$SS(\alpha, \beta) := -\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = -\sum_{i=1}^n \epsilon_i^2.$$

Therefore, the maximum likelihood estimators for α and β will coincide with the values resulting from the ***least squares*** criterion, which minimizes the sum of squares of the difference between the data points Y_i and the values predicted by the regression.

We then compute the partial derivative of the score function with respect to α and impose the optimality condition to find the critical points $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\sigma}^2$

$$\partial_\alpha \ln\Big[L(\widehat{\alpha},\widehat{\beta},\widehat{\sigma}^2|\boldsymbol{X},\boldsymbol{Y})\Big] = \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}(y_i-\widehat{\alpha}-\widehat{\beta}x_i) = 0.$$

From this condition it follows that

$$\sum_{i=1}^{n}(y_i-\widehat{\alpha}-\widehat{\beta}x_i) = n\overline{y}_n - n\widehat{\alpha} - n\widehat{\beta}\,\overline{x}_n = 0.$$

Hence, the sample means $\overline{x}_n$ and $\overline{y}_n$ satisfy the regression equation

$$\overline{y}_n = \widehat{\alpha} + \widehat{\beta}\,\overline{x}_n, \tag{14.4}$$

and therefore will be located right along the regression line (this is illustrated in the right panel of Figure 14.1). We can verify that the critical point satisfying the equation above is indeed a maximum, since

$$\partial^2_\alpha \ln\Big[L(\widehat{\alpha},\widehat{\beta},\widehat{\sigma}^2|\boldsymbol{X},\boldsymbol{Y})\Big] = -\frac{n}{\widehat{\sigma}^2} < 0.$$

We now use the fact that $\widehat{\alpha} = \overline{y}_n + \widehat{\beta}\,\overline{x}_n$ to modify the score function as

$$\ln\big[L(\alpha,\beta,\sigma^2|\boldsymbol{X},\boldsymbol{Y})\big] = -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\overline{y}_n-\beta(x_i-\overline{x}_n))^2.$$

Note that if α and β are indeed the maximizers then the expression above is identical to the original score function. We then compute the partial derivative of the score function with respect to β and evaluate at the critical points to obtain

$$\partial_\beta \ln\Big[L(\widehat{\alpha},\widehat{\beta},\widehat{\sigma}^2|\boldsymbol{X},\boldsymbol{Y})\Big] = \frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}(y_i-\overline{y}_n-\widehat{\beta}(x_i-\overline{x}_n))(x_i-\overline{x}_n) = 0$$

Solving for $\widehat{\beta}$ in the equation above leads to

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(y_i-\overline{y}_n)(x_i-\overline{x}_n)}{\sum_{i=1}^{n}(x_i-\overline{x}_n)^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y}_n)(x_i-\overline{x}_n)}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}. \tag{14.5}$$

Once again, we can verify that $\widehat{\beta}$ is indeed a maximizer by computing

$$\partial^2_{\widehat{\beta}} \ln\Big[L(\widehat{\alpha},\widehat{\beta},\widehat{\sigma}^2|\boldsymbol{X},\boldsymbol{Y})\Big] = -\frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 < 0.$$

Finally, to determine the MLE for the variance we compute

$$\partial_{\sigma^2} 2\ln\Big[L(\widehat{\alpha},\widehat{\beta},\widehat{\sigma}^2|\boldsymbol{X},\boldsymbol{Y})\Big] = -\frac{n}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{n}(y_i-\widehat{\alpha}-\widehat{\beta}x_i)^2 = 0,$$

from which we conclude that

$$\widehat{\sigma}^2 = \frac{1}{}\sum_{i=1}^{n}(y_i-\widehat{\alpha}-\widehat{\beta}x_i)^2 = \frac{1}{}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 \tag{14.6}$$

where we denoted by $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ the value of y predicted by the regression line evaluated at $x = x_i$.

14.3 Exercises

1. Consider an exponential random variable $f(x|\theta) = \theta e^{-\theta x}, \quad x > 0$ and a sample of size n.

   (a) Find the maximum likelihood estimator of θ.

   (b) Find the maximum likelihood estimator for $E[X]$. How does this compare to the result of the previous problem?

2. Find the maximum likelihood estimator of θ based on a sample of size n for the random variable with density $f(x|\theta) = (1+\theta)x^{\theta}, \quad x \in [0,1], \quad \theta > 0$.

3. Consider a geometric random variable with density $f(x|\theta) = \theta(1-\theta)^x, \quad x = 0, 1, 2, 3, \ldots$ Geometric random variables model the number of attempts, given by x, required **before** obtaining the first success in a binomial process with probability of success θ.

   (a) Compute the maximum likelihood estimator of θ for this distribution.

   (b) Perform the following experiment. Using a six-faced dice, roll the die and record the number of attempts before a four appear on the dice. Repeat this experiment 20 times to obtain a sample $(x_1, \ldots, x_{20})$ where every x_i is the number of rolls required before success. Then use this information and the result from the previous part to obtain the MLE for θ.

4. Consider the PDF $f(x|\theta) = \theta^2 x e^{-\theta x}, \quad x \geq 0$.

   (a) Using the PDF,compute an expression for the mean and the variance of X.

   (b) Compute the maximum likelihood estimators for the mean and the variance.

Lecture 15

Estimation V

Contents

15.1 Bayesian estimation

Up to this point in our introduction to estimation, we have considered that the parameter θ that we are attempting to estimate has a fixed value. However, the Bayesian paradigm considers that the parameter might in fact be a realization from an independent random variable that varies according to the probability density function $\pi(\theta)$ known as the ***prior distribution***, or **prior** for short.

When considering two random variables X and Θ, we need to consider their ***joint probability density function*** which we will denote by $f_{X,\Theta}(x,\theta)$. According to Baye's rule, the joint PDF can be written in terms of the product between a conditional density and a marginal density in two different ways:

$$\overbrace{f_{X|\Theta}(x|\theta)}^{\text{Likelihood function}}\underbrace{\pi(\theta)}_{\text{prior}} = \overbrace{f_{X,\Theta}(x,\theta)}^{\text{Joint PDF}} = \overbrace{f_{\Theta|X}(\theta|x)}^{\text{Posterior}}\underbrace{f_X(x)}_{\text{Marginal density}} . \tag{15.1}$$

On the left hand side, we express the joint distribution as the product between the likelihood function (i.e. the conditional probability for X given an assumed or measured value of the parameter θ) and the prior density (i.e. the *assumed* marginal behavior of the random variable θ *before* making any data measurements). On the right hand side, we express the joint PDF as the product between the conditional density for the random variable θ conditioned to a given set of measurements of x (known as the ***posterior density***, it captures the behavior of θ *after* taking into account the measured information of x) and the marginal density of X.

The paradigm of ***Bayesian estimation*** is to consider that the particular realization of the random variable Θ that gave rise to certain measurements $\boldsymbol{X} = (X_1, \ldots, X_n)$ is the value $\widehat{\theta}$ that minimizes the mean risk

$r(\pi, d)$. Hence, to determine what would the Bayesian estimate $\widehat{\theta}$ be, we analize the mean risk as follows:

$$\begin{aligned}
r(\pi, d) &= E\left[R(\theta, d)\right] \\
&= \int_{\Theta} R(\theta, d)\pi(\theta)\, d\theta \\
&= \int_{\Theta} E\left[\mathcal{L}(\theta, d)\right]\pi(\theta)\, d\theta \\
&= \int_{\Theta}\left(\int_{\boldsymbol{X}} \mathcal{L}(\theta, d(\boldsymbol{x}))f(\boldsymbol{x}|\theta)\, d\boldsymbol{x}\right)\pi(\theta)\, d\theta \\
&= \int_{\Theta}\int_{\boldsymbol{X}} \mathcal{L}(\theta, d(\boldsymbol{x}))\underbrace{f(\boldsymbol{x}|\theta)\,\pi(\theta)}_{=\text{Joint PDF}}\, d\boldsymbol{x}\, d\theta \\
&= \int_{\Theta}\int_{\boldsymbol{X}} \mathcal{L}(\theta, d(\boldsymbol{x}))f_{\boldsymbol{X},\Theta}(\boldsymbol{x}, \theta)\, d\boldsymbol{x}\, d\theta \qquad \text{(Using 15.1)} \\
&= \int_{\Theta}\int_{\boldsymbol{X}} \mathcal{L}(\theta, d(\boldsymbol{x}))f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}\, d\theta \qquad \text{(Using 15.1)} \\
&= \int_{\boldsymbol{X}}\left(\int_{\Theta} \mathcal{L}(\theta, d(\boldsymbol{x}))f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta\right) f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}
\end{aligned}$$

Note that all the functions, $\mathcal{L}(\theta, d(\boldsymbol{x}))$, $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$, and $f_{\boldsymbol{X}}(\boldsymbol{x})$, involved in the integral above are non-negative. Moreover, we have not control over the random variables, hence we can not optimize neither $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$, nor $f_{\boldsymbol{X}}(\boldsymbol{x})$. Hence the mean risk will be minimized if we can choose $d(\boldsymbol{X})$ such that the innermost integral

$$\int_{\Theta} \mathcal{L}(\theta, d(\boldsymbol{x}))f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta \tag{15.2}$$

is minimized. This expression is in fact the definition of the ***conditional expectation*** of the loss. Hence, the Bayes estimator $d(\boldsymbol{x})$ should be chosen such that the conditional expectation of the loss is minimized. If we consider a squared loss function, then (15.2) becomes

$$\int_{\Theta} (\theta - d(\boldsymbol{x}))^2 f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta,$$

and hence the optimality condition is

$$\frac{\partial}{\partial d}\int_{\Theta} (d(\boldsymbol{x}) - \theta)^2 f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta = 2\int_{\Theta} (d(\boldsymbol{x}) - \theta) f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta = 0$$

which leads to

$$d(\boldsymbol{x}) = d(\boldsymbol{x})\underbrace{\int_{\Theta} f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta}_{=1} = \int_{\Theta} d(\boldsymbol{x}) f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta = \int_{\Theta} \theta f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta =: E\left[\theta|\boldsymbol{x}\right].$$

Therefore, the Bayesian estimator should be chosen to be the ***conditional expectation***

$$d(\boldsymbol{x}) = E\left[\theta|\boldsymbol{x}\right] := \int_{\Theta} \theta f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta. \tag{15.3}$$

15.2 Computing the conditional expectation

As we showed in the previous section, when loss is measured through squared loss, the Bayesian estimator is given by the conditional expectation. Hence, at least in principle, the only thing that we need to do is to

compute the integral

$$\int_\Theta \theta f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\, d\theta,$$

involving the ***posterior density* function** $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$. However, in most application this density is not known directly. Instead what we have at our disposal are: 1) the likelihood function, 2) the prior distribution, and 3) a set of measurements $\boldsymbol{x} = (x_1, \ldots, x_n)$. Fortunately, the posterior distribution can be written in terms of these three quantities as follows.
From equation (15.1) we can conclude that

$$\begin{aligned} f_{\Theta|\boldsymbol{X}}(\theta, \boldsymbol{x}) &= \frac{f_{\boldsymbol{X},\Theta}(\boldsymbol{x}, \theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} && \text{(Posterior = Joint / Marginal)} \\ &= \frac{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\pi(\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})} && \text{(From (15.1))} \\ &= \frac{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\pi(\theta)}{\int_\Theta f_{\boldsymbol{X},\Theta}(\boldsymbol{x}, \theta)\, d\theta} && \text{(Definition of marginal PDF)} \\ &= \frac{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\pi(\theta)}{\int_\Theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\pi(\theta)\, d\theta} && \text{(From (15.1)).} \end{aligned}$$

Hence the conditional expectation in equation (15.3) may be computed using only the prior, the likelihood and the data measurements according to the formula

$$E\left[\theta|\boldsymbol{x}\right] = \frac{\int_\Theta \theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\pi(\theta)\, d\theta}{\int_\Theta f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}, \theta)\pi(\theta)\, d\theta}, \tag{15.4}$$

where the integrals are carried over the set Θ consisting of all admissible values of θ.

15.3 Examples

Example 1. We start by considering the problem of estimating the probability of success, p, on a Bernoulli trial based on n independent experiments. The PDF for a single experiment is given by

$$f_{X|P}(x, p) = p^x(1-p)^{1-x},$$

therefore, using the fact that the experiments are independent, the likelihood function for n trials $\boldsymbol{x} = (x_1, \ldots, x_n)$ is given by

$$f_{\boldsymbol{X}|P}(\boldsymbol{x}, p) = p^{x_1}(1-p)^{1-x_1} \times \cdots \times p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

We can go one step further by realizing that, since in a Bernoulli trial $x_i = 1$ only if the attempt was a success, the term $s = \sum_{i=1}^n x_i$ will equal the number of successes obtained after n trials. Hence

$$f_{\boldsymbol{X}|P}(\boldsymbol{x}, p) = p^s(1-p)^{n-s}.$$

We need now a PDF that encodes out knowledge (or belief) about the behavior of P as a random variable. For this example we will consider that the true value of P can be anywhere in the interval $[0, 1]$ with equal probability, leading to the uniform prior $\pi(p) = 1$ for $p \in [0, 1]$[1]. To obtain the Bayesian estimate, we must then compute the integrals

$$\int_0^1 p f_{\boldsymbol{X}|P}(\boldsymbol{x}, p)\pi(p)\, dp = \int_0^1 p^{s+1}(1-p)^{n-s}\, dp \quad \text{and} \quad \int_0^1 f_{\boldsymbol{X}|P}(\boldsymbol{x}, p)\pi(p)\, dp = \int_0^1 p^s(1-p)^{n-s}\, dp.$$

[1] A more general approach would be to assume that P follows a Beta distribution. This choice allows for some more flexibility regarding the shape of the distribution while still ensuring that the integral of the product of prior and likelihood remains somehow manageable. Very loosely speaking, the families of priors and likelihoods that "go well together" (in the sense that the integrals resulting from their product can be computed analytically with relative ease) are called ***conjugate densities***.

Even though in this particular case we could compute these integrals by expanding the binomial term and integrating the resulting polynomial, we will make use of the following properties of the Gamma function[2].

1. For any complex number z whose real part is positive $\mathrm{Re}[z > 0]$, the Gamma function is defined as
$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt. \tag{15.5}$$
2. For any real numbers a and b
$$\int_0^1 p^{a-1}(1-p)^{b-1}\, dp = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \tag{15.6}$$
3. For any real number r
$$\Gamma(r+1) = r\Gamma(r). \tag{15.7}$$

Using the property (15.6) we see that

$$\int_0^1 p f_{\boldsymbol{X}|P}(\boldsymbol{x}, p)\pi(p)\, dp = \int_0^1 p^{s+1}(1-p)^{n-s}\, dp = \int_0^1 p^{s+2-1}(1-p)^{n-s+1-1}\, dp = \frac{\Gamma(s+2)\Gamma(n-s+1)}{\Gamma(n+3)}$$

and

$$\int_0^1 f_{\boldsymbol{X}|P}(\boldsymbol{x}, p)\pi(p)\, dp = \int_0^1 p^{s}(1-p)^{n-s}\, dp = \int_0^1 p^{s+1-1}(1-p)^{n-s+1-1}\, dp = \frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)},$$

and therefore according to equation (15.4) the conditional expectation is

$$\begin{aligned}
E\left[p|\boldsymbol{x}\right] &= \frac{\int_0^1 \theta f_{\boldsymbol{X}|P}(\boldsymbol{x}, p)\pi(p)\, dp}{\int_0^1 f_{\boldsymbol{X}|P}(\boldsymbol{x}, p)\pi(p)\, dp} \\
&= \frac{\int_0^1 p^{s+1}(1-p)^{n-s}\, dp}{\int_0^1 p^{s}(1-p)^{n-s}\, dp} \\
&= \frac{\Gamma(s+2)\Gamma(n-s+1)}{\Gamma(n+3)} \left(\frac{\Gamma(s+1)\Gamma(n-s+1)}{\Gamma(n+2)}\right)^{-1} \\
&= \frac{\Gamma(n+2)\Gamma(s+2)}{\Gamma(n+3)\Gamma(s+1)} \\
&= \frac{\Gamma(n+2)(s+1)\Gamma(s+1)}{(n+2)\Gamma(n+2)\Gamma(s+1)} \qquad \text{(Using property (15.7))} \\
&= \frac{s+1}{n+2}.
\end{aligned}$$

Hence, the Bayesian estimator for the probability of success of a Bernoulli random variable, based on n experiments is

$$d(\boldsymbol{x})_{\text{Bayes}} = \frac{s+1}{n+2}.$$

This expression would seem a little bit odd, specially when compared with the one in equation (14.2) obtained through maximum likelihood

$$d(\boldsymbol{x})_{\text{MLE}} = \frac{s}{n},$$

[2] This would be the *only way* to perform the integral if the prior were chosen to be the more general Beta density.

which seems to be more natural: when trying to estimate the probability of success, one carries out n experiments, counts the number of successes s and determines the what percentage of the total turned out to be successes.

The advantage of the Bayesian estimator becomes clear when we consider the estimates produced when a single experiment is performed. Since the number of successes in a single experiment is either 0 or 1, then the maximum likelihood estimator would prescribe a probability of success equal to 0 if the experiment fails or a probability of 1 if the experiment succeeds. It should be clear that estimating that a coin is completely rigged one way or another one based on a single experiment is a little too extreme. On the other hand, the Bayesian estimator would predict a probability of success equal to $1/3$ if the experiment fails, or a probability of $2/3$ if the experiment succeeds. The Bayesian estimates seem more measured (and believable) even with the extremely low amount of information provided by a single experiment.

15.4 Exercises

1. Consider the problem of minimizing the integral in equation (15.2).

   (a) Show that if the loss is measured through the absolute error loss

   $$\mathcal{L}(\theta, d(\boldsymbol{x})) = |d(\boldsymbol{x}) - \theta|,$$

   then the Bayesian estimator is given by the conditional median, defined as the value θ_m such that $P(\theta < \theta_m | \boldsymbol{X}) = 1/2$.

   (b) Show that if the loss is measured through the absolute error loss

   $$\mathcal{L}(\theta, d(\boldsymbol{x})) = \begin{cases} 0 & \text{if } d(\boldsymbol{x}) = \theta \\ 1 & \text{if } d(\boldsymbol{x}) \neq \theta \end{cases},$$

   then the Bayesian estimator is given by the conditional mode defined as the most common value θ_M given a set of measurements $\boldsymbol{X}$.

2. Find the Bayes estimator for the parameter p of a Bernoulli random variable based on a random sample of size n if $\mathcal{L}(p, a) = (a - p)^2$ and $\pi(p) = 3p(1 - p)$ for $p \in [0, 1]$.

3. Given the likelihood $f_{X|\Theta}(x|\theta) = 1/\theta$ for $0 \leq x \leq \theta$, the loss function $\mathcal{L}(p, a) = (a - p)^2$, the prior $\pi(\theta) = \theta e^{-\theta}$ for $\theta > 0$, and considering a sample of size n:

   (a) Compute the marginal density $f_X(x)$.

   (b) Use $f_X(x)$ to compute the posterior density $f_{\Theta|X}(\theta|x)$.

   (c) Use the posterior density to compute the conditional expectation $E[\Theta|X]$.

   [Hint: Use the properties of the Gamma function $\Gamma(r)$ to compute the integrals involved].

4. Find the Bayes estimator of the Poisson parameter μ based on a random sample of size n if squared loss is used and the prior is $\pi(\mu) = e^{-\mu}$ for $\mu > 0$. [Hint: Use the properties of the Gamma function $\Gamma(r)$ to compute the integrals involved].

5. Find the Bayes estimator for the parameter θ in the likelihood function $f_{X|\Theta}(x|\theta) = \theta e^{-\theta x}$ for $x > 0$ based on a single observation of X if the loss is measured as the square loss and $\pi(\theta) = 1$ for $\theta \in [0, 1]$.

6. You will need a computer to do the integrals involved in this exercise. We will consider the problem studied in Example 1 with a prior that is not a conjugate density of the Bernoulli distribution. The likelihood for a single trial is $f_{X|P}(x,p) = p^x(1-x)^{1-x}$ where the probability of success p is known to vary around fairness according to a normal-like distribution $\pi(p) = C\exp\left(-4(p-1/2)^2\right)$ for $0 \leq p \leq 1$ and C is a constant to be determined.

   (a) Compute the value of the constant C so that $\pi(p)$ is a probability density function.

   (b) Consider a squared loss function and compute the Bayesian estimator for p based on an independent sample of size n.

Lecture 16

Confidence intervals

Contents

16.1 Estimating expectation when the variance is known

Consider the problem of estimating the value of the mean value of a random variable X with finite mean μ and variance σ^2 based on a simple random sample of size n. We will assume that the variance is known. As we have seen in the preceding sections, the sample mean $\overline{X}_n$ is the best unbiased estimator (i.e. it has the minimum variance allowed by the information inequality). However, it would be nice to know how likely the resulting estimation $\overline{x}_n$ is to be within a certain range of the actual value μ. We will use the fact that the sample mean has the probability density function

$$f_{\overline{X}_n}(\overline{x}, \mu, \sigma^2, n) = \sqrt{\frac{n}{2\pi\sigma^2}}\, e^{-\frac{(\overline{x}_n-\mu)^2}{2(\sigma^2/n)}}$$

we can compute the probability that a particular realization of $\overline{x}_n$ falls within a certain distance from the true value μ. We first observe that if we define

$$z := \frac{\overline{x}_n - \mu}{(\sigma/\sqrt{n})}, \tag{16.1}$$

the resulting random variable Z (often known as the ***z-score*** of the sample) is also normally distributed, has mean zero, unit variance, and PDF

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Hence the probability that a particular realization of Z falls within two units (two standard deviations) of zero (the mean value of Z) is

$$P(|Z| < 2) \approx 0.95 \tag{16.2}$$

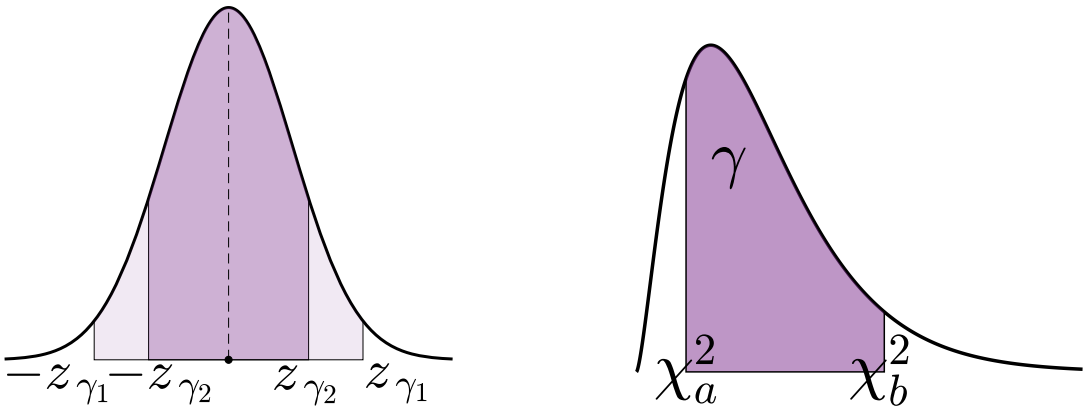

Figure 16.1: Left: The area under the standard normal curve and between the points $-z_\gamma$ and z_γ determines the confidence level of the interval. If $z_{\gamma_1} > z_{\gamma_2}$ the confidence associated to z_{γ_1} is larger, but the margin of error increases as well. Right: The area under the curve of the χ^2 density between the points χ^2_γ determines the confidence level for the variance interval. The values of χ^2_a and χ^2_b will change depending on the target confidence level γ and the sample size n (degrees of freedom).

Substituting the value of z and observing that

$$\frac{|\overline{x}_n - \mu|}{(\sigma/\sqrt{n})} < 2 \iff -\frac{2\sigma}{\sqrt{n}} + \mu < \overline{x}_n < \mu + \frac{2\sigma}{\sqrt{n}}$$

this is equivalent to

$$P\left(-\frac{2\sigma}{\sqrt{n}} + \overline{x}_n < \mu < \overline{x}_n + \frac{2\sigma}{\sqrt{n}}\right) \approx 0.95.$$

Hence, approximately 95% of the realizations of the sample mean $\overline{x}_n$ will be within a distance $2\sigma/\sqrt{n}$ of the true value. Note that the values the distance $|Z| < 2$ and the probability 0.95 in equation are related. If we allow the distance to grow, the probability will increase and vice versa.
Hence, given a target probability $\gamma \in [0, 1]$ (known as the ***confidence level***) there will be a corresponding distance z_γ that will guarantee that

$$P\left(-\frac{z_\gamma\sigma}{\sqrt{n}} + \overline{x}_n < \mu < \overline{x}_n + \frac{z_\gamma\sigma}{\sqrt{n}}\right) = \gamma, \tag{16.3}$$

where, for a target ***confidence level*** γ, the critical value z_γ is found by applying the inverse of the cumulative distribution function of the standard normal distribution. We then define the ***z-confidence interval for the mean*** with confidence level γ as

$$\overline{x}_n \pm z_\gamma \frac{\sigma}{\sqrt{n}}, \tag{16.4}$$

where the term $z_\gamma\sigma/\sqrt{n}$ is known as the ***z-margin of error***. When providing an estimate of the mean, we would like to have the smallest possible margin of error. Since we have no control over the variance σ^2 of the random variable, we can achieve this by two mechanisms: (1) decreasing the magnitude of the critical values z_γ, or (2) increasing the sample size n. Given that the parameter z_γ controls the size of the region under the normal curve, as can be seen in the left pab=nel of Figure16.1 decreasing this value has the disadvantage of decreasing the confidence level of the estimate. Increasing the sample size on the other hand will not decrease the confidence level, but may be expensive or impossible in practice.

16.2 The χ^2 distribution

Consider a random variable X and define $Y := X^2$. The CDF $F_Y(y)$ of Y is given by

$$F_Y(y) = P(Y \leq y) = P\left(X^2 \leq y\right) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \quad \text{for } y > 0.$$

Noting that since Y is non negative, it follows that $F_Y(y) \equiv 0$ for $y \leq 0$. From here differentiation with respect to y yields the PDF for Y as

$$f_Y(y) = \begin{cases} 0 & \text{for } y \leq 0, \\ \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right) & \text{for } y > 0, \end{cases} \tag{16.5}$$

We will now assume that $X \sim N(0,1)$, then the random variable X^2 is particularly important in statistics and receives the name of ***chi squared distribution*** with one degree of freedom and is denoted by $\chi^2(1)$ (we will come back to the point of the degrees of freedom below). We can use (16.5) to express the PDF for $\chi^2(1)$ as:

$$f_{\chi^2(1)}(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} & \text{for } x > 0. \end{cases} \tag{16.6}$$

The density above is a particular case of a Gamma density with parameters $\alpha = 1/2$ and $\beta = 1/2$. One property of Gamma densities is that if $X_1, \ldots, X_n$ are independent random variables each of them following the Gamma density $\Gamma(\alpha_i, \beta)$ then the random variable $Z = X_1 + \cdots + X_n$ follows the Gamma density $\Gamma(\alpha, \beta)$, where $\alpha = \sum_{i=0}^{n} \alpha_i$. As a consequence of this property we can state the following

Theorem 16.1. *If $X_1, \ldots, X_n$ are independent random variables each of them following the chi squared density $\chi^2(1)$, then*

$$Z = X_1 + \cdots + X_n$$

follows also a **chi squared distribution** *with n degrees of freedom with PDF given by $\Gamma(n/2, 1/2)$. This distribution is denoted as $\chi^2(n)$.*

In summary we can say that, if $X_1, \ldots, X_n$ are IID from a normal distribution with mean μ and standard deviation σ, then the random variable

$$\chi^2(n) := \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2,$$

has a ***chi squared distribution*** with n degrees of freedom given by the Gamma distribution $\Gamma(n/2, 1/2)$. The expected value and the mean of $\chi^2(n)$ can be shown to be given by

$$E[X] = n \qquad \text{and} \qquad \text{Var}[X] = 2n.$$

16.3 Estimating variance when the expectation is known

We will now construct a confidence interval for the variance σ^2 of a normal density when the mean μ of the distribution is known. We have determined that $\sum_{i=1}^{n}(X_i - \mu)^2/n$ is the best unbiased estimator for the variance and we will use this estimator to construct a confidence interval.
As we have seen in the previous section, given a simple random sample $X_1, \ldots, X_n$ drawn from a normal distribution, the set

$$\chi_i^2 := (X_i - \mu)^2/\sigma^2 \qquad \text{for } i = 1, \ldots, n,$$

follows a chi squared density with n degrees of freedom as shown on the right panel of Figure 16.1. Hence, using the inverse CDF for $\chi^2(n)$ (numerically if necessary), it is possible to find numbers χ^2_a and χ^2_b such that, for a target confidence level $\gamma \in [0, 1]$

$$P\left(\chi^2_a < \chi^2 < \chi^2_b\right) = \gamma.$$

Substituting the definition of χ^2, the expression above is equivalent to

$$P\left(\sum_{i=1}^{n}(X_i - \mu)^2/\chi^2_b < \sigma^2 < \sum_{i=1}^{n}(X_i - \mu)^2/\chi^2_a\right) = \gamma.$$

Hence, the γ-confidence interval for σ^2 is

$$\left(\sum_{i=1}^{n}(X_i - \mu)^2/\chi^2_b \ , \ \sum_{i=1}^{n}(X_i - \mu)^2/\chi^2_a\right)$$

16.4 Estimating the variance when the expectation is unknown

It can be shown that the random variable

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{x}_n)^2,$$

where $\overline{x}_n$ denotes the sample mean, **follows a chi squared distribution with $n-1$ degrees of freedom**. Hence, if the value of μ is not known, we can approximate $\mu \approx \overline{x}_n$ by the sample mean and repeat the process outlined in the previous section **using a chi squared distribution with $n-1$ degrees of freedom and adjusting the values of χ^2_a and χ^2_b accordingly.**

16.5 Student's T distribution

When neither the mean nor the variance of a distribution are known it is natural to use the sample mean $\overline{x}_n$ and the sample variance $\overline{s}^2_n := \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \overline{x}_n)^2$ as estimators for the mean and variance respectively. To determine a confidence interval for the mean similar to the one given in (16.3) but replacing σ by the sample standard deviation $\overline{s}_n$ we need to study the properties of the random variable

$$T := \frac{X}{\overline{s}_n/\sqrt{n}}.$$

We will start by defining a random variable formed by taking the quotient between two chi squared random variables. Let us Y_1 and Y_2 be follow chi squared distributions with k_1 and k_2 degrees of freedom respectively and define

$$Z := \frac{Y_1/k_1}{Y_2/k_2}. \tag{16.7}$$

Using Corollary 4.4 it is possible to show that Z has the PDF

$$f_Z(x, k_1, k_2) = \frac{(k_1/k_2)\Gamma[(k_1+k_2)/2](k_1x/k_2)^{(k_1+k_2)/2-1}}{\Gamma(k_1/2)\Gamma(k_2/2)(1+(k_1x/k_2))^{(k_1+k_2)/2}}, \qquad \text{for } x > 0. \tag{16.8}$$

This random variable is said to have an ***F-distribution*** and its CDF is denoted as $F(k_1, k_2)$.

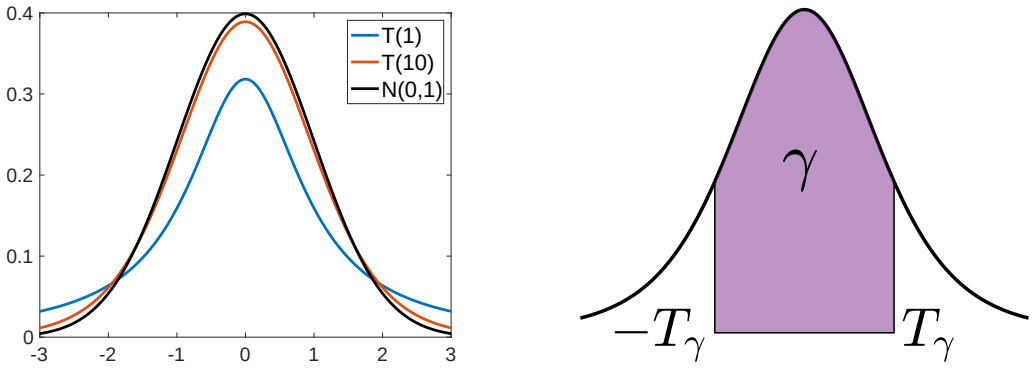

Figure 16.2: Left: Comparison between a standard normal distribution (black line) and T distributions with different degrees of freedom. Right: The value of T_γ is determined by the location such that the area under the curve and between $-T_\gamma$ and T_γ is equal to γ.

We then consider a simple random sample $X_1, \ldots, X_n$ taken from a normal distribution with mean μ and standard deviation σ. We know from the previous section that if we take a subset of the sample (say from 1 to j) the random variables

$$Y_1 := \frac{1}{\sigma^2}\sum_{i=1}^{j}(X_i - \mu)^2 \qquad Y_2 := \frac{1}{\sigma^2}\sum_{i=j+1}^{n}(X_i - \mu)^2$$

are independent and follow the distributions $\chi_1^2(j)$ and $\chi_2^2(n-j)$. Consider then the case $j = 1$ (i.e. the sample is split into one group with a single observation and one group with all the rest). Defining $Y_1 \sim \chi^2(1)$ and $Y_2 \sim \chi^2(n-1)$ as above and substituting into equation (16.7) yields

$$Z^2 := \frac{(X_1 - \mu)^2}{\sum_{i=2}^{n}(X_i - \mu)^2/(n-1)}$$

which is a random variable with $F(1, n-1)$ distribution. Recalling that all X_i are normally distributed and therefore have a symmetric PDF, we can conclude that the random variable

$$Z := \frac{(X_1 - \mu)}{\sqrt{\sum_{i=2}^{n}(X_i - \mu)^2/(n-1)}} \tag{16.9}$$

has a symmetric PDF as well. This fact, combined with the formula (16.5) (where we make the substitution $y = z^2$) and the (16.8), can be used to obtain the PDF for Z

$$f_Z(z) = \frac{\Gamma\left[(n+1)/2\right]\left(1 + z^2/n\right)^{-(n+1)/2}}{\sqrt{n\pi}\,\Gamma(n/2)} \qquad \text{for } y \in (-\infty, \infty). \tag{16.10}$$

A random variable of the form (16.9) with PDF given by (16.10), is said to have a ***T-distribution*** with n degrees of freedom. As shown in Figure 16.2, the T-distribution looks remarkably similar to the standard normal distribution as the number of degrees of freedom increases. However, for $n < \infty$ the two distributions are indeed different and the difference manifests itself more clearly for larger values of $|x|$ as the fact that if m is a large integer, and T and Z follow respectively a T distribution and a standard normal distribution then $P(|T| > m) > P(|Z| > m)$.

16.6 Estimating the mean when the variance is unknown

When using the sample variance $\overline{s}_n^2$ instead of the variance σ^2 to determine a confidence interval, a very similar argument to the one we used in Section 16.1 (using the T distribution with n degrees of freedom

instead of the normal distribution) can be used to show that for a target confidence level γ the confidence interval takes the form

$$\overline{x}_n \pm T_\gamma \frac{\overline{s}_n}{\sqrt{n}}, \tag{16.11}$$

where, for a given confidence level γ the number T_γ is such that an area equal to γ is enclosed by the curve of the T-density function (16.10) and the values $-T_\gamma$ and T_γ, as shown in Figure 16.2.

16.7 Summary

Consider a simple random samples consisting of measurements $X_1, \ldots, X_n$ from a random variable X with mean μ and standard deviation σ and a target confidence level $\gamma \in [0, 1]$. We recall the definition of the sample mean and the sample variance respectively

$$\overline{x}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad \overline{s}_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x}_n)^2.$$

Confidence intervals for the mean.

- If the variance is known, then the relevant distribution for z_γ is $N(0, 1)$ and the interval is

$$\left(\overline{x}_n - z_\gamma \frac{\sigma}{\sqrt{n}}, \overline{x}_n + z_\gamma \frac{\sigma}{\sqrt{n}}\right)$$

- If the variance is unknown, then the relevant distribution for T_γ is $T(n)$ and the interval is

$$\left(\overline{x}_n - T_\gamma \frac{\overline{s}_n}{\sqrt{n}}, \overline{x}_n + T_\gamma \frac{\overline{s}_n}{\sqrt{n}}\right)$$

Confidence intervals for the variance.

- If the mean is known, then the relevant distribution for χ_a^2 and χ_b^2 is $\chi^2(n)$ and the interval is

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_b^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_a^2}\right)$$

- If the mean is unknown, then the relevant distribution for χ_a^2 and χ_b^2 is $\chi^2(n-1)$ and the interval is

$$\left(\frac{\sum_{i=1}^n (X_i - \overline{x}_n)^2}{\chi_b^2}, \frac{\sum_{i=1}^n (X_i - \overline{x}_n)^2}{\chi_a^2}\right)$$

16.8 Exercises

1. Consider that a sample of size $n = 25$ was drawn from the distribution

$$f(x|\mu) = \frac{1}{\sqrt{32\pi}} e^{-\frac{(x-\mu)^2}{32}},$$

and has a sample mean $\overline{x}_n = 20$.

   (a) Find a 90% confidence interval for μ.

   (b) How large should the sample be if the confidence interval is to be reduce to half its length.

2. Construct a 95% confidence interval for the parameter p of a Bernoulli density

$$f(x|p) = p^x(1-p)^{1-x},$$

if a random sample of size $n = 50$ yielded $\sum_{i=1}^{50} X_i = 15$. Use the normal approximation for $\sum_{i=1}^{50} X_i/50$ and replace its variance by its sample variance $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$.

3. Consider a sample of size $n = 20$ drawn from

$$f(x|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-10)^2}{2\sigma^2}}.$$

If $\sum_{i=1}^{20} X_i = 180$ and $\sum_{i=1}^{20} X_i^2 = 2000$:

   (a) Find a 90% confidence interval for σ^2.

   (b) Find a 90% confidence interval for σ under the assumption that $\mu = 0$ but it is not known that the true value of the variance is $\sigma^2 = 1$.

   (c) Find a 90% confidence interval for σ under the assumption that the value of μ is unknown.

4. Build an 80% confidence interval for the parameter θ in the density

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta} \quad \text{for } x > 0,$$

if a random sample of size $n = 20$ yielded $\sum_{i=1}^{20} X_i = 80$. Use the normal approximation for $\sum_{i=1}^{20} X_i/20$.

Appendix A

Examples of random variables

The following list is of course incomplete, and should be considered a reference. You are not expected to memorize all, but many of these random variables are so relevant in applications that it is to your best interest to at least be acquainted with them. In all instances below, p represents the probability mass function, and D represents the (discrete) set of possible values of x. The natural numbers (i.e. positive integers) are represented by $\mathbb{N} := \{1, 2, 3, \ldots\}$.

A.1 Discrete random variables

1. **Uniform**. $U(a, b)$, $D = \{a, a + 1, \ldots, b - 1, b\}$.

$$p(x) = \frac{1}{b - a + 1}.$$

   The probability of drawing an integer between a and b in one attempt when all the integers in the interval are equally likely to be drawn.

2. **Bernoulli**. $\text{Ber}(p)$, $D = \{0, 1\}$.

$$p(x) = p^x(1 - p)^{1-x}.$$

   Represents the probability of attaining either success ($x = 1$) or failure ($x = 0$) in a single attempt with constant probability of success p.

3. **Binomial**. $\text{Bin}(n, p)$, $D = \{0, 1, \ldots, n\}$

$$p(x) = \binom{n}{x} p^x(1 - p)^{n-x}.$$

   Note that $\text{Ber}(p) = \text{Bin}(1, p)$. A binomial RV is usually interpreted as giving the probability of succeeding exactly x times in an experiment with n attempts each of which has probability p of succeeding.

4. **Geometric** $\text{Geo}(p)$, $D = \mathbb{N}$.

$$p(x) = p(1 - p)^x.$$

   Represents the probability that exactly x Bernoulli trials are required to produce the first success. Alternatively, it represents the probability that $x - 1$ failures occur in a Bernoulli trial before the first success.

5. **Hypergeometric**. $\text{Hyp}(N, n, k)$, $N \in \mathbb{N}$, $k, n \in \{0, 1, \ldots, N\}$,

$D = \{\max\{0, n - N + k\}, \ldots, \min\{n, K\}\}$.

$$p(x) = \frac{\binom{n}{x}\binom{N-n}{k-x}}{\binom{N}{n}}.$$

The hypergeometric distribution considers a population of size N containing n individuals with a "special" property (for instance k red balls in a bag otherwise containing N blue balls) and returns the probability of finding x individuals with the special property within a sample of size n when drawing is done *without replacement*.

6. **Negative binomial**. Negbin(n, p), $D = \mathbb{N}$.

$$p(x) = \binom{n + x - 1}{x} p^n (1 - p)^x$$

This distribution pertains to the probability that x successes are achieved before n failures happen on on independent Bernoulli trials with probability of success p. Note that $Geo(p) = Negbin(1, p)$.

7. **Poisson**. Pois(λ), $D = \mathbb{N}$, $\lambda \in (0, \infty)$.

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Describes the probability that x independent events take place within a fixed interval, if the events are known to happen randomly with constant mean rate λ.

A.2 Continuous random variables

1. **Uniform**. $U(a, b)$, $x \in [a, b]$.

$$f(x) = \frac{1}{b - a}.$$

2. **Beta**. Beta(α, β), $x \in [0, 1]$, $\alpha > 0$, $\beta > 0$.

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1}.$$

3. **Cauchy**. Cau(μ, σ^2), $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma x \in (0, \infty)$.

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

4. **Chi-squared**. χ^2_a, $x \in [0, \infty)$, $a \in \mathbb{N}$.

$$f(x) = \frac{x^{\frac{a}{2} - 1}}{2^{a/2}\Gamma(a/2)} e^{-x/2}.$$

5. **Exponential**. Exp(θ), $x \in [0, \infty)$, $\theta > 0$.

$$f(x) = \theta e^{-\theta x}.$$

6. **Fischer's F**. $F_{q,a}$, $x \in [0,\infty)$, $q > 0$, $a > 0$.

$$f(x) = \frac{\Gamma((q+a)/2) q^{q/2} a^{a/2}}{\Gamma(q/2)\Gamma(a/2)} \, x^{\frac{q}{2}-1} (a+qx)^{-(q+a)/2}.$$

7. **Gamma**. $\Gamma(\alpha, \beta)$, $x \in [0,\infty)$, $\alpha > 0$, $\beta > 0$.

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}.$$

   Observe that $\text{Exp}(\theta) = \Gamma(1, \theta)$.

8. **Inverse Gamma**. $\Gamma^{-1}(\alpha, \beta)$, $x \in [0,\infty)$, $\alpha > 0$, $\beta > 0$.

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \alpha^{-\alpha-1} e^{-\beta/x}.$$

9. **Laplace**. $\text{Lap}(\mu, \sigma)$, $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in (0,\infty)$

$$f(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}.$$

10. **Normal**. $N(\mu, \sigma^2)$, $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in (0,\infty)$

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

11. **Pareto**. $\text{Par}(\alpha, c)$, $x \in [c,\infty)$, $c \in (0,\infty)$, $\alpha \in (0,\infty)$.

$$f(x) = \frac{c^\alpha \alpha}{x^{\alpha+1}}.$$

Bibliography

[1] P.G. Hoel. *Introduction to Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 1984. ISBN: 9780471890454. URL: https://books.google.com/books?id=lBPvAAAAMAAJ.

[2] E. L. Lehmann J. L. Hodges. *Basic Concepts Of Probability And Statistics*. 2nd ed. Classics in applied mathematics 48. Society for Industrial and Applied Mathematics, 2005. ISBN: 9780898715750,089871575X. URL: https://epubs.siam.org/doi/book/10.1137/1.9780898719123.

[3] Charles J. Stone Paul G. Hoel Sidney C Port. *Introduction to statistical theory*. 1st ed. The Houghton-Mifflin series in statistics. Houghton-Mifflin, 1971. ISBN: 9780395046371,0395046378. URL: https://books.google.com/books/about/Introduction_to_Statistical_Theory.html?id=N5_At9Jq2eoC.

[4] Charles J. Stone Paul G. Hoel Sidney C. Port. *Introduction to Probability Theory*. 1st ed. Brooks Cole, 1972. ISBN: 039504636X,9780395046364. URL: https://books.google.com/books/about/Introduction_to_Probability_Theory.html?id=LxPvAAAAMAAJ.

[5] J.S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, 2006. ISBN: 9789812703705. URL: https://books.google.com/books?id=PYNAgzDffDMC.

[6] Tonatiuh Sánchez-Vizuet. *Minimal analysis I: A concise introduction to Real Analysis of a single variable*. 2025.

[7] Tonatiuh Sánchez-Vizuet. *Minimal analysis II: A concise introduction to Real Analysis of multiple variables*. 2025.

[8] Larry Wasserman. *All of Statistics A Concise Course in Statistical Inference*. Springer, 2004. URL: https://books.google.com/books/about/All_of_Statistics.html?id=RMdgcgAACAAJ.

[9] J Watkins. *An Introduction to the Science of Statistics: From Theory to Implementation*. Unpublished, 2005. URL: https://www.math.arizona.edu/~jwatkins/statbook.pdf.

Alphabetical Index