Studies in Statistics for Economists (212.503 - 001)

Prof. Myung Hwan Seo
Department of Economics
Seoul National University

March, 2023

Contents

The lecture for most part of the semester will be based on this note. However, here are some references for those who want to possess a/some books. They may cover more materials than required in this course:

References

[1] *Probability*, Bruce Hansen, 2021, https://www.ssc.wisc.edu/~bhansen/probability/

[2] G. Casella and R.L. Berger. 2001. *Statistical Inference*. Duxbury Press. 2nd edition. for Probability and Statistics.

[3] R. Hogg and A. Craig 1995. *Introduction to Mathematical Statistics* Prentice Hall.

[4] H. White 2000. *Asymptotic Theory for Econometricians.* Academic Press. 2nd edition. for asymptotic theory.

[5] J. Davidson 1994. *Stochastic Limit Theory.* Oxford University Press.

[6] J. Shao 2003. *Mathematical Statistics.* Springer. 2nd edition.

1 Probability

1.1 Probability Space

A *probability space* is the triple $(\Omega, \mathcal{F}, \mathbf{P})$, each of which will be introduced below.

First, Ω is a set called the *sample space*. It is the set consisting of all the possible outcomes of a random experiment. An element ω of Ω is called an *outcome*. An *event,* E is a collection of possible outcomes.

Second, $\mathcal{F}$ is a collection of subsets of Ω, called a σ-field. More precisely,

Definition 1.1. A class $\mathcal{F}$ of subsets of Ω is called a σ*-field* if it satisfies the following three properties:
(a) $\Omega \in \mathcal{F}$.
(b) $E \in \mathcal{F}$ implies $E^c \in \mathcal{F}$.
(c) $E_1, E_2, \ldots \in \mathcal{F}$ implies $\bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$.

Remarks It follows immediately that
(a) $\emptyset \in \mathcal{F}$ since $\emptyset = \Omega^c$.
(b) $E_1, E_2, \ldots \in \mathcal{F}$ implies $\bigcap_n E_n \in \mathcal{F}$, because $\bigcap_n E_n = (\bigcup_n E_n^c)^c$.
(c) The smallest σ-field associated with Ω is $\mathcal{F} = \{\emptyset, \Omega\}$.

Example If you toss a coin, $\Omega = \{H, T\}$, where H and T stand for head and tail of the coin, respectively. If you toss two coins, $\Omega = \{HH, HT, TH, TT\}$. You can think of various events such as $E = \{HH, HT\}$, that is, an event where the first coin shows the head of the coin.

Example The *Borel σ-field,* $\mathcal{B}$, on a topological space is the σ-field *generated* by the family of open subsets. For example, let $\Omega = (-\infty, \infty)$, the real line. Then, $\mathcal{B}$ contains all sets of the form

$$[a, b]\,, (a, b], (a, b)\,, \text{ and } [a, b)$$

for all real numbers a and b. It follows from the property of the σ-field that $\mathcal{B}$ contains all sets that can be formed by taking unions and intersections of the above varieties. But, it does not contain all the subsets of the real line.

Third, $\mathbf{P}$ is a real-valued function defined on a σ-field $\mathcal{F}$. We define

Definition 1.2. A set function $\mathbf{P}$ on the σ-field $\mathcal{F}$ is a probability or probability measure if it satisfies the following conditions:
(a) $\mathbf{P}(E) \geq 0$ for all $E \in \mathcal{F}$.
(b) $\mathbf{P}(\Omega) = 1$.
(c) If $E_1, E_2, \ldots \in \mathcal{F}$ are disjoint, then

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(E_n).$$

Example A coin, possibly biased, is tossed. Let $\Omega = \{H, T\}$, $\mathcal{F} = \{\emptyset, \Omega, \{H\}, \{T\}\}$ and the probability with which H occurs is p. The probability for this experiment is $\mathbf{P}$:$\mathcal{F} \to [0, 1]$, given by $\mathbf{P}(\emptyset) = 0$, $\mathbf{P}(H) = p$, $\mathbf{P}(T) = 1 - p$, $\mathbf{P}(\Omega) = 1$.

Properties Many ‘already-well-known’ properties of probability follow directly from the above axioms.

1. For $E, F \in \mathcal{F}$ such that $E \subset F$, it follows that $\mathbf{P}(E) + \mathbf{P}(F - E) = \mathbf{P}(F)$. We therefore have

   $$E \subset F \text{ implies } \mathbf{P}(E) \leq \mathbf{P}(F)$$

   i.e., probability is *monotone*. It follows further that $\mathbf{P}(F - E) = \mathbf{P}(F) - \mathbf{P}(E)$, and we have as special cases

   $$\mathbf{P}(\emptyset) = 0 \quad \text{and} \quad \mathbf{P}(E^c) = 1 - \mathbf{P}(E)$$

2. It can also be easily deduced that

   $$\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F)$$

   the common value of the two sides being $\mathbf{P}(E \cap F^c) + \mathbf{P}(E \cap F) + \mathbf{P}(E^c \cap F)$. Note that $E = (E \cap F) \cup (E \cap F^c)$ and $F = (E \cap F) \cup (E^c \cap F)$. Consequentially,

   $$\mathbf{P}(E \cup F) \leq \mathbf{P}(E) + \mathbf{P}(F),$$

which is the subadditivity of the probability (aka *Boole's Inequality* or *union bound*). And

$$\mathbf{P}\left(E \cap F\right) \geq \mathbf{P}(E) + \mathbf{P}(F) - 1,$$

as $\mathbf{P}(E \cup F) \leq 1$. This inequality is known as *Bonferroni's Inequality* (a special case).

cf. If $\mathbf{P}(A) = 1$, we say that A occurs *almost surely* (a.s.).

1.2 Conditional Probability and Independence

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.

Definition 1.3. For an event F such that $\mathbf{P}(F) > 0$, we define the conditional probability of E given F by

$$\mathbf{P}(E \,|\, F) = \frac{\mathbf{P}(E \cap F)}{\mathbf{P}(F)}.$$

Properties

1. For a fixed F the function $\mathbf{Q}(\cdot) = \mathbf{P}(\cdot|\, F)$ is a set function which satisfies the three axioms of probability with F as the new sample space. We may indeed easily show that $\mathbf{Q}(E) \geq 0$, $\mathbf{Q}(F) = 1$ and, for disjoint events $E_1, E_2, \ldots$, $\mathbf{Q}(\bigcup E_n) = \sum \mathbf{Q}(E_n)$. Being a probability, all the properties of probability introduced earlier hold for $\mathbf{Q}$. For instance, it is immediate that $\mathbf{Q}(\emptyset) = 0$ and $\mathbf{Q}(E^c) = 1 - \mathbf{Q}(E)$, which implies that $\mathbf{P}(\emptyset \,|\, F) = 0$ and $\mathbf{P}(E^c \,|\, F) = 1 - \mathbf{P}(E \,|\, F)$.

2. It is often convenient to define a probability of intersection via a conditional probability, i.e.,

$$\mathbf{P}(E \cap F) = \mathbf{P}(E \,|\, F)\,\mathbf{P}(F)$$

which can easily be extended to

$$\mathbf{P}(E \cap F \cap G) = \mathbf{P}(E \,|\, F \cap G)\,\mathbf{P}(F \,|\, G)\,\mathbf{P}(G)$$

or more general cases

$$\mathbf{P}(E_1 \cap E_2 \cap \cdots \cap E_n) = \prod_{i=2}^{n} \mathbf{P}(E_i | I_{i-1})\mathbf{P}(E_1) \text{ where } I_{i-1} = E_{i-1} \cap E_{i-2} \cap \cdots \cap E_1.$$

3. It is said that $\{F_n\}$ is a *partition* of Ω such that $\mathbf{P}(F_n) > 0$ for all n when F_n's are mutually exclusive and exhaustive, i.e., Ω is their disjoint union. In general, partitions are very useful, allowing us to divide the sample space into small, non-overlapping pieces. If a sequence $\{F_n\}$ of events is a partition of Ω, then

$$\mathbf{P}(E) = \sum_n \mathbf{P}(E \,|\, F_n)\mathbf{P}(F_n)$$

for any event E. This property is often referred to as the *theorem of total probability*. The *Bayes' Formula*

$$\mathbf{P}(F_k \,|\, E) = \frac{\mathbf{P}(E \,|\, F_k)\mathbf{P}(F_k)}{\sum_n \mathbf{P}(E \,| F_n)\mathbf{P}(F_n)}$$

is also immediate.

Definition 1.4. Events E and F are called *independent* if

$$\mathbf{P}(E \cap F) = \mathbf{P}(E)\mathbf{P}(F).$$

Remarks

1. We may equivalently reformulate the definition of independent events as

$$\mathbf{P}(E \,|\, F) = \mathbf{P}(E) \quad \text{or} \quad \mathbf{P}(F \,|\, E) = \mathbf{P}(F)$$

when $\mathbf{P}(F) > 0$ or $\mathbf{P}(E) > 0$, respectively. Thus, the occurrence of E has no effect on F and vice versa.

2. A null set and the sample space are independent of any event.

3. Independence of E and F implies independence of the complements also as we can write

$$\mathsf{Pr}\left(E^c|F\right) = 1 - \mathsf{Pr}\left(E|F\right) = 1 - \mathsf{Pr}\left(E\right) = \mathsf{Pr}\left(E^c\right),$$

assuming $\mathsf{Pr}\left(F\right) > 0$ wlog.

4. In general, $E_1, E_2, \ldots$ are said to be independent if, for any subcollection, $E_{i_1}, E_{i_2}, \cdots, E_{i_k}$

$$\mathbf{P}(E_{i_1} \cap E_{i_2} \cap \cdots E_{i_n}) = \prod_{k=1}^{n} \mathbf{P}(E_{i_k}).$$

Note that

$$\mathbf{P}(E_1 \cap E_2 \cap \cdots E_n) = \prod_{i=1}^{n} \mathbf{P}(E_i)$$

cannot guarantee pairwise independence. (Example 1.3.10)

2 Random Variables, Distributions and Densities

2.1 Random Variables and Distributions

Let a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ be given. We define

Definition 2.1. A random variable X is a measurable function from Ω to $\mathcal{R}$.

Remarks:

(a) A random variable X is a *function* from Ω to $\mathcal{R}$, i.e., it assigns a number to each outcome.

(b) It is *measurable*, i.e.,

$$X^{-1}(A) = \{\omega |\, X(\omega) \in A\} \in \mathcal{F}$$

for any $A \in \mathcal{B}(\mathcal{R})$.

Examples:

(a) For the random experiment of tossing a coin, we may define a random variable

X by $X(H) = 1$ and $X(T) = 0$, where H and T denote outcomes of the coin landing on its head and tail, respectively.
(b) For the random experiment of tossing a coin 50 times, if we define $Y =$ the number of heads in 50 tosses, then Y takes on the integers $\{0, 1, ..., 50\}$.

Definition 2.2. The distribution P_X of a random variable X is the probability measure on $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$ induced by X. It is thus defined by

$$\mathsf{P}_X(A) = \mathbf{P}\left(X^{-1}(A)\right) = \mathbf{P}\{\omega | \, X(\omega) \in A\}$$

Note that $\mathbf{P}$ is the probability measure on $\mathcal{F}$, whereas P_X is the probability measure on $\mathcal{B}(\mathcal{R})$, i.e. *Borel σ-field.*

We often write $\mathsf{P}_X = \mathbf{P} \circ X^{-1}$. When there is no ambiguity about the underlying random variable, we usually write P in the place of P_X for simplicity.

Remarks: We may easily show that P is indeed a probability measure, i.e., it satisfies three axioms of probability. Notice that

(a) $\mathsf{P}(A) = \mathbf{P}\left(X^{-1}(A)\right) \geq 0$ for any $A \in \mathcal{B}(\mathcal{R})$.
(b) $\mathsf{P}(\mathcal{R}) = \mathbf{P}(\Omega) = 1$.
(c) If $A_1, A_2, \ldots \in \mathcal{B}(\mathcal{R})$ are disjoint, then

$$\begin{aligned}
\mathsf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) &= \mathbf{P}\left(X^{-1}\left(\bigcup_{n=1}^{\infty} A_n\right)\right) \\
&= \mathbf{P}\left(\bigcup_{n=1}^{\infty} X^{-1}(A_n)\right) \\
&= \sum_{n=1}^{\infty} \mathbf{P}\left(X^{-1}(A_n)\right) \\
&= \sum_{n=1}^{\infty} \mathsf{P}(A_n)
\end{aligned}$$

The properties of probability introduced earlier thus apply to P_X.

Example: Consider the random variable X introduced in the above example, and let $\mathbf{P}\{H\} = 1/3$ and $\mathbf{P}\{T\} = 2/3$. The distribution P_X of X is then given by

$$\mathbf{P}\{T\} = \mathsf{P}_X\{0\} = \frac{2}{3} \quad \text{and} \quad \mathbf{P}\{H\} = \mathsf{P}_X\{1\} = \frac{1}{3}$$

More precisely, $\mathsf{P}_X(A) = 1/3, 2/3, 1$ depending whether A contains only 1, only 0 or both. If A contains neither 0 or 1, then $\mathsf{P}_X(A) = 0$.

Definition 2.3. The distribution function F_X of a random variable X is defined by

$$F_X(x) = \mathsf{P}_X(-\infty, x]$$

As in the case of distribution, we often omit the subscript X of F_X and write F when there is no danger of confusion. We write $X \sim F$ to indicate that X has a distribution given by F.

Remarks:
(a) It is noted that P_X is the probability measure on *Borel σ-field* whereas F_X is the function on a real line.
(b) It is known that the distribution P is uniquely determined by the distribution function F.

Properties of Distribution Function:

(a) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
(b) $F(x) \le F(y)$ if $x \le y$.
(c) F is right continuous.

Proof: (a) Let $x_n \uparrow \infty$. Then,

$$1 = \mathsf{P}\{R\} = \sum_{n=1}^{\infty} \mathsf{P}(x_{n-1}, x_n] = \lim_{N\to\infty} \sum_{n=1}^{N} \mathsf{P}(x_{n-1}, x_n] = \lim_{N\to\infty} F(x_N),$$

where $x_0 = -\infty$. The case $x \to -\infty$ is similar.
(b) Let $x \le y$. Then $(-\infty, x] \subset (-\infty, y]$ and it follows that

$$F(x) = \mathsf{P}(-\infty, x] \le \mathsf{P}(-\infty, y] = F(y)$$

by the monotonicity of P.
(c) Fix an arbitrary x. For the right continuity of F at x, it suffices to show that $F(x_n) \to F(x)$ for any sequence $\{x_n\}$ such that $x_n \downarrow x$. However, this is obvious since

$$F(x_n) - F(x) = \mathsf{P}(x, x_n] \to 0$$

for the same argument as (a), that is, as any real number can be contained to the complement of $(x, x_n]$ for all large enough n. ■

Remark: F is not necessarily left-continuous. For $x_n \uparrow x$, we have $(-\infty, x_n] \uparrow (-\infty, x)$. It thus follows that

$$F(x_n) = \mathsf{P}(-\infty, x_n] \to \mathsf{P}(-\infty, x) \neq F(x)$$

We indeed have

$$F(x) = F(x^-) + \mathsf{P}\{x\}$$

and therefore, F is continuous at x if and only if P does not have any point mass at x.

The indicator function $1\{A\}$, which assigns 1 if A is true and 0 otherwise, will be used repeatedly in the sequel. The letter I is also used instead of the number 1. It can be also written as $1_A\{\omega\}$ to make explicit the argument of the function.

2.2 Random Vectors and Joint Distributions

Often, we think of more than two random variables as components of a random vector $(X_1, X_2, ..., X_n)$ taking values in $\mathcal{R}^n$, rather than being unrelated random variables each taking values in $\mathcal{R}$.

Let $X_1, \ldots, X_n$ be random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We now define a *random vector* X by

$$X(\omega) = \begin{pmatrix} X_1(\omega) \\ \vdots \\ X_n(\omega) \end{pmatrix}$$

An n-dimensional random vector is thus just an n-tuple of random variables. We may of course view X as a measurable function from Ω to $\mathcal{R}^n$.

Examples:

(a) Consider the random experiment of flipping a coin. Denote by X_1 a random variable given by $X_1(H) = 1$ and $X_1(T) = 0$. Also, let X_2 be another random variable given by $X_2(H) = 0$ and $X_2(T) = 1$. Define a random vector X by $X = (X_1, X_2)'$. We may view X as a function from from Ω to $\mathcal{R}^2$, since

$$X(H) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad X(T) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

(b) Look at the random experiment of tossing two coins. Let X_1 be a random variable taking value 1 if the first coin is flipped down with head and 0 otherwise. Similarly, denote by X_2 the random variable assigning value 1 if the second coin is flipped down with head and 0 otherwise. The random vector $X = (X_1, X_2)'$ is then given by

$$X(HH) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \; X(HT) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \; X(TH) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; X(TT) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

(c) A series of economic time series such as GDP series may be viewed as a random vector.

The distribution of a random vector is defined similarly as for the case of a random variable. The distribution P_X of an n-dimensional random vector $X =$

$(X_1, \ldots, X_n)'$ is a probability measure on $\mathcal{R}^n$, which is given by

$$\mathsf{P}_X(A) = \mathbf{P}\{\omega |\, X(\omega) \in A\}$$

for $A \in \mathcal{B}(\mathcal{R}^n)$. The distribution of a random vector X is often called the *joint distribution* of random variables $X_1, \ldots, X_n$. This is to make it explicit that more than one random variable are involved. The distribution of a subvector of X is called the *marginal* distribution in this context.

For an n-dimensional random vector $X = (X_1, \cdots, X_n)$, the distribution function F_X is defined by

$$F_X(x_1, \cdots, x_n) = \mathbf{P}\{\omega |\, X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\}$$

It is thus a real valued function on $\mathcal{R}^n$. The distribution function of a random vector X is often called the *joint distribution function* of the component random variables $X_1, \ldots, X_n$.

2.3 Density

The existence of a density is provided formally by the following theorem.

Theorem 2.4. *Let μ and ν be two nonnegative measures on a measure space $(M, \mathcal{M})$. If ν is absolutely continuous with respect to μ, then ν can be represented as*

$$\nu(A) = \int_A f \, d\mu$$

for a measurable f.

The function f is called the *Radon-Nikodym derivative* or *density* of ν with respect to μ.

In probability, $\nu(A) = \mathsf{P}(A)$ or $\nu(A) = F_X(x)$ for $A = (-\infty, x)$. If the distribution of a random variable X is absolutely continuous with respect to the Lebesque measure, then it has a density with respect to the Lebesque measure, which is often called *probability density function (pdf).* That is, the pdf f_X of a random variable X is the function that satisfies

$$F_X\left(x\right) = \int_{-\infty}^{x} f_X\left(s\right) ds \text{ for all } x \in \mathcal{R},$$

and thus, $\mathsf{P}\left(A\right) = \int_A f\left(s\right) ds$ for any $A \in \mathcal{B}$. And such a random variable is called *continuous.* On the other hand, a distribution has a density with respect to the counting measure, it is called *probability mess function (pmf).* Such a random variable is *discrete.* In this case,

$$\mathsf{P}\left(A\right) = \sum_{x \in A} f_X\left(x\right).$$

2.4 Independence

Let X be a random variable. We define

Definition 2.5. The σ-field $\sigma(X)$ generated by X is given by

$$\sigma(X) = \{X^{-1}(A) \,|\, A \in \mathcal{B}(\mathcal{R})\}$$

Remarks

(a) It is easy to check that $\sigma(X)$ is indeed a σ-field.

(b) The σ-field $\sigma(X)$ is the smallest σ-field that makes X measurable.

(c) The σ-field generated by a random vector $X = (X_1, \cdots, X_n)$ is defined similarly, i.e., it is given by

$$\sigma(X) = \sigma(X_1, \ldots, X_n) = \{X^{-1}(A) \,|\, A \in \mathcal{B}(\mathcal{R}^n)\}$$

(d) Intuitively, $\sigma(X)$ is precisely the collection of events E such that, for a given ω, we can tell whether or not $\omega \in E$ solely on the basis of the value $X(\omega)$.

Example Let E be an event and define a random variable by $X = I(E)$. The σ-field $\sigma(X)$ generated by X would then be given by

$$\sigma(X) = \{\emptyset, \Omega, E, E^c\}$$

Obviously, given the value $X(\omega)$, it can be unambiguously determined whether or not ω is in any of the sets in $\sigma(X)$.

Definition 2.6. Sub-σ-fields $\mathcal{F}_1, \mathcal{F}_2, \ldots$ of $\mathcal{F}$ are said to be independent if for any index set $\{i_1, ..., i_n\}$ and any $E_{i_k} \in \mathcal{F}_{i_k}$,

$$\mathbf{P}\left(\bigcap_{k=1}^{n} E_{i_k}\right) = \prod_{k=1}^{n} \mathbf{P}(E_{i_k}).$$

Random variables $X_1, X_2, \ldots$ are independent if the σ-fields $\sigma(X_1), \sigma(X_2), \ldots$ generated by them are independent.

By the abuse of notation, we use $p(x_{i_1}, \ldots, x_{i_n})$ to denote the joint density of the random variables $X_{i_1}, \ldots, X_{i_n}$. We may show the *factorization* theorem

Theorem 2.7. *The random variables $X_1, X_2, \ldots$ are independent if and only if*

$$p(x_{i_1}, \ldots, x_{i_n}) = \prod_{k=1}^{n} p(x_{i_k}).$$

The same can be said for the distribution function.

Proof: We only show the sufficiency. Also, we only consider the case of two continuous random variables X and Y with continuous $p()$. Let $Z = (X, Y)'$ be a two dimensional random vector and X be independent of Y. Then, for any $A, B \in \mathcal{B}(\mathcal{R})$, we get

$$\begin{aligned}
\int_{A\times B} p\,(x,y)\,dxdy = \mathsf{P}_Z(A \times B) &= \mathbf{P}\left(Z^{-1}(A \times B)\right) \\
&= \mathbf{P}\left(X^{-1}(A) \bigcap Y^{-1}(B)\right) \\
&= \mathbf{P}\left(X^{-1}(A)\right)\mathbf{P}\left(Y^{-1}(B)\right) \\
&= \mathsf{P}_X(A)\mathsf{P}_Y(B) \\
&= \int_{A\times B} p\,(x)\,p\,(y)\,dxdy.
\end{aligned}$$

Since A and B are arbitrary, $p(x, y) = p(x)\, p(y)$. ■

cf. Note that $p(x) = \int_{-\infty}^{\infty} p(x, y)\, dy$ since

$$P(A) = \int_A p(x)\, dx = \int_A \int_{-\infty}^{\infty} p(x, y)\, dy dx$$

for any $A \in \mathcal{B}(\mathcal{R})$.

3 Expectations

3.1 Expectation

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We define the *expectation* $\mathbf{E}(X)$ of X by

$$\mathbf{E}(X) = \int X \, d\mathbf{P}.$$

More generally, define for $g(X) = g \circ X$ with a measurable $g : \mathcal{R} \to \mathcal{R}$

$$\mathbf{E}g(X) = \int g(X) \, d\mathbf{P}$$

to be the expectation of a random variable $g(X)$. It is straightforward to show that $g(X)$ is indeed a random variable. Note for any $A \in \mathcal{B}(\mathcal{R})$ that

$$(g \circ X)^{-1}(A) = X^{-1}\left(g^{-1}(A)\right) \in \mathcal{F}$$

since g is assumed to be measurable and $g^{-1}(A) \in \mathcal{B}(\mathcal{R})$.

The following theorem tells us how to compute expectations. That is, it allows us to use the Riemann integral instead of the Lebesgue integral.

Theorem 3.1.

$$\int g(X)\,d\mathbf{P} = \int g\,d\mathsf{P}_X = \int g p_X\,d\mu.$$

That is, $\mathbf{E}\left(g\left(X\right)\right) = \int g\left(x\right) p_X\left(x\right) dx$ whenever it is defined.

All the properties of the integral, including linearity, apply to the expectation operation.

Expectations of some special functions have special names and meaning. In particular, we call

$$\mu = \mathbf{E}(X) \quad \text{and} \quad \sigma^2 = \text{var}(X) = \mathbf{E}(X-\mu)^2$$

the *mean* and the *variance*, respectively, of the random variable X. Moreover, if we let X and Y be two random variables with means μ_x and μ_y, then

$$\sigma_{xy} = \text{cov}(X,Y) = \mathbf{E}(X-\mu_x)(Y-\mu_y) \;\; \text{and} \;\; \rho_{xy} = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$$

are called, respectively, *covariance* and *correlation* of X and Y.

Remarks The following are their properties.

(a) var$(aX) = a^2$var(X) and var$(X + b) =$ var(X).

(b) cov$(a_1X_1 + a_2X_2, Y) = a_1cov(X_1, Y) + a_2cov(X_2, Y)$. Notice also that

$$\text{cov}(X, b_1Y_1 + b_2Y_2) = b_1\text{cov}(X, Y_1) + b_2\text{cov}(X, Y_2)$$

due to symmetry cov$(X, Y) =$ cov(Y, X).

(c) If X and Y are independent, cov$\,(X, Y) = 0$.

3.2 Expectational Inequalities

In this section, we derive some useful expectational inequalities. Again, we let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Theorem 3.2. *Let* $\varepsilon > 0$*. Then*

$$\mathbf{P}\{|X| \geq \varepsilon\} \leq \frac{\mathbf{E}|X|^k}{\varepsilon^k}.$$

Proof Notice that

$$\varepsilon^k I\{|X| \geq \varepsilon\} \leq |X|^k$$

Take expectation on both sides and notice that $\mathbf{E}I(x \in A) = \mathbf{P}(A)$ from the Radon-Nikodym Theorem to get the stated result. ■

Remark We have as a special case of the inequality

$$\mathbf{P}\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

where μ and σ^2 are the mean and variance of X. The inequality, which goes under the name Chebyshev, thus yields an upper bound for tail probabilities of a random variable with finite variance.

Let X and Y be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then we have

Theorem 3.3.

$$(\mathbf{E}XY)^2 \leq (\mathbf{E}X^2)(\mathbf{E}Y^2)$$

Proof Write

$$Y = \frac{\mathbf{E}XY}{\mathbf{E}X^2} X + \left(Y - \frac{\mathbf{E}XY}{\mathbf{E}X^2} X \right)$$

Note that the first term represents an orthogonal projection of Y on the range of X and the second term represents the orthogonal projection of Y on the null space of X. Squaring and taking expectations on both sides yield

$$\mathbf{E}Y^2 = \mathbf{E} \left(\frac{\mathbf{E}XY}{\mathbf{E}X^2} X \right)^2 + \mathbf{E} \left(Y - \frac{\mathbf{E}XY}{\mathbf{E}X^2} X \right)^2$$

Notice that the expectation of the cross product term vanishes. We have that

$$\mathbf{E}Y^2 \geq \frac{(\mathbf{E}XY)^2}{\mathbf{E}X^2}$$

from which the stated result follows directly. It is obvious that the equality holds when and only when Y is a constant multiple of X. ■

Remark If we apply Cauchy-Schwarz to two random variables $X - \mu_x$ and $Y - \mu_y$ centered around their means, then

$$\mathrm{cov}(X, Y)^2 \;\leq\; \mathrm{var}(X)\mathrm{var}(Y)$$

It follows in particular that $|corr(X, Y)| \leq 1$ with equality when and only when Y is a linear function of X.

Recall that a set $A \subset \mathcal{R}^n$ is said to be *convex* if $\alpha x + (1 - \alpha)y \in A$ for any $x, y \in A$ and $\alpha \in [0, 1]$. A function $f : \mathcal{R} \to \mathcal{R}$ is called *convex* if the set $\{(x, y) |\, y \geq f(x)\} \subset \mathcal{R}^2$ is convex.

Theorem 3.4. *Let $f : \mathcal{R} \to \mathcal{R}$ be convex. Then*

$$f(\mathbf{E}X) \leq \mathbf{E}f(X)$$

Proof Since $f : \mathcal{R} \to \mathcal{R}$ is convex, there exists a linear function $\ell : \mathcal{R} \to \mathcal{R}$ such that

$$\ell \leq f \quad \text{and} \quad \ell(\mathbf{E}X) = f(\mathbf{E}X)$$

It follows that

$$\begin{aligned}\mathbf{E}f(X) &\geq \mathbf{E}\ell(X)\\ &= \ell(\mathbf{E}X)\\ &= f(\mathbf{E}X)\end{aligned}$$

as was to be shown. ■

Remarks

(a) Functions such as $f(x) = |x|,\ x^2,\ e^x$ are convex.

(b) We call a function $f : \mathcal{R} \to \mathcal{R}$ is *concave* if the set $\{(x, y) : y \leq f(x)\} \subset \mathcal{R}^2$ is convex. Obviously, the inequality is reversed for a concave function such as $f(x) = \log x$.

3.3 Functions

Moments Let X be a random variable. We define the *k-th moment* μ_k and the *k-th central moment* μ_k^* respectively by

$$\mu_k = \mathbf{E}(X^k) \quad \text{and} \quad \mu_k^* = \mathbf{E}\Big(X - \mathbf{E}(X)\Big)^k$$

Proposition 1. *We have*
(a) If $\mu_p < \infty$, then $\mu_q < \infty$ for all $q \leq p$.
(b) $\mu_k < \infty$ if and only if $\mu_k^* < \infty$.

Proof We have from Jensen's inequality

$$(\mathbf{E}|X|^q)^{p/q} \;\leq\; \mathbf{E}|X|^p \;<\; \infty$$

for all $q \leq p$, since $f(x) = x^{p/q}$ is convex. This proves the part (a). The part (b) is immediate, upon noticing that μ_k^* is the expected value of a k-th polynomial in X. ∎

For a random vector $X = (X_1, \ldots, X_n)'$, we define

$$\mathbf{E}(X) = \begin{pmatrix} \mathbf{E}(X_1) \\ \vdots \\ \mathbf{E}(X_n) \end{pmatrix} \quad \text{and} \quad \text{var}(X) = \mathbf{E}\left(X - \mathbf{E}(X)\right)\left(X - \mathbf{E}(X)\right)'$$

Moment Generating Function Let X be a random variable with density p. The *moment generating function* of X is defined as

$$m(t) = \mathbf{E}\Big(e^{tX}\Big) = \int_{-\infty}^{\infty} e^{tx} p(x)\, d\mu(x)$$

Remarks

1. The moment generating function is the *Laplace transform* of the density.

2. One may easily see that

$$\frac{d^k}{dt^k}\, m(0) = \mathbf{E}(X^k)$$

whenever differentiation and integration are interchangeable, to which the name of moment generating function is due.

3. The moment generating function of $\mathbf{N}\left(\mu, \sigma^2\right)$ is given by $\exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$, for instance.

Characteristic Function The *characteristic function* of a random variable X is defined by

$$\varphi(t) = \mathbf{E}\Big(e^{itX}\Big) = \int_{-\infty}^{\infty} e^{itx} p(x) d\mu(x)$$

Remarks

1. The characteristic function is the *Fourier transform* of the density.

2. The moment generating function does not exist in some cases. However, one may easily see that the characteristic function always exists because $e^{itx} = \cos x + i \sin x$.

3.4 Conditional Expectation

Conditional expectation plays a central role in econometrics since it is the best predictor under certain criterion and many economic models builds on the concept of rational expectation. We give the formal definition and explore its key properties.

Let X and Y be random variables defined on a commom probability space with $\mathbf{E}(X) < \infty$ and $\sigma(Y)$ denote the generated σ-field of Y. We define

Definition 3.5. The conditional expectation $\mathbf{E}(X|Y)$ of X given Y is a random variable such that
(a) $\mathbf{E}(X|Y)$ is a function of Y,
(b) For every $F \in \sigma(Y)$,

$$\int_F \mathbf{E}(X|Y)\, d\mathbf{P} = \int_F X\, d\mathbf{P}$$

The conditional expectation is often defined as $\mathbf{E}\left(X|\mathcal{F}\right)$ for a given σ-field $\mathcal{F}$. What matters here is the information structure revealed by the conditioning random variable, in other words, the σ-field.

To help grasp the meaning of the definition, we begin with a discrete random variable $Y = \sum_{k=1}^{n} y_k 1\left\{F_k\right\}$, where $\{F_k\}$ is a partition of Ω. Then, condition (a) requires that $\mathbf{E}(X|Y)$ must be constant over each of F_k's in the partition, i.e., it must be given as

$$\mathbf{E}(X|Y) = \sum c_k \mathsf{I}(F_k) \text{ or } \mathbf{E}(X|Y) = \left(\begin{array}{c} c_1 \text{ if } F_1 \\ c_2 \text{ if } F_2 \\ \vdots \\ c_n \text{ if } F_n \end{array}\right)$$

with some constants c_k's. In other words, $\mathbf{E}\left(X|Y\right)$ is defined on each y_k that Y can assume as it is a function of Y.

Condition (b) shows how to calculate c_k's. In particular,

$$\int_{F_k} \mathbf{E}(X|Y) d\mathbf{P} = c_k \int_{F_k} d\mathbf{P} = c_k \mathbf{P}(F_k) = \int_{F_k} X \; d\mathbf{P}$$

for all k. We thus have

$$c_k = \frac{1}{\mathbf{P}(F_k)} \int_{F_k} X \, d\mathbf{P}$$

which can be thought of as the average of X over F_k weighed by the probability $\mathbf{P}\left(F_k\right)$.

The conditional expectation $\mathbf{E}(X|Y)$ may therefore be viewed as a random variable taking values that are local averages of X over the partitions made by $\mathcal{F}$.

Example Let E and F be two events such that

$$\mathbf{P}(E) = \mathbf{P}(F) = \frac{1}{2} \quad \text{and} \quad \mathbf{P}(E \cap F) = \frac{1}{3}$$

Define $X = I(E), Y = I(F)$ and let $\mathcal{F} = \sigma(Y) = \{\emptyset, \Omega, F, F^c\}$. On F, X takes on 1 if $\omega \in E \cap F$ and 0 otherwise. Then, $\int_F X d\mathbf{P} = \mathbf{P}\left(E \cap F\right) = 1/3$. Similarly, we can compute $\int_{F^c} X d\mathbf{P} = \mathbf{P}\left(E \cap F^c\right) = 1/2 - 1/3 = 1/6$. The conditional expectation $\mathbf{E}(X|Y)$ is therefore given by

$$\mathbf{E}(X|Y) = \frac{2}{3} I(F) + \frac{1}{3} I(F^c) \text{ or } \frac{2}{3} Y + \frac{1}{3}(1 - Y) = \frac{1}{3}(Y + 1) \, .$$

Now consider continuous random variables X and Y that have joint density $p(x, y)$. In this case, this method does not apply since $\mathbf{P}(Y = c) = 0$ for any $c \in \mathbb{R}$. First define the conditional density

$$p(x|y) = \frac{p(x, y)}{p(y)},$$

and then

$$\mathbf{E}(X|Y = y) = \int xp(x|y)\, dx,$$

which is traditionally defined as the conditional expectation of X given $Y = y$. The integral yields a function of y, which we denote by $f(y)$, i.e.,

$$f(y) = \mathbf{E}(X|Y = y).$$

Thus, we see again that $\mathbf{E}(X|Y)$ is a random variable as it is a transformation of Y. In addition, the conditional expectation for a fixed y is an expectation as the conditional distribution is a probability measure.

For each $F \in \sigma(Y)$, there exists $B \in \mathcal{B}(\mathcal{R})$ such that $F = Y^{-1}(B)$. Therefore, we have

$$\begin{aligned}\int_F f(Y)\,d\mathbf{P} &= \int_B f(y)p(y)\,dy\\ &= \int_B \left(\int_{\mathcal{R}} xp(x|y)\,dx\right) p(y)\,dy\\ &= \iint_{\mathcal{R}\times B} xp(x,y)\,dx\,dy\\ &= \int_F X\,d\mathbf{P}\end{aligned}$$

for all $F \in \sigma(Y)$. Note that the inverse image of $\mathcal{R} \times B$ through the mapping (X, Y) is F. It follows that

$$f(Y) = \mathbf{E}(X|Y).$$

Indeed, it satisfies the conditions of our definition.

Example Let

$$p(x,y) = (x+y)\,I\{0 \le x, y \le 1\}$$

Then, for $0 \leq y \leq 1$,

$$\begin{aligned}\mathbf{E}(X|Y=y) &= \int xp(x|y)dx \\ &= \int_0^1 x\,\frac{x+y}{\frac{1}{2}+y}\,dx \\ &= \frac{\frac{1}{3}+\frac{y}{2}}{\frac{1}{2}+y}\end{aligned}$$

We now have

$$\mathbf{E}(X|Y) = \frac{\frac{1}{3}+\frac{Y}{2}}{\frac{1}{2}+Y}$$

Properties We now list some of the useful properties of conditional expectation.

(a) *Law of Iterated Expectation* $\mathbf{E}\Big(\mathbf{E}(X|Y)\Big) = \mathbf{E}(X)$, which is straightforward from the definition of conditional expectation with $F = \Omega$.

(b) For a function g, $\mathbf{E}\left(g\left(X,Y\right)|Y=y\right) = \mathbf{E}\left(g\left(X,y\right)|Y=y\right)$.

(c) *Linearity* $\mathbf{E}(\alpha X + \beta Y|Z) = \alpha \mathbf{E}(X|Z) + \beta \mathbf{E}(Y|Z)$, which is obvious.

(d) *Conditional Variance Identity*

$$\text{var}\,(Y) = \mathbf{E}\,[\text{var}\,(Y|X)] + \text{var}\,[\mathbf{E}\,(Y|X)]\,,$$

which follows immediately from the projection theory.

4 Functions of Random Variables

Let the distribution of a random variable or a random vector X be known. We show in this section how to obtain the distribution of a random variable Y defined by $Y = f(X)$ for a measurable f. There are three commonly used methods: *distribution function technique, moment generating function technique*, and *transformation technique*.

Distribution Function Technique It is sometimes possible to compute the distribution function F_Y of Y directly from

$$F_Y(y) = \mathbf{P}\{Y \leq y\} = \mathbf{P}\{f(X) \leq y\}.$$

Examples:

(a) Let $X \sim U[0,1]$ and $Y = -\log X$. Then for $y \geq 0$

$$\begin{aligned} F_Y(y) &= \mathbf{P}\{-\log X \leq y\} \\ &= \mathbf{P}\{X \geq e^{-y}\} \\ &= 1 - e^{-y}. \end{aligned}$$

(b) Let $X_1, \ldots, X_n$ be i.i.d. random variables with common distribution function F. Define $Y = \max\{X_1, \ldots, X_n\}$. Then the distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= \mathbf{P}\left(\bigcap_{i=1}^{n}\{X_i \leq y\}\right) \\ &= \prod_{i=1}^{n} \mathbf{P}\{X_i \leq y\} \\ &= \Big(F(y)\Big)^n \end{aligned}$$

(c) Let $X = (X_1, X_2)'$ be a random vector with joint distribution P and joint density p. Then the distribution function of $Y = X_1 + X_2$ is given by

$$\begin{aligned} F_Y(y) &= \mathbf{P}\{X_1 + X_2 \leq y\} \\ &= \mathsf{P}\{(x_1, x_2)|\, x_1 + x_2 \leq y\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{y-x_2} p(x_1, x_2) dx_1 dx_2 \end{aligned}$$

which can easily be obtained if p is specified.

Moment Generating Function Technique The distribution of sums of *independent* random variables can readily be obtained via moment generating function. If we let $X_1, \ldots, X_n$ be independent random variables with moment generating function m_i, $i = 1, \ldots, n$, and $Y = X_1 + \cdots + X_n$, then the moment generating function m of Y is given by

$$m_Y(t) = \mathbf{E}\Big(e^{t(X_1+\cdots+X_n)}\Big) = \prod_{i=1}^{n} m_i(t)$$

A generalization can be done easily. Let $Z = (a_1X_1 + b_1) + \cdots + (a_nX_n + b_n)$ where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be fixed constants. Then the moment generating function m of Z is

$$m_Z(t) = \mathbf{E}\Big(e^{t(a_1X_1+\cdots+a_nX_n)+t(b_1+\cdots+b_n)}\Big) = e^{t\sum_{i=1}^{n} b_i} \prod_{i=1}^{n} m_i(a_i t)$$

Examples:

(a) Let $X_i \sim$ Poisson(λ_i), $i = 1, \ldots, n$, be *independent*. Then the moment generating function of $Y = X_1 + \cdots + X_n$ is

$$m(t) = \prod_{i=1}^{n} \text{exp}\left(\lambda_i(e^t - 1)\right) = \text{exp}\left((e^t - 1)\sum_{i=1}^{n} \lambda_i\right)$$

from which we may deduce that Y has Poisson distribution with parameter $\sum \lambda_i$. cf. $p_\lambda(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0, 1, 2, ...$

(b) Let $X_i \sim \mathbf{N}(\mu_i, \sigma_i^2)$ be independent. The moment generating function of $Y = c_1X_1 + \cdots + c_nX_n$ is given by

$$\begin{aligned} m(t) &= \prod_{i=1}^{n} \text{exp}\left(c_i\mu_i t + \frac{c_i^2\sigma_i^2 t^2}{2}\right) \\ &= \text{exp}\left(t\sum_{i=1}^{n} c_i\mu_i + \frac{t^2 \sum_{i=1}^{n} c_i^2\sigma_i^2}{2}\right) \end{aligned}$$

which implies that

$$Y \sim \mathbf{N}\left(\sum_{i=1}^{n} c_i\mu_i, \sum_{i=1}^{n} c_i^2\sigma_i^2\right).$$

Transformation Technique If the function f is one-to-one, we may find the density of Y from that of X by the transformation technique. Denote by P_X and P_Y the distributions of X and Y, respectively. Furthermore, we assume that P_X and P_Y have densities p_X and p_Y with respect to the Lebesque measure. Let $A \in \mathcal{B}(\mathcal{R})$ and $B = f(A)$. We have

$$\begin{aligned} P_Y(B) &= P_X(A) \\ &= \int_A p_X(x)\,dx \\ &= \int_B p_X(f^{-1}(y)) \left| \frac{1}{f'(f^{-1}(y))} \right| dy \end{aligned}$$

where the last equality follows by the change of variable $x = f^{-1}(y)$. We now deduce that

$$p_Y(y) = p_X\Big(f^{-1}(y)\Big) \left| \frac{1}{f'\Big(f^{-1}(y)\Big)} \right|$$

The term

$$\left|\frac{1}{f'\Big(f^{-1}(y)\Big)}\right|$$

here is often called the *Jacobian of transformation*. More gerenally, when $f : \mathcal{R}^n \to \mathcal{R}^n$, the Jacobian is given by

$$\left| det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \\ \frac{\partial x_n}{\partial y_1} & & \frac{\partial x_n}{\partial y_n} \end{pmatrix} \right| .$$

Example The joint density of X_1 and X_2 is given by

$$\begin{aligned} p(x_1, x_2) &= 4x_1x_2 \text{ if } 0 < x_1, x_2 < 1, \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Let $Y_1 = X_1/X_2$ and $Y_2 = X_1X_2$. Now we will obtain the joint density of Y_1 and Y_2. Substituting $x_1 = \sqrt{y_1y_2}$ and $x_2 = \sqrt{y_2/y_1}$ into $p(x_1, x_2)$ and using the

Jacobian of the transformation $|J| = \frac{1}{2y_1}$ which is computed by

$$\begin{aligned} J &= \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix} \\ &= \det \begin{pmatrix} \frac{y_2}{2\sqrt{y_1 y_2}} & \frac{y_1}{2\sqrt{y_1 y_2}} \\ -\frac{1}{2}\frac{\sqrt{y_1 y_2}}{y_1^2} & \frac{1}{2\sqrt{y_1 y_2}} \end{pmatrix} \\ &= \frac{1}{2y_1} \end{aligned}$$

so that the joint density of Y_1 and Y_2 is

$$p(y_1, y_2) = \frac{2y_2}{y_1}$$

for $0 < y_1$ and $0 < y_2 < \left(y_1 \wedge y_1^{-1}\right)$.

5 Multivariate Normal Distribution

5.1 Introduction

An n-dimensional random vector X is said to have (multivariate) normal distribution with parameters μ and $\mathbf{\Sigma}$, and denoted by $X \sim \mathbf{N}(\mu, \mathbf{\Sigma})$ (or $\mathbf{N}_n(\mu, \mathbf{\Sigma})$ to emphasize the dimension of X), if it has probability density

$$p(x) = \frac{1}{(2\pi)^{n/2}} (\det \mathbf{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)' \mathbf{\Sigma}^{-1} (x-\mu) \right)$$

where $\mu \in \mathcal{R}^n$ and $\mathbf{\Sigma}$ is an $n \times n$ positive definite matrix.

Let Z be an n-vector of independent standard normal variates. Clearly, $Z \sim \mathbf{N}_n(0, I)$. The following lemma shows that any normal random vector can be written as a linear transformation of Z.

Lemma 5.1. *Let Z be defined as above and*

$$X = \mu + \mathbf{\Sigma}^{1/2} Z$$

Then

$$X \sim \mathbf{N}(\mu, \mathbf{\Sigma})$$

Proof Notice that the density of Z is given by

$$p(z) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}z'z\right)$$

and the Jacobian of the transformation $z \mapsto x = \mu + \Sigma^{1/2}z$ is

$$(\det \Sigma)^{-1/2}$$

The stated result then follows immediately. ■

The representation of a normal vector as an affine function of a random vector consisting of independent standard normals is very useful. It is indeed straightforward from such representation that

Corollary 2. *Let $X \sim \mathrm{N}(\mu, \Sigma)$. Then*

$$\mathbf{E}\,(X) = \mu \quad \text{and} \quad \mathrm{var}\,(X) = \Sigma$$

Proof Due to Lemma 5.1, we can write $X = \mu + \Sigma^{1/2}Z$, where Z is a random vector of independent standard normals. We have

$$\begin{aligned} \mathbf{E}\,(X) = \mu + \Sigma^{1/2}\mathbf{E}(Z) &= \mu \\ \mathrm{var}\,(X) = \Sigma^{1/2}\mathrm{var}(Z)\,\Sigma^{1/2} &= \Sigma \end{aligned}$$

as was to be shown. ■

We may also easily derive the moment generating function of multivariate normal distribution, as shown below.

Corollary 3. *Let $X \sim \mathbf{N}(\mu, \Sigma)$. Then the moment generating function m_x of X is given by*

$$m_x(t) = \mathbf{E}\left(e^{t'X}\right) = \exp\left(\mu' t + \frac{1}{2} t' \mathbf{\Sigma} t\right).$$

Proof Let Z be defined as above, and denote by m_z the moment generating function of Z. Also, let

$$t = (t_1, \ldots, t_n)' \quad \text{and} \quad Z = (Z_1, \ldots, Z_n)'$$

From the independence of $Z_1, \ldots, Z_n$, we have

$$\begin{aligned} m_z(t) &= \mathbf{E}\left(e^{t'Z}\right) \\ &= \mathbf{E}\left(e^{t_1 Z_1}\right) \cdots \mathbf{E}\left(e^{t_n Z_n}\right) \\ &= \exp\left(\frac{1}{2} t' t\right) \end{aligned}$$

It follows that

$$\begin{aligned} m_x(t) &= \mathbf{E}\left(e^{t'X}\right) \\ &= e^{t'\mu} m_z(\boldsymbol{\Sigma}^{1/2} t) \\ &= \exp\left(t'\mu + \frac{1}{2} t'\boldsymbol{\Sigma} t\right) \end{aligned}$$

as is required to be shown. ■

Remarks

We have thus far assumed that $\boldsymbol{\Sigma}$ is nonsingular. If $\boldsymbol{\Sigma}$ is singular, then there exists some $c \in \mathcal{R}^n$ such that

$$\mathrm{var}\,(c'X) = c'\boldsymbol{\Sigma}\, c = 0$$

This implies that the distribution of X is concentrated on a subset of $\mathcal{R}^n$. We often say that the distribution of X is degenerate in this case.

5.2 Marginal and Conditional Distributions

Throughout this section, we let $X \sim \mathrm{N}(\mu, \Sigma)$. Partition X as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

where X_1 and X_2 are, respectively, n_1- and n_2-dimensional. Let mean and variance of X be partitioned conformably as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

The following lemma shows that any affine transformation of a multivariate normal variate is normal.

Lemma 5.2. *If $Y = AX + b$, then $Y \sim \mathrm{N}(A\mu + b, A\Sigma A')$.*

Proof The moment generating function of Y is given by

$$\begin{aligned} m_y(t) &= \mathbf{E}\left(e^{t'Y}\right) \\ &= e^{t'b}\mathbf{E}\left(e^{t'AX}\right) \\ &= e^{t'b}m_x(A't) \\ &= e^{t'b}\exp\left(t'A\mu + \frac{1}{2}t'A\Sigma A't\right) \\ &= \exp\left(t'(A\mu + b) + \frac{1}{2}t'A\Sigma A't\right) \end{aligned}$$

which shows that the distribution of Y is multivariate normal with given mean and variance. ■

We have as a special case that

Corollary 4.

$$X_1 \sim \mathbf{N}(\mu_1, \Sigma_{11})$$

Proof Apply Lemma 5.2 with a selection matrix $A = (I_{n_1}, 0)$ and $b = 0$. ■

Corollary 4 shows that the marginal distribution of multivariate normal distribution is also normal.

Theorem 5.3. *X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.*

Proof The *only if* part is obvious. To prove *if* part, let $\Sigma_{12} = 0$ and write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

Denote by $p(x_1)$ and $p(x_2)$ the probability densities of X_1 and X_2. It follows that

$$\begin{aligned} p(x) &= (2\pi)^{-n/2}(\det\Sigma)^{-1/2}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right) \\ &= (2\pi)^{-n_1/2}(\det\Sigma_{11})^{-1/2}\exp\left(-\frac{1}{2}(x_1-\mu_1)'\Sigma_{11}^{-1}(x_1-\mu_1)\right) \\ &\cdot (2\pi)^{-n_2/2}(\det\Sigma_{22})^{-1/2}\exp\left(-\frac{1}{2}(x_2-\mu_2)'\Sigma_{22}^{-1}(x_2-\mu_2)\right) \\ &= p(x_1)p(x_2) \end{aligned}$$

and therefore X_1 and X_2 are independent. ■

Theorem 5.4. *The conditional distribution of X_1 given X_2 is*

$$\mathbf{N}(\mu_{1\cdot 2}, \Sigma_{11\cdot 2})$$

where

$$\begin{aligned} \mu_{1\cdot 2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2) \\ \Sigma_{11\cdot 2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

Proof Consider a random vector given by

$$\begin{pmatrix} X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2 \\ X_2 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

which is normal as a linear transformation of a normal random vector X. The two sub-vectors $X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$ and X_2 are uncorrelated, and therefore independent.

Write

$$X_1 = \Sigma_{12}\Sigma_{22}^{-1}X_2 + \left(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2\right).$$

The last term is independent of X_2. Its conditional distribution given X_2 is consequently the same as its unconditional distribution, which is normal with mean and variance

$$\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \quad \text{and} \quad \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

The first term can be treated as constant when X_2 is given. It therefore just shifts the mean of conditional distribution of X_1 given X_2.

■

Remarks

(a) The conditional mean given X_2 is a linear function of X_2.

(b) The conditional variance given X_2 does not depend on X_2 (conditional homoskedasticity).

5.3 Quadratic Forms

We consider the distribution of the quadratic form $X'AX$ in a normal random vector X, defined with nonrandom matrix A. It is well known from elementary statistics that for a vector $Z = (Z_1, \ldots, Z_n)'$ of independent standard normals $Z'Z = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$. This is generalized below.

Proposition 5. *Let $X \sim \mathrm{N}_n(0, \Sigma)$. Then*

$$X'\Sigma^{-1}X \sim \chi_n^2$$

Proof Notice that $\Sigma^{-1/2}X \sim \mathrm{N}(0, I)$. ∎

Moreover, we have

Theorem 5.5. *Let $Z \sim \mathrm{N}_n(0, I)$ and P be an m-dimensional orthogonal projection matrix in $\mathcal{R}^n$. Then we have*

$$Z'PZ \sim \chi_m^2$$

Proof Let

$$P = H_m H_m'$$

where H_m is an orthogonal matrix such that $H_m' H_m = I_m$. Write

$$Z'PZ = (H_m'Z)'(H_m'Z)$$

and observe that $H_m'Z \sim \mathrm{N}_m(0, I)$ to finish the proof. ∎

Theorem 5.6. *Let $Z \sim \mathrm{N}(0, I)$, and let A and B be nonrandom matrices of full column ranks. Then, $A'Z$ and $B'Z$ are independent if and only if $A'B = 0$.*

Proof Let

$$C = (A, B)$$

Clearly, $C'Z$ is multivariate normal with sub-vectors $A'Z$ and $B'Z$, the covariance of which is zero if and only if $A'B = 0$. The stated result follows from Theorem 5.3. ∎

Corollary 6. *Let $Z \sim \mathrm{N}(0, I)$, and let P and Q are orthogonal projections such that $PQ = 0$. Then $Z'PZ$ and $Z'QZ$ are independent.*

Proof Since $Z'PZ = (PZ)'PZ$ and $Z'QZ = (QZ)'QZ$, $Z'PZ$ and $Z'QZ$ are functions, respectively, of PZ and QZ. It therefore suffices to show that PZ and QZ are independent, which follows from $PQ = 0$ and Theorem 5.6. ■

Theorem 5.7. *Let $X \sim \mathrm{N}(0, I)$, and let P and Q be orthogonal projections of dimension p and q, respectively. If $PQ = 0$, then*

$$\frac{\frac{X'PX}{p}}{\frac{X'QX}{q}} \sim F_{p,q}.$$

Proof The stated result follows directly from Theorem 5.5 and Corollary 6, and the definition of F-distribution. ■

5.4 An Important Example

We assume throughout this section that $X_i's$ are independent and identically distributed as $\mathrm{N}(\mu, \sigma^2)$. Define

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

In the following theorem are presented the main results of elementary statistics.

Theorem 5.8. *We have*
(a) $\overline{X}_n \ \sim \ \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$
(b) $\frac{(n-1)S_n^2}{\sigma^2} \ \sim \ \chi^2_{n-1}$
(c) $\overline{X}_n$ *and* S_n^2 *are independent*
(d) $\quad \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \ \sim \ t_{n-1}$.

Proof Homework !! ■

6 Estimation

6.1 Sufficient Statistic

Let $T = \tau(X)$ be a statistic, and $\mathcal{P} = \{\mathsf{P}_\theta | \theta \in \Theta\}$ be a family of *distributions* of X. To avoid confusion over notations, note that P_θ denotes a distribution whereas p_θ denotes a joint density. Now, we define

Definition 6.1. We say that T (or τ) is *sufficient* for $\mathcal{P}$ (or for θ) if the conditional distribution of X given $T = t$ does not depend on θ for all t.

The distribution of X is unknown. It can be any member of the family $\mathcal{P}$. Therefore, the conditional distribution of X given $T = t$ would generally be dependent upon θ. If T is a sufficient statistic, however, it is uniquely determined irrespective of the value of θ.

Remarks If T is sufficient, then we may write

$$p_\theta(x, t) = p(x|t)p_\theta(t)$$

Note that we omit the subscript θ in representing the conditional density $p(x|t)$, since it is independent of θ. We observe that

(a) The information of θ in the observation of X is concentrated in that of T. Usually, T is of lower dimension than X since the former is a function of the latter. Hence, the observation of T is less costly, though it includes the same amount of information on θ. Usefulness of a sufficient statistic lies in such data reduction.

(b) Once we know the value of T, then we may perform a random experiment and define a random variable $\tilde{X}$, say, such that the conditional distribution of $\tilde{X}$ given $T = t$ has density $p(x|t)$. We may design such a random experiment and a random variable, since $p(x|t)$ is independent of θ and known. Observed value $\tilde{x}$, though it is generally different from the original observation x, can obviously be regarded as an observation from the same distribution as X. We may therefore recover data in this sense from the observation of T.

Examples

(a) Suppose that $X \sim \mathrm{N}(0, \sigma^2)$ and $T = |X|$. For $T = t$, X can take only two values t and $-t$. Furthermore, since the distribution of X is symmetric about the origin, each point has conditional probability 1/2. This is so, regardless of the value of σ^2. The statistic T is therefore sufficient. For the data recovery, we now consider a random experiment of flipping a fair coin, and a random variable $\tilde{X}$ that takes value t if head is up and $-t$ if tail is up. The observed value of $\tilde{X}$ can be regarded as from the same distribution as X.

(b) Let X_1 and X_2 be independent Poisson(λ). We will show that $T = X_1 + X_2$ is sufficient. First, the joint density of X_1 and X_2 is

$$p_\lambda(x_1, x_2) = e^{-2\lambda} \frac{\lambda^{x_1+x_2}}{x_1! x_2!} I\{x_1, x_2 = 0, 1, \ldots\},$$

and T is Poisson(2λ). Since

$$\begin{aligned} p_\lambda(x_1, t) &= e^{-2\lambda} \frac{\lambda^t}{x_1!(t - x_1)!} I\{x_1 = 0, \ldots, t, t = 0, 1, 2, \ldots\} \\ p_\lambda(t) &= e^{-2\lambda} \frac{(2\lambda)^t}{t!} I\{t = 0, 1, \ldots\} \end{aligned}$$

the conditional density of X_1 given $T = t$ is given by

$$\begin{aligned} p(x_1|t) &= \frac{t!}{x_1!(t-x_1)!}\left(\frac{1}{2}\right)^t I\{x_1 = 0, \ldots, t\} \\ &= \binom{t}{x_1}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{2}\right)^{t-x_1} I\{x_1 = 0, \ldots, t\} \end{aligned}$$

so that the conditional distribution of X_1 given $T = t$ is Binomial$(t, \frac{1}{2})$. For the data recovery, consider a random experiment of t tosses of a fair coin and define random variables $\tilde{X}_1$ and $\tilde{X}_2$ as the numbers of head and tail, respectively.

(c) Let X_1 and X_2 be independent $\mathrm{N}(\mu, 1)$. Let us verify that $T = X_1 + X_2$ is sufficient. For this purpose, define $S = X_1 - X_2$. Then we can easily see that T and S are independent because the covariance of T and S are zero. Then the conditional distribution of S given T is equal to the distribution of S, and $S \sim \mathrm{N}(0, 2)$. From this we can deduce that the conditional distribution of X_1 and X_2 given T does not depend on μ as X_1 or X_2 is a linear transformation of S given T.

However, for a particular model, it becomes tedious to find a sufficient statistic by using the definition of a sufficient statistic. The following Theorem provides a very convenient way of finding sufficient statistics. Let the distribution of X now be given by a family of *densities*

$$\mathcal{P} = \{p_\theta |\, \theta \in \Theta\}$$

Then

Theorem 6.2. *A statistic $T = \tau(X)$ is sufficient if and only if the joint density of a sample X is factorized as*

$$p_\theta(x) = f(\tau(x), \theta)g(x).$$

Proof The proof of if part involves a higher level of mathematics. But if X is discrete, then for $\tau = t$,

$$p_\theta\left(\tau = t\right) = \sum_{x' : \tau(x) = t} p\left(x'\right) = f\left(t, \theta\right) \sum_{x' : \tau(x) = t} g\left(x'\right)$$

$$p_\theta\left(x|\tau = t\right) = \frac{p_\theta\left(x\right)}{p_\theta\left(\tau = t\right)} = \frac{g\left(x\right)}{\sum_{x' : \tau(x) = t} g\left(x'\right)}$$

The only if part is immediate as $p_\theta\left(x, t\right) = p\left(x|t\right)p_\theta\left(t\right)$, letting $g\left(x\right) = p\left(x|t\right)$ for which we note that t is a function of x. ■

Examples

(a) Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$ so that the joint density is given by

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\mu^2\right) \end{aligned}$$

We can define $g\,(x) = 1$, $\tau(x) = (\tau_1, \tau_2) = \left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$, and

$$f(\tau, \theta) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\tau_2 + \frac{\mu}{\sigma^2}\tau_1 - \frac{n}{2\sigma^2}\mu^2\right).$$

Then, it can be easily seen that

$$p_\theta(x) = f(\tau(x), \theta)g(x).$$

Therefore, it follows that $\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2\right)$ is sufficient for $\theta = (\mu, \sigma^2)'$.

(b) It is always true that the complete sample, X is a sufficient statistic. We can factor the pdf of X as $p_\theta(x) = f(\tau(x), \theta)g(x)$ by defining $\tau(x) = x$ and $g(x) = 1$

for all x.

(c) Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(0, \sigma^2)$. It is easy to show that $\tau(x) = \sum_{i=1}^n x_i^2$ is sufficient for σ^2. However, it is not the only sufficient statistic. The statistics given by

$$\begin{aligned}
\tau_1(x) &= (x_1, \ldots, x_n)' \\
\tau_2(x) &= (x_1^2, \ldots, x_n^2)' \\
\tau_3(x) &= (x_1^2 + \cdots + x_m^2, x_{m+1}^2 + \cdots + x_n^2)'
\end{aligned}$$

are all sufficient. For instance, if we let $T = \tau_1(X)$, then the conditional distribution of X given $T = t$ is t with probability 1. Therefore $T = \tau_1(X)$ is clearly sufficient.

Let T and S be two statistics. The last example demonstrates that

Lemma 6.3. *If $T = \varphi(S)$ for some function φ and T is sufficient, then S is also sufficient.*

Proof Suppose $T = \varphi(S)$. By the Factorization Theorem, $p_\theta(x) = f(\varphi(S), \theta)g(x) = f^*(S, \theta)g(x)$ where $f^*(S, \theta) = f(\varphi(S), \theta)g(x)$. ■

Remark If φ is many-to-one, T provides further reduction of data. Indeed, we say that a sufficient statistic T is *minimal* if it is a function of every sufficient statistic. A minimal sufficient statistic thus achieves the greatest reduction of data.

We now introduce

Definition 6.4. A family $\{\mathsf{P}_\theta | \, \theta \in \Theta\}$ of distributions is said to form an m-parameter *exponential family* if the distributions have densities of the form

$$p_\theta(x) = \exp\left(\sum_{i=1}^{m} f_i(\theta)\tau_i(x) + g(\theta)\right) h(x).$$

The exponential families include many of the distributions that we know such as Poisson(λ) , Bernoulli(p) ,Normal $\left(\mu, \sigma^2\right)$, Gamma (α, β) , and so on. It is easy to find a sufficient statistic for an exponential family of distributions using the Factorization Theorem.

Remarks

(a) Note that for the exponential families

$$\tau(x) = (\tau_1(x), \ldots, \tau_m(x))'$$

is a sufficient statistic, which follows directly from the Factorization Theorem.

(b) If $X_1, \ldots, X_n$ are i.i.d. with density

$$p_\theta(x_i) = \exp\Big(f(\theta)\tau(x_i) + g(\theta)\Big)h(x_i)$$

then the density of $X = (X_1, \ldots, X_n)'$ is

$$p_\theta(x) = \exp\left(f(\theta)\sum_{i=1}^{n}\tau(x_i) + ng(\theta)\right)h(x_1)\cdots h(x_n)$$

from which we deduce that $\sum_{i=1}^{n}\tau(x_i)$ is sufficient. In fact, it is also known to be minimal.

6.2 Sample Analogue Estimation

Let $x = (x_1, \ldots, x_n)'$ be an observation from $X = (X_1, \ldots, X_n)'$. Assume that the random variables $X_1, \ldots, X_n$ are independent and has common underlying distribution, which we parameterize as

$$\mathcal{P} = \{\mathsf{P}_\theta | \, \theta \in \Theta\}$$

The underlying distribution is often called the *population*.

We denote by E_θ the integration with respect to the probability distribution P_θ on $\mathcal{R}$, i.e.,

$$\mathsf{E}_\theta(f) = \int f \, d\mathsf{P}_\theta$$

for a measurable function $f : \mathcal{R} \to \mathcal{R}$. Furthermore, we define P_n be a probability distribution, called the *empirical distribution*, which assigns probability mass $1/n$ on each of points x_i, $i = 1, \ldots, n$. Then let E_n be the integration with respect to P_n on $\mathcal{R}$, i.e.,

$$\mathsf{E}_n(f) = \int f \, d\mathsf{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

again for any measurable function $f : \mathcal{R} \rightarrow \mathcal{R}$. However, more often, the empirical distribution may not be satisfactory, because it assigns the same weight on each sample points.

Estimator Suppose

$$\pi = \mathsf{E}_\theta(f)$$

is the parameter of interest. The sample analogue principle suggests that we estimate π by

$$\hat{\pi} = \mathsf{E}_n(f)$$

Loosely speaking, the principle proposes to estimate a population characteristic by its sample analogue.

Remark If $f(x) = x^k$ and π is a moment of the underlying distribution, then the method reduces to what is known as the method of moments.

Examples

(a) Let X_i be i.i.d. Poisson(λ). To get the sample analogue estimator $\hat{\lambda}$ of λ, we note that

$$\mathsf{E}_\lambda(f) = \lambda$$

for $f(x) = x$. It therefore follows that

$$\hat{\lambda} = \mathsf{E}_n(f) = \bar{X}$$

(b) Let X_i be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. Suppose we want to estimate two parameters μ and σ^2. Notice that

$$\mathsf{E}_{\mu,\sigma^2}(f) = \mu \quad \text{and} \quad \mathsf{E}_{\mu,\sigma^2}(g) = \sigma^2$$

for $f(x) = x$ and $g(x) = (x - \mu)^2$. However,

$$\mathsf{E}_n(f) = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \mathsf{E}_n(g) \ = \ \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

The sample analogue estimators for μ and σ^2 are therefore given by

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

6.3 Maximum Likelihood Estimation

As before, let $x = (x_1, \dots, x_n)'$ be an observation from $X = (X_1, \dots, X_n)'$, whose density is assumed to belong to the family

$$\mathcal{P} = \{p(\cdot, \theta) | \, \theta \in \Theta\}$$

Note that here we use the notation $p(\cdot, \theta)$, instead of $p_\theta(\cdot)$, to denote a member in the class. This is because we now wish to regard a density p also as a function of θ. If a density is thought of as a function of the unknown parameter θ, then it is called the *likelihood function*.

Estimator The *maximum likelihood estimator* (MLE) of θ is defined by

$$\hat{\theta}_{\mathsf{ML}} = \mathsf{argmax}_{\theta \in \Theta} \, p(x, \theta)$$

Remarks

(a) Computationally, it is often much easier to maximize the *log-likelihood function*

$$\ell(x, \theta) = \log \, p(x, \theta)$$

which is legitimate since log function is monotone increasing.
(b) Usually, the function $\ell(x, \cdot)$ is differentiable and globally concave for every x. The maximizer can therefore be found simply by solving the first-order condition (FOC)

$$\frac{\partial}{\partial \theta} \ell(x, \theta) = 0$$

for θ in terms of x.
(c) The MLE of a function of θ, say $\pi = f(\theta)$, is given by

$$\hat{\pi}_{\text{ML}} = f\left(\hat{\theta}_{\text{ML}}\right)$$

since any other value of π results in values of θ different from $\hat{\theta}_{\text{ML}}$ which yield smaller likelihood, in terms of the induced likelihood. It can, in particular, be said that the ML estimation is invariant with respect to reparametrization.

Examples
(a) Let X_i, $i = 1, \ldots, n$, be i.i.d. $\mathrm{N}(\mu, \sigma^2)$. Then the log-likelihood function is given by

$$\ell(x, \mu, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

and solving the FOC's yields

$$\hat{\mu}_{\mathsf{ML}} = \bar{x}$$
$$\hat{\sigma}^2_{\mathsf{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

simultaneously.

(b) Let X_i, $i = 1, \ldots, n$, be i.i.d. $U[0, \theta]$. Then

$$\begin{aligned} p(x, \theta) &= \frac{1}{\theta^n}\prod_{i=1}^{n} \mathsf{I}\{0 \leq x_i \leq \theta\} \\ &= \frac{1}{\theta^n}\mathsf{I}\left\{\min_{1\leq i\leq n} x_i \geq 0\right\}\mathsf{I}\left\{\max_{1\leq i\leq n} x_i \leq \theta\right\} \end{aligned}$$

it follows that $\hat{\theta}_{\mathsf{ML}} = \max\{x_1, \ldots, x_n\}$. Note also that $\hat{\theta}_{\mathsf{ML}}$ is a sufficient statistic since $p(x, \theta) = g(x)f(\max x_i, \theta)$ where $g(x) = \mathsf{I}\left\{\min_{1\leq i\leq n} x_i \geq 0\right\}$ and $f(\max x_i, \theta) = \frac{1}{\theta^n}\mathsf{I}\left\{\max_{1\leq i\leq n} x_i \leq \theta\right\}$.

6.4 Uniformly Minimum Variance Unbiased Estimators

As desirable properties for a good estimator, we introduce the concepts of *unbiasedness* and *minimum mean squared error*. We first define unbiasedness. Here and elsewhere, we denote by $\mathbf{P}_\theta$ the probability in Ω. Expectation with respect to $\mathbf{P}_\theta$ is denoted by $\mathbf{E}_\theta$.

Definition 6.5. An estimator $T = \tau(X)$ is called unbiased if

$$\mathbf{E}_\theta(T) = \theta$$

for all $\theta \in \Theta$.

The *mean squared error* (MSE) of an estimator $T = \tau(X)$ can be decomposed as the sum of the variance and the squared bias, as shown below

$$\mathbf{E}_\theta(T-\theta)^2 = \mathbf{E}_\theta\left(T - \mathbf{E}_\theta(T)\right)^2 + \left(\mathbf{E}_\theta(T) - \theta\right)^2 .$$

Remarks

(a) For an unbiased estimator, mean squared error reduces to variance.

(b) MSE is, in general, a function of the unknown parameter θ. An estimator that has the smallest MSE for all $\theta \in \Theta$ does not exist. This is because the trivial estimator $\hat{\theta} = \theta_1$ for some fixed value θ_1 has zero MSE at $\theta = \theta_1$. The MSE of any other estimator is strictly positive at $\theta = \theta_1$, since it takes other values with positive probabilities. For example, the estimator $\hat{\theta} = 1$ cannot be beaten in MSE sense at $\theta = 1$ but is a terrible estimator otherwise.

It is sometimes, if not always, possible to find an estimator with minimum variance for all $\theta \in \Theta$, if we restrict ourselves to the class of unbiased estimators. We are thus led to define

Definition 6.6. An estimator $T_* = \tau_*(X)$ is called a *uniformly minimum variance unbiased (UMVU)* estimator if it satisfies
(a) T_* is unbiased, and
(b) $\mathbf{E}_\theta(T_* - \theta)^2 \leq \mathbf{E}_\theta(T - \theta)^2$ for any unbiased estimator $T = \tau(X)$.

Finding a best unbiased estimator, if any, is not an easy task for an array of reasons including computational difficulty and hence a more comprehensive method is necessary. A UMVU estimator can be obtained from the following information inequality. This is the approach taken with the use of the Cramer-Rao bound.

6.5 Information Inequality

We first define

Definition 6.7. We define
(a) *(score function)* $s(x, \theta) = \frac{\partial}{\partial\theta}\ell(x, \theta)$
(b) *((Fisher) information)* $I(\theta) = \mathbf{E}_\theta\Big(s(X, \theta)s(X, \theta)'\Big)$
(c) *(Hessian)* $H(x, \theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\ell(x, \theta)$
(d) *(expected Hessian)* $H(\theta) = \mathbf{E}_\theta\, H(X, \theta)$

The following set of assumptions are imposed to derive the information inequality.

Assumption 1. *In what follows, we assume*

1. $\frac{\partial}{\partial\theta}\int p(x, \theta)d\mu(x) = \int \frac{\partial}{\partial\theta}p(x, \theta)d\mu(x)$

2. $\frac{\partial^2}{\partial\theta\partial\theta'}\int p(x, \theta)d\mu(x) = \int \frac{\partial^2}{\partial\theta\partial\theta'}p(x, \theta)d\mu(x)$

3. *$\int \tau(x) \frac{\partial}{\partial \theta'} p(x,\theta) d\mu(x) = \frac{\partial}{\partial \theta'} \int \tau(x) p(x,\theta) d\mu(x)$ for density $p(x,\theta)$ and any estimator $\tau(x)$.*

Proposition 7. *Suppose that Assumption 1 hold. Then*

$$\mathbf{E}_\theta \, s(X,\theta) = 0.$$

Proof Notice that

$$\begin{aligned}
\int s(x,\theta) p(x,\theta) d\mu(x) &= \int \frac{\partial}{\partial \theta} \ell(x,\theta) p(x,\theta) d\mu(x) \\
&= \int \frac{\frac{\partial}{\partial \theta} p(x,\theta)}{p(x,\theta)} p(x,\theta) d\mu(x) \\
&= \frac{\partial}{\partial \theta} \int p(x,\theta) d\mu(x) \\
&= 0
\end{aligned}$$

as we wanted to show. ■

Remark The information $I(\theta)$ is therefore the variance of random score $s(X,\theta)$, which has expectation zero.

Proposition 8. *Suppose that Assumption (2) holds. Then*

$$I(\theta) = -H(\theta)$$

Proof Notice that

$$\frac{\partial^2}{\partial\theta\partial\theta'}\ell(x,\theta) = \frac{\frac{\partial^2}{\partial\theta\partial\theta'}p(x,\theta)}{p(x,\theta)} - \frac{\partial}{\partial\theta}\log p(x,\theta)\frac{\partial}{\partial\theta'}\log p(x,\theta)$$

we get

$$\begin{aligned} H(\theta) &= \int \left(\frac{\partial^2}{\partial\theta\partial\theta'}\ell(x,\theta)\right) p(x,\theta)d\mu(x) \\ &= \frac{\partial^2}{\partial\theta\partial\theta'}\int p(x,\theta)d\mu(x) - I(\theta) \\ &= -I(\theta) \end{aligned}$$

as was to be shown. ■

Remark Let X_i, $i = 1, \ldots, n$, be independent random variables with the information and expected Hessian denoted respectively by $\iota_i(\theta)$ and $h_i(\theta)$. Then the information $I(\theta)$ and expected Hessian $H(\theta)$ of $X = (X_1, \ldots, X_n)'$ are given by

$$I(\theta) = \sum_{i=1}^{n} \iota_i(\theta) \quad \text{and} \quad H(\theta) = \sum_{i=1}^{n} h_i(\theta)$$

If X_i's have identical distribution with $\iota(\theta)$ and $h(\theta)$, then $I(\theta) = n\,\iota(\theta)$ and $H(\theta) = n\,h(\theta)$.

Lemma 6.8. *Let $T = \tau(X)$ be an unbiased estimator for θ for which Assumption (3) holds. Then*

$$\mathbf{E}_\theta\Big(\tau(X)s(X,\theta)'\Big) = I \textit{ where } I \textit{ denotes } Identity \textit{ Matrix}$$

Proof Note that

$$\begin{aligned}
\mathbf{E}_\theta\Big(\tau(X)s(X,\theta)'\Big) &= \int \tau(x)\frac{\frac{\partial}{\partial\theta'}p(x,\theta)}{p(x,\theta)}p(x,\theta)d\mu(x) \\
&= \frac{\partial}{\partial\theta'}\int \tau(x)p(x,\theta)d\mu(x) \\
&= \frac{\partial}{\partial\theta'}\mathbf{E}_\theta(T) \\
&= I
\end{aligned}$$

from which the stated result is immediate. ■

Remark Since the expectation of random score is zero and $Cov(X, Y) = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y$, the lemma implies that covariance between an unbiased estimator and random score is identity for all θ.

Theorem 6.9. *Let $T = \tau(X)$ be an unbiased estimator of θ, and suppose that Assumption (3) holds. Then*

$$\text{var}_\theta T \geq I(\theta)^{-1}$$

Proof Note that

$$\text{var}_\theta \begin{pmatrix} \tau(X) \\ s(X, \theta) \end{pmatrix} = \begin{pmatrix} \text{var}_\theta \tau(X) & I \\ I & I(\theta) \end{pmatrix}$$

To get the stated result, observe that for any matrix

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \geq 0$$

we have $A_{11} \geq A_{12}A_{22}^{-1}A_{21}$. This is because we may write

$$A_{11} - A_{12}A_{22}^{-1}A_{21} = B'AB \geq 0$$

with $B' = \left(I,\ -A_{12}A_{22}^{-1}\right)$. ■

Examples

(a) Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ). Then

$$\begin{aligned} \ell(x_i, \lambda) &= -\lambda + x_i \log \lambda - \log x_i! \\ s(x_i, \lambda) &= -1 + \frac{x_i}{\lambda};\ h\left(x_i, \lambda\right) = -\frac{x_i}{\lambda^2} \\ \iota(\lambda) &= \frac{1}{\lambda^2}\mathbf{E}_\lambda\left(x_i\right) \ = \ \frac{1}{\lambda} \end{aligned}$$

and therefore,

$$I(\lambda) = n\,\iota(\lambda) = \frac{n}{\lambda}$$

The unbiased estimator $\tau(X) = \bar{X}$ achieves the Cramer-Rao bound, since

$$\text{var}_\lambda(\overline{X}) = \frac{\lambda}{n}$$

It is thus an UMVU estimator.

(b) Let $X_1, \ldots, X_n$ be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. Then

$$\ell(x_i, \mu, \sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$s(x_i, \mu, \sigma^2) = \begin{pmatrix} \frac{x_i-\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} + \frac{(x_i-\mu)^2}{2\sigma^4} \end{pmatrix}$$

$$H(x_i, \mu, \sigma^2) = \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{x_i-\mu}{\sigma^4} \\ -\frac{x_i-\mu}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(x_i-\mu)^2}{\sigma^6} \end{pmatrix}$$

We may easily deduce that

$$I(\mu, \sigma^2) = -H(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

It can now be easily seen that the estimator $\overline{X}$ for μ achieves the Cramer-Rao bound, and an UMVU estimator. However, the variance of the estimator S^2 for σ^2, which is $2\sigma^4/(n-1)$, is strictly greater than the Cramer-Rao bound. It can also be deduced that $\mathsf{E}\,(x_i-\mu)^4 = 3\sigma^4$. In that case, we are left with more questions. Is the Cramer-Rao bound with respect to σ^2 unattainable? and if so is there an unbiased estimator of σ^2 which is better than S^2?

(c) Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli (p). Then

$$\begin{aligned} \ell(x_i, p) &= x_i \log p + (1 - x_i) \log (1 - p) \\ s(x_i, p) &= \frac{x_i}{p} - \frac{1 - x_i}{1 - p} \\ H(x_i, p) &= -\frac{x_i}{p^2} - \frac{1 - x_i}{(1 - p)^2}. \end{aligned}$$

Thus,

$$I\left(p\right) = -H\left(p\right) = n\left(p^{-1} + (1 - p)^{-1}\right)$$

and the lower bound becomes

$$\frac{p\left(1 - p\right)}{n}.$$

6.6 Sufficiency and Unbiasedness

This section relates sufficient statistics to unbiased estimators. Indeed, the concept of sufficiency can be an immensely powerful tool for finding a UMVU estimator.

Theorem 6.10. *[Rao Blackwell]Suppose S is a sufficient statistic. Let $T = \tau(X)$ be an unbiased estimator for θ with finite mean and variance. Define a statistic $T_* = \mathbf{E}_\theta(T|S)$ and write $T_* = \tau_*(X)$. Then T_* is unbiased and $\mathbf{E}_\theta(T_* - \theta)^2 \leq \mathbf{E}_\theta(T - \theta)^2$ for all θ.*

Proof It follows from the Jensen's inequality that

$$\begin{aligned}(T_* - \theta)^2 &= [\mathbf{E}_\theta(T|S) - \theta]^2 \\ &\leq \mathbf{E}_\theta\left\{(T-\theta)^2|S\right\},\end{aligned}$$

where equality holds when T is a function of S. Taking expectations on both sides completes the proof. ■

Remarks

(a) If we write $f(S) = \mathbf{E}_\theta(T|S)$ and $S = \sigma(X)$, then $T_* = \tau_*(X)$ with $\tau_* = f \circ \sigma$. Note that $f\left(S\right)$, i.e. $\mathbf{E}_\theta(T|S)$, is a function of S only, and not of θ but only of a sample, since S is sufficient. Therefore, T_* is a statistic.

(b) A UMVUE for θ must be a function of every sufficient statistic, and hence of minimal sufficient statistic. Otherwise, we may always improve it by taking expectation conditional on a sufficient statistic. Hence, we need to consider only statistics that are functions of a sufficient statistic in our search for a UMVUE.

(c) $\mathbf{E}_{\theta_0}(\tau_* - \theta_0)^2 < \mathbf{E}_{\theta_0}(\tau - \theta_0)^2$ unless $\tau_* = \tau$ a.s. $\mathbf{P}_{\theta_0}$

Definition 6.11. A statistic T is called *complete* if $\mathbf{E}_\theta\Big(f(T)\Big) = 0$ for all $\theta \in \Theta$ implies $f = 0$ a.s. P_θ.

Remarks

(a) A statistic T is complete if and only if there exists a unique function of T that is unbiased. To see this, let $f_1(T)$ and $f_2(T)$ be unbiased, i.e., $\mathbf{E}_\theta f_1(T) = \mathbf{E}_\theta f_2(T) = \theta$. Then $\mathbf{E}_\theta f(T) = 0$ with $f = f_1 - f_2$, and completeness of T implies

that $f = 0$, that is $f_1 = f_2$ a.s. P_θ. For the other direction, suppose $f_1(T)$ is the unique unbiased estimator and $E_\theta f(T) = 0$ but $f \neq 0$. Then, $f_2 = f_1 + f$ yields another unbiased estimator which is not equal to f_1, leading to a contradiction.
(b) Notice that completeness is a property of a family of probability distributions, not of a particular distribution.

Theorem 6.12. *[Lehmann Scheffe]If S is complete and sufficient and $T_* = f(S)$ is unbiased, then T_* is the UMVU estimator for θ.*

Proof Obvious from Remark (b) of Rao-Blackwell(Theorem 6.10) and completeness(Definition 6.11). ■

Remark Given a complete and sufficient statistic S, it is now easy to obtain the UMVU estimator. We may indeed take any unbiased estimator U and let $T_* = \mathbf{E}_\theta(U|S)$. The resulting estimator T_* is the UMVU estimator, as one can easily see.

Examples:

(a) Let X_i, $i = 1, \ldots, n$, be i.i.d. $U(0, \theta)$. Recall that $T = \max_{1 \le i \le n} X_i$ is sufficient. The statistic T is indeed complete. To see this, note that for $t \le \theta$

$$\begin{aligned} \mathbf{P}_\theta\{T \le t\} &= \mathbf{P}_\theta\{X_1 \le t, \ldots, X_n \le t\} \\ &= \Big(\mathbf{P}_\theta\{X_i \le t\}\Big)^n \\ &= \left(\frac{t}{\theta}\right)^n \end{aligned}$$

and T has density

$$p_\theta(t) = \frac{nt^{n-1}}{\theta^n} I\{0 \le t \le \theta\}$$

Now, $\mathbf{E}_\theta f(T) = 0$ for all θ implies

$$\int_0^\theta t^{n-1} f(t)\, dt = 0$$

for all θ, from which we deduce that $f = 0$ a.s.

(b) Let X_i, $i = 1, \ldots, n$, and T be given as above. Define an unbiased estimator $U = 2X_1$ of θ. Suppose $T = t$. Then X_1 can take t with probability $1/n$, since every X_i, $i = 1, \ldots, n$, is equally likely to have the maximum value. Moreover, when $X_1 \neq t$ with probability $(n-1)/n$, X_1 is uniformly distributed on $(0, t)$. Therefore,

$$\begin{aligned}
\mathbf{E}_\theta(U|T=t) &= 2\mathbf{E}_\theta(X_1|T=t) \\
&= 2\mathbf{E}_\theta(X_1|T=t, X_1=t)P\left(X_1=t|T=t\right) \\
&+ 2\mathbf{E}_\theta\left(X_1|T=t, X_1<t\right)P\left(X_1<t|T=t\right) \\
&= 2\left(\frac{1}{n}t + \frac{n-1}{n}\frac{t}{2}\right) \\
&= \frac{n+1}{n}t
\end{aligned}$$

Thus

$$\tau_*(x) = \frac{n+1}{n}\max_{1\leq i\leq n} x_i$$

is the UMVU estimator of θ.

In general, it is rather complicated analysis problem to show that a statistic is complete. However, there is a powerful result for the exponential family.

Theorem 6.13. *Let $\{X_i\}_{i=1}^n$ be a random sample from an exponential family with pdf or pmf of the form*

$$f(x|\theta) = h(x)\, c(\theta) \exp\left(\sum_{j=1}^{k} w_j(\theta)\, t_j(x)\right),$$

where $\theta = (\theta_1, ..., \theta_k)$. Then, the statistic

$$T = \left(\sum_{i=1}^{n} t_j(X_i), \quad j = 1, ..., k\right)$$

is complete if $\{(w_1(\theta), ..., w_k(\theta)) : \theta \in \Theta\}$ contains an open set in $\mathbf{R}^k$.

For example, it follows from this theorem that $T = \left(\bar{X}_n, S_n^2\right)$ is a transformation of a complete and sufficient statisic for a random sample $\{X_i\}_{i=1}^n$ from $\mathbf{N}\left(\mu, \sigma^2\right)$. And it is unbiased and thus a UMVUE.

7 Hypothesis Testing

7.1 Introduction

Let a partition of Θ be made and given by

$$\Theta = \Theta_0 \cup \Theta_1$$

The *null hypothesis* is given by

$$H_0 : \ \theta \in \Theta_0$$

and is tested against the *alternative hypothesis*

$$H_1 : \ \theta \in \Theta_1$$

The null hypothesis H_0 is maintained unless it is rejected in favor of the alternative hypothesis H_1. When Θ_0 and Θ_1 are singleton sets, we say that the hypothesis is *simple*. Otherwise, they are *composite*.

The statistical hypothesis testing is usually based on a *test statistic* τ. According to the value of τ, the state space $\mathcal{X}$ is partitioned as the disjoint union of the *critical region* C and *acceptance region* A, i.e.,

$$\mathcal{X} = C \cup A$$

If $x \in C$, then H_0 is rejected in favor of H_1. If, on the other hand, $x \in A$, then H_0 is continued to be maintainted. A 'test' is thus completely synonymous to a 'critical region'. We will therefore refer to a test with its critical region.

Let $X = (X_1, \ldots, X_n)'$ be a random sample, and suppose the distribution of X is given by a parametric family $\mathcal{P} = \{\mathsf{P}_\theta | \, \theta \in \Theta\}$. The *power function* $\pi(\theta)$ of the test C is defined by

$$\pi(\theta) = \mathsf{P}_\theta(C)$$

Moreover,

$$\max_{\theta \in \Theta_0} \pi(\theta)$$

is called the *size* of the test, while the values of π at $\theta \in \Theta_1$ are called the *power* of the test. We define

Definition 7.1. The test C^* is Uniformly Most Powerful (UMP) if $\forall \theta \in \Theta_1$,

$$\mathsf{P}_\theta(C^*) \geq \mathsf{P}_\theta(C)$$

for any test represented by C of the same size.

When both the null and alternative hypotheses are simple, we may write $H_0 = \theta_0$ and $H_1 = \theta_1$. Since $\Theta = \{\theta_0, \theta_1\}$ in this case, $\mathcal{P}$ consists of two distributions, which we write as

$$\mathsf{P}_{\theta_0} = \mathsf{P}_0 \quad \text{and} \quad \mathsf{P}_{\theta_1} = \mathsf{P}_1$$

for the null and alternative distributions, respectively. Clearly, $\mathsf{P}_0(C)$ and $\mathsf{P}_1(C)$ are the size and power of the test. Notice that $\mathsf{P}_0(C)$ is the probability of rejecting H_0 when it is true. On the other hand, $\mathsf{P}_1(A)$ is the probability of accepting H_0 when H_0 is false (and H_1 is true). Both of $\mathsf{P}_0(C)$ and $\mathsf{P}_1(A)$ are the probabilities of making errors, which we refer to as the *type* I and *type* II errors, respectively.

7.2 The Neyman-Pearson Lemma

Assume that both the null and alternative hypotheses are simple, and the distributions P_0 and P_1 are given by the likelihood functions $p(x, \theta_0)$ and $p(x, \theta_1)$.

Lemma 7.2. *[Neyman-Pearson]The test which rejects $\mathcal{H}_0$ when*

$$\lambda(x) = \frac{p(x,\theta_1)}{p(x,\theta_0)} \geq c$$

for a constant c is most powerful.

Proof Let $C^* = \{x |\, \lambda(x) \geq c\}$, and suppose C is any test other than C^* with the same size, i.e. $\mathsf{P}_0(C) = \mathsf{P}_0(C^*)$. We need to show $\mathsf{P}_1(C) \leq \mathsf{P}_1(C^*)$. Assume without loss of generality that C and C^* are disjoint. Then

$$p(x,\theta_1) \geq c\, p(x,\theta_0) \quad \text{over} \quad C^*$$

and

$$p(x,\theta_1) < c\, p(x,\theta_0) \quad \text{over} \quad C$$

We thus have

$$\mathsf{P}_1(C^*) = \int_{C^*} p(x,\theta_1)dx \geq \int_{C^*} cp(x,\theta_0)dx = c\mathsf{P}_0(C^*)$$

and

$$\mathsf{P}_1(C) = \int_{C} p(x,\theta_1)dx \leq \int_{C} cp(x,\theta_0)dx = c\mathsf{P}_0(C)$$

The stated result then follows immediately from

$$\mathsf{P}_1(C^*) \geq c\mathsf{P}_0(C^*) = c\mathsf{P}_0(C) \geq \mathsf{P}_1(C)$$

■

Remark The above lemma guarantes the existence of a best test under simple null and alternative hypotheses. The criterion used in the construction of the likelihood ratio (LR) test $\lambda(x) \geq c$ indeed tells us about the form of the partition of $\mathcal{X}$. We can view $\lambda(x)$ as a ratio of marginal power to marginal size. Then the LR test includes only those points in $\mathcal{X}$ that have significant enough power increase per unit of size increase.

The LR test is generalized as

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_1} p(x, \theta)}{\sup_{\theta \in \Theta_0} p(x, \theta)}$$

for composit hypotheses. The generalized LR test rejects $\mathcal{H}_0$ when $\tau(x) \geq c$, where $\tau(x)$ is any monotone increasing function of $\lambda(x)$ and c is given for a prescribed size. Note that the Neyman-Pearson lemma does not apply to the generalized LR test. Optimality properties of the generalized LR tests are much harder to show.

Examples

(a) Let $X_1, \dots, X_n$ be i.i.d. $\mathbf{N}(\mu, 1)$, and consider

$$\mathcal{H}_0 : \ \mu = 0 \quad \text{against} \quad \mathcal{H}_1 : \ \mu = 1$$

Since both null and alternative hypotheses are simple, we can apply the Neyman-

Pearson lemma to obtain a best test. The likelihood ratio is

$$\begin{aligned}
\lambda(x) &= \frac{p(x,1)}{p(x,0)} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^{n} \exp\left(-\frac{1}{2}\sum(x_i-1)^2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^{n} \exp\left(-\frac{1}{2}\sum x_i^2\right)} \\
&= \exp\left(-\frac{1}{2}\sum(x_i-1)^2+\frac{1}{2}\sum x_i^2\right)
\end{aligned}$$

giving

$$\ln\lambda(x) = \sum x_i - \frac{n}{2}$$

Since

$$\tau(x) = \ln\lambda(x) + \frac{n}{2} = \sum x_i$$

is an increasing function of $\lambda(x)$, we can use $\tau(x)$ to construct the LR test as

$$C = \{x|\tau(x) \geq c\}$$

We now determine the value of the constant c. Given a prescribed size, say 5%, we can compute c using the null distribution to ensure

$$\mathsf{P}_0(C) = 0.05.$$

Note that $\tau(X) \sim \mathbf{N}(0, n)$ under $\mathcal{H}_0$, which implies

$$\frac{\tau(X)}{\sqrt{n}} = \frac{\sum X_i}{\sqrt{n}} \sim \mathbf{N}(0, 1).$$

Then we know from the $\mathbf{N}(0, 1)$ table that

$$\mathbf{P}_0 \left\{ \frac{\tau(X)}{\sqrt{n}} \geq 1.645 \right\} \sim 0.05$$

or

$$\mathsf{P}_0 \left\{ x \,\middle|\, \tau(x) \geq 1.645\sqrt{n} \right\} \sim 0.05$$

and this gives the value of $c = 1.645\sqrt{n}$.

Next we consider

$$\mathcal{H}_0 : \ \mu = 0 \quad \text{against} \quad \mathcal{H}_1 : \ \mu > 0$$

with the composite alternative hypothesis. Note, however, that for any value of $\mu_1 > 0$, the most powerful test for

$$\mathcal{H}_0 : \mu = 0 \quad \text{against} \quad \mathcal{H}_1 : \ \mu = \mu_1$$

is given by $C = \{x |\, \tau(x) = \sum x_i \geq c\}$. Therefore, C is the uniformly most powerful test.

This is also a *one-sided hypothesis testing*, while the alternative that negates the null hypothesis, that is, $\mathcal{H}_1 : \mu \neq 0$, makes it a *two-sided hypothesis testing*. Under the two-sided testing, one may employ the generalized LR test statistic.

(b) Let $X_1, \ldots, X_n$ be i.i.d. **N**$(0, \sigma^2)$, and consider

$$\mathcal{H}_0 : \sigma^2 = \sigma_0^2 \quad \text{against} \quad \mathcal{H}_1 : \ \sigma^2 = \sigma_1^2 > \sigma_0^2$$

The likelihood ratio is

$$\lambda(x) = \frac{\left(\frac{1}{2\pi\sigma_1^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_1^2}\sum x_i^2\right)}{\left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_0^2}\sum x_i^2\right)}$$

$$= \left(\frac{\sigma_0^2}{\sigma_1^2}\right)^{\frac{n}{2}} \exp\left(\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\sum x_i^2\right)$$

Note that $\tau(x) = \sum x_i^2$ is an increasing function of $\lambda(x)$. Then the LR test is given by $C = \{x|\,\tau(x) \geq c\}$.

Now we find the constant c for a 5% test. Under $\mathcal{H}_0$,

$$\frac{1}{\sigma_0^2}\sum X_i^2 = \frac{1}{\sigma_0^2}\tau(X) \sim \chi_n^2$$

i.e., $\tau(X) \sim \sigma_0^2\chi_n^2$. Then it follows directly that

$$\mathsf{P}_0(C) = \mathsf{P}_0\{x|\,\tau(x) \geq c\} = 0.05$$

giving $c = \sigma_0^2 \chi^2_{n(0.05)}$, where $\chi^2_{n(0.05)}$ can be found from the χ^2 table.

We note that for all $\mathcal{H}_1 : \ \sigma_1^2$ such that $\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) > 0$, the above LR test C continues to be the most powerful test by the Neyman-Pearson lemma. Hence, C is UMP test for the composite alternative hypothesis

$$\mathcal{H}_0 : \ \sigma^2 = \sigma_0^2 \quad \text{against} \quad \mathcal{H}_1 : \ \sigma^2 > \sigma_0^2$$

Similarly, you can show that for the composite hypothesis

$$\mathcal{H}_0 : \ \sigma^2 \leq \sigma_0^2 \quad \text{against} \quad \mathcal{H}_1 : \ \sigma^2 > \sigma_0^2$$

the above test C is also UMP test.

(c) Let $X_1, \ldots, X_n$ be i.i.d. $\mathbf{N}(\mu, \sigma^2)$. We first consider

$$\mathcal{H}_0 : \ \mu = \mu_0 \quad \text{against} \quad \mathcal{H}_1 : \ \mu \neq \mu_0$$

where the null and alternative hypotheses are, respectively, one-way and two-way composite. Thus, we must consider generalized LR test with the following general likelihood ratio

$$\lambda(x) = \frac{\sup\limits_{\mu,\sigma^2} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)}{\sup\limits_{\sigma^2} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2\right)}$$

We showed earlier that

$$\hat{\mu}_{\mathsf{ML}} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2_{\mathsf{ML}} = \frac{1}{n}\sum(x_i-\bar{x})^2$$

maximize the likelihood function. Thus we can compute the generalize LR test using the above estimates as

$$\begin{aligned}\lambda(x) &= \frac{\left(\frac{n}{\sum(x_i-\bar{x})^2}\right)^{\frac{n}{2}} \mathsf{exp}\left(-\frac{n}{2\sum(x_i-\bar{x})^2}\sum(x_i-\bar{x})^2\right)}{\left(\frac{n}{\sum(x_i-\mu_0)^2}\right)^{\frac{n}{2}} \mathsf{exp}\left(-\frac{n}{2\sum(x_i-\mu_0)^2}\sum(x_i-\mu_0)^2\right)} \\ &= \left(\frac{\sum(x_i-\mu_0)^2}{\sum(x_i-\bar{x})^2}\right)^{\frac{n}{2}} \\ &= \left(1+\frac{n(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2}\right)^{\frac{n}{2}}\end{aligned}$$

Then the generalized LR may be defined as

$$C = \left\{ x \,\middle|\, \tau(x) = \frac{n(\bar{x}-\mu_0)^2}{\sum(x_i-\bar{x})^2} \geq c \right\}$$

Now, we have under $\mathcal{H}_0$ that

$$\frac{n(\bar{X}-\mu_0)^2}{\sigma^2} \sim Z^2 \quad \text{and} \quad \frac{\sum(X_i-\bar{X})^2}{\sigma^2} \sim W$$

where $Z \equiv \mathbf{N}(0,1)$ and $W \equiv \chi^2_{n-1}$. Then it follows that

$$(n-1)\tau(X) = \frac{\dfrac{n(\bar{X}-\mu_0)^2}{\sigma^2}}{\dfrac{\dfrac{\sum(X_i-\bar{X})^2}{\sigma^2}}{(n-1)}} \sim \left(\frac{Z}{\sqrt{W}}\right)^2 \equiv (t_{n-1})^2 \equiv F(1, n-1)$$

In order to find the critical values, we look at

$$\mathsf{P}_0 \left\{ x \,\middle|\, \tau(x) \geq \frac{t^2_{n-1}(\alpha)}{n-1} \right\} = \alpha$$

for a size α% test.

Next, we consider

$$\mathcal{H}_0 : \ \sigma^2 = \sigma_0^2 \quad \text{against} \quad \mathcal{H}_1 : \ \sigma^2 \neq \sigma_0^2$$

where both hypothesis are again composite. The likelihood ratio is written as

$$\lambda(x) = \frac{\sup\limits_{\mu,\sigma^2} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \text{exp}\left(-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2\right)}{\sup\limits_{\mu} \left(\frac{1}{2\pi\sigma_0^2}\right)^{\frac{n}{2}} \text{exp}\left(-\frac{1}{2\sigma_0^2}\sum(x_i-\mu)^2\right)}$$

$$= \frac{\left(\frac{n}{\sum(x_i-\bar{x})^2}\right)^{\frac{n}{2}} \text{exp}\left(-\frac{n}{2\sum(x_i-\bar{x})^2}\sum(x_i-\bar{x})^2\right)}{\left(\frac{1}{\sigma_0^2}\right)^{\frac{n}{2}} \text{exp}\left(-\frac{1}{2\sigma_0^2}\sum(x_i-\bar{x})^2\right)}$$

$$= \text{ constant} \left(\frac{1}{2\sigma_0^2}\sum(x_i-\bar{x})^2\right)^{-\frac{n}{2}} \text{exp}\left(\frac{1}{2\sigma_0^2}\sum(x_i-\bar{x})^2\right)$$

Thus, we can express $\lambda(x)$ as $f(z) = z^{-n/2}e^z$, with $z = \sum(x_i - \bar{x})^2$. A generalized LR test may then be defined as $C = \{z|\, f(z) \geq c\}$ or $C = \{z \,|\, z \leq c_1 \text{ or } z \geq c_2\}$, where c_1 and c_2 are such that $f(c_1) = f(c_2)$. In practice, we usually use the following result for $\tau(X) = \sum(X_i - \bar{X})^2$:

$$\frac{\tau(X)}{\sigma_0^2} \sim \chi^2_{n-1}$$

under $\mathcal{H}_0$. The generalized LR test is then defined as

$$C = \left\{ x \;\middle|\; \tau(x) \leq \chi^2_{(n-1)\left(1-\frac{\alpha}{2}\right)} \text{ or } \tau(x) \geq \chi^2_{(n-1)\frac{\alpha}{2}} \right\}$$

for a size α% test.

7.3 p-value

Another way of reporting the result of a hypothesis testing is to report the p-value

$$\sup_{\Theta_0} P_\theta \left\{ \tau\left(X\right) > \tau\left(x\right) \right\}$$

where $\tau\left(X\right)$ is the test statistic such that the critical region of the test is specified by $\{\tau\left(x\right) > c\}$. Certainly the smaller the p-value the stronger the evidence against the null hypothesis. The p-value is more informative than the dichotonous decision "Accept $\mathcal{H}_0$" or "Reject $\mathcal{H}_0$".

8 Asymptotic Theory

8.1 Limit Concepts in Probability

Let $\{E_n\}$ be a sequence of events. We say that $\{E_n\}$ is *monotone* when

$$E_1 \subset E_2 \subset \cdots \quad \text{or} \quad E_1 \supset E_2 \supset \cdots$$

For a monotone sequence $\{E_n\}$ of events, we define

$$\lim_{n\to\infty} E_n = \bigcup_{n=1}^{\infty} E_n \quad \text{or} \quad \lim_{n\to\infty} E_n = \bigcap_{n=1}^{\infty} E_n$$

depending upon whether the sequence is increasing or decreasing.

Theorem 8.1. *Let E_n be a monotone sequence of events. Then*

$$\mathbf{P}\left(\lim_{n\to\infty} E_n\right) = \lim_{n\to\infty} \mathbf{P}(E_n)$$

Proof Assume $\{E_n\}$ is monotonically increasing. Define

$$F_n = E_n - E_{n-1}$$

for $n = 1, 2, \ldots$, with the convention that $E_0 = \emptyset$. It follows that

$$\begin{array}{ccc} \mathbf{P}\left(\bigcup_{n=1}^{\infty} F_n\right) & = & \sum_{n=1}^{\infty} \mathbf{P}(F_n) \\ \| & & \| \\ \mathbf{P}\left(\lim_{n\to\infty} E_n\right) & & \lim_{n\to\infty} \mathbf{P}(E_n) \end{array}$$

as was to be shown. For a decreasing sequence E_n of events, let $F_n = E_1 - E_n$ and apply the above result.(Homework) ■

For a sequence $\{E_n\}$ of events, we generally define

$$\limsup_{n\to\infty} E_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_k$$

$$\liminf_{n\to\infty} E_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} E_k$$

Note that

$\omega \in \limsup E_n$ iff for any n, there exists $k \geq n$ such that $\omega \in E_k$

$\omega \in \liminf E_n$ iff there exists n such that $\omega \in E_k$ for all $k \geq n$

For obvious reasons, we often write

$$\begin{aligned}\limsup E_n &= \{E_n \quad \text{i.o.}\}\\ \liminf E_n &= \{E_n \quad \text{ev}\}\end{aligned}$$

where "i.o." stands for "infinitely often" and "ev" is the abbreviation for "eventually". When $\limsup E_n = \liminf E_n$, we say that E_n has limit $E = \limsup E_n = \liminf E_n$.

Corollary 9. *Let $\{E_n\}$ be a sequence of events. We have*

$$\begin{aligned}\mathbf{P}(\liminf E_n) &\le \liminf \mathbf{P}(E_n)\\ &\le \limsup \mathbf{P}(E_n) \le \mathbf{P}(\limsup E_n)\end{aligned}$$

Proof Since

$$\begin{array}{ccc}\bigcap_{k=n}^{\infty} E_k & \uparrow & \bigcup_n \bigcap_{k=n}^{\infty} E_k\\ \bigcup_{k=n}^{\infty} E_k & \downarrow & \bigcap_n \bigcup_{k=n}^{\infty} E_k,\end{array}$$

it follows that

$$\mathbf{P}(E_n) \geq \mathbf{P}\left(\bigcap_{k=n}^{\infty} E_k\right) \to \mathbf{P}(\liminf E_n)$$
$$\mathbf{P}(E_n) \leq \mathbf{P}\left(\bigcup_{k=n}^{\infty} E_k\right) \to \mathbf{P}(\limsup E_n).$$

Then, take liminf and limsup on both sides, respectively, to each equation to complete the proof. ■

Theorem 8.2. *[Borel-Cantelli]Let $\{E_n\}$ be a sequence of events. We have*

$$\sum_{n=1}^{\infty} \mathbf{P}(E_n) < \infty \quad \textit{implies} \quad \mathbf{P}\left(\limsup_{n\to\infty} E_n\right) = 0$$

Proof Notice that

$$\begin{aligned}
\mathbf{P}(\limsup E_n) &= \mathbf{P}\left(\bigcap_n \bigcup_{k\geq n} E_k\right) \\
&\leq \mathbf{P}\left(\bigcup_{k\geq n} E_k\right) \\
&\leq \sum_{k\geq n} \mathbf{P}(E_k) \to 0
\end{aligned}$$

as was to be shown. ∎

8.2 Modes of Convergence

In this section, we will study various modes of convergence for a sequence $\{X_n\}$ of random variables.

Definition 8.3. Let X_n be defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say that $\{X_n\}$ converges in probability to X if for any $\varepsilon > 0$

$$\mathbf{P}\left\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\right\} \ \to \ 0$$

and denote by $X_n \xrightarrow{p} X$.

Definition 8.4. Let X_n be defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say that $\{X_n\}$ converges almost surely to X if

$$\mathbf{P}\left\{\omega |\ X_n(\omega) \to X(\omega)\right\} \ = \ 1$$

and write $X_n \xrightarrow{a.s.} X$.

Remark We may equivalently formulate the a.s. convergence as

$$\begin{aligned}\mathbf{P}\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon \ \ i.o.\} &= \mathbf{P} \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} \{\omega : |X_k(\omega) - X(\omega)| > \varepsilon\} \\ &= \mathbf{P} \limsup_{n \to \infty} \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\} \\ &= 0\end{aligned}$$

or similarly as

$$\begin{aligned}\mathbf{P}\{\omega : |X_n(\omega) - X(\omega)| < \varepsilon \ \ ev.\} &= \mathbf{P} \bigcup_{n=1}^{\infty} \bigcap_{k \geq n} \{\omega : |X_k(\omega) - X(\omega)| < \varepsilon\} \\ &= \mathbf{P} \liminf_{n\to\infty} \{\omega : |X_n(\omega) - X(\omega)| < \varepsilon\} \\ &= 1\end{aligned}$$

for any $\varepsilon > 0$.

Examples:

(a) Define a sequence of random variables $\{X_n\}$ whose distribution is givenby

$$X_n = \langle \begin{array}{l} 1 \text{ with probability } n^{-1} \\ 0 \text{ with probability } 1 - n^{-1}. \end{array}$$

For any $\varepsilon > 0$,

$$\mathbf{P}(|X_n - 0| \geq \varepsilon) \leq \mathbf{P}\left(X_n = 1\right) = n^{-1} \to 0,$$

as $n \to \infty$. Therefore, $X_n \xrightarrow{p} 0$.

(b) Let $\Omega = [0,1]$ with the Lebesque measure. Define random variables $X_n(x) = x + x^n$ and $X(x) = x$. For every $x \in [0,1)$, $x^n \to 0$ as $n \to \infty$ and hence $X_n(x) \to x = X(x)$. As $P([0,1)) = 1$, $X_n \xrightarrow{a.s.} X$.

(c) Define a sequence of random variables $\{X_n\}$ whose distribution is givenby

$$X_n = \langle \begin{array}{l} n \text{ with probability } n^{-2} \\ 0 \text{ with probability } 1 - n^{-2}. \end{array}$$

Then, we can show that $X_n \xrightarrow{a.s.} X$.

Definition 8.5. Let X_n be defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We say that $\{X_n\}$ converges in L^p to X if

$$\mathbf{E}\,|X_n - X|^p \;\to\; 0$$

and write $X_n \xrightarrow{\mathcal{L}^p} X$.

Remark Most commonly considered is the case $p = 2$, i.e. L^2-convergence, which we often refer to as the *mean squared error* convergence.

For a random variable X defined on $(\Omega, \mathcal{F}, \mathbf{P})$, we define the *distribution* of X to be the probability measure given by

$$P_X(B) = \mathbf{P} \circ X^{-1}(B)$$

for any Borel set $B \in \mathcal{B}(\mathcal{R})$, the Borel σ-field on $\mathcal{R}$.

Definition 8.6. We say that $\{X_n\}$ converges in distribution or in law to X if

$$\mathbf{E}f(X_n) \ \to \ \mathbf{E}f(X)$$

for every function f that is bounded and continuous a.s. in P_X, in symbols $X_n \xrightarrow{d} X$.

Remarks

(a) In the above definition, the function f need not be continuous at every point. It suffices that f is continuous with P_X probability 1, i.e., f may be discontinuous on a set N such that $P_X(N) = 0$.

(b) For the convergence in distribution, $\{X_n\}$ need not be defined on a common probability space. It is not a convergence of $\{X_n\}$, but that of probability measures $\{P_n\}$ induced by $\{X_n\}$. We may regard it as a convergence in the

set of probability measures with some weak topology. For this reason, it is often referred to as the *weak convergence*.
(c) Let P and Q be two probability measures on $(\mathcal{R}, \mathcal{B}(\mathcal{R}))$. Then $P = Q$ if and only if $P(B) = Q(B)$ for all $B \in \mathcal{B}(\mathcal{R})$. We may show that the condition holds if and only if $\int f\,dP = \int f\,dQ$ for every bounded and continuous f. Our definition of convergence in distribution may similarly be motivated.

In what follows, we let F_n (and F) and φ_n (and φ), respectively, be the distribution and characteristic functions of X_n (and X).

Lemma 8.7. *The following are equivalent:*
(a) $X_n \xrightarrow{d} X$.
(b) $\mathbf{E}f(X_n) \to \mathbf{E}f(X)$ *for every bounded and uniformly continuous* f.
(c) $F_n(t) \to F(t)$ *for every continuity point* t *of* F.
(d) $\varphi_n(t) \to \varphi(t)$ *for all* t.

Proof It is trivial that (a) ⇒ (b). To see that (a) ⇒ (c), consider

$$f_t(x) = 1\{x \leq t\}$$

Clearly, f_t is bounded, and continuous a.s. in P_X whenever t is a continuity point of F. It thus follows that $F_n(t) = \mathbf{E} f_t(X_n)$ converges to $F(t) = \mathbf{E} f_t(X)$ for such point t. For the implication (a) ⇒ (d), look at the class of functions

$$f_t(x) = e^{itx}$$

For every t, the function f_t is bounded and continuous. Therefore, $\varphi_n(t) = \mathbf{E} f_t(X_n)$ converges to $\varphi(t) = \mathbf{E} f_t(X)$. The proofs for other implications are more involved and omitted. ■

Definition 8.8. We say that $\{X_n\}$ converges to X in the p-th moment if

$$\mathbf{E}\, X_n^p \;\to\; \mathbf{E}\, X^p.$$

Remark Let $\hat{\theta}_n$ be an estimator of the parameter θ. For $\hat{\theta}_n$ to be a good estimator, it must be *consistent* and *asymptotically normal*, if properly standardized. The estimator $\hat{\theta}_n$ is said to be consistent if it converges to θ. It is called *strongly* or *weakly* consistent, depending upon whether the mode of convergence is almost sure or in probability. The asymptotic normality implies

$$\sqrt{n}(\hat{\theta}_n - \theta) \;\xrightarrow{d}\; \mathsf{N}(0, \Sigma)$$

for some covariance matrix Σ.

8.3 Relationships among Modes of Convergence

In below we summarize the relationships existing among various modes of convergence introduced in the previous section.

Theorem 8.9.

$$\begin{array}{cccccc}
\xrightarrow{a.s.} & \text{(a)} & & \text{(c)} & \\
 & \Longrightarrow & \xrightarrow{p} & \Longrightarrow & \xrightarrow{d} \\
\xrightarrow{\mathcal{L}^p} & \text{(b)} & & &
\end{array}$$

Proof To show (a), notice that

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{P}\left\{|X_n - X| > \varepsilon\right\} &\leq \lim_{n\to\infty} \mathbf{P} \bigcup_{k\geq n} \left\{|X_k - X| > \varepsilon\right\} \\
&= \mathbf{P} \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} \left\{|X_k - X| > \varepsilon\right\}.
\end{aligned}$$

Part (b) follows directly from the Chebyshev inequality

$$\mathbf{P}\{|X_n - X| > \varepsilon\} \leq \frac{\mathbf{E}|X_n - X|^p}{\varepsilon^p}.$$

For (c), pick an arbitrary f that is bounded and uniformly continuous. Let $M = \sup|f(x)|$, and for any $\varepsilon > 0$ choose δ such that

$$|X_n - X| \leq \delta \quad \text{implies} \quad |f(X_n) - f(X)| \leq \varepsilon$$

We have

$$|f(X_n) - f(X)| \leq \varepsilon + 2M\, \mathsf{I}\{|X_n - X| > \delta\}$$

Then it follows that

$$\begin{aligned} |\mathbf{E}f(X_n) - \mathbf{E}f(X)| &\leq \mathbf{E}|f(X_n) - f(X)| \\ &\leq \varepsilon + 2M\, \mathbf{P}\{|X_n - X| > \delta\} \end{aligned}$$

from which the stated implication is immediate. ■

Other implications do not hold, as we will see in the following counterexamples.

Counterexamples

Consider a probability space $([0,1], \mathcal{B}[0,1], \lambda)$, where λ is the Lebesgue measure and $\mathcal{B}[0,1]$ is the Borel σ-field on $[0,1]$. Define a sequence $\{X_n\}$ of random variables by

$$X_n(\omega) = n^{\frac{1}{p}} \mathsf{I}\left\{0 \le \omega \le \frac{1}{n}\right\}$$

for $p > 0$, and a sequence $\{Y_n\}$ by

$$Y_n = \mathsf{I}\left\{\frac{b-1}{a} \le \omega \le \frac{b}{a}\right\}$$

for $n = \frac{a(a-1)}{2} + b$ with $a = 1, 2, \ldots$ and $1 \le b \le a$.

It is not difficult to see that

$$X_n \xrightarrow{a.s.} 0 \quad \text{but not} \quad X_n \xrightarrow{\mathcal{L}^p} 0$$

We indeed have $X_n(\omega) \to\ 0$ for all $\omega \in (0,1]$, but $\mathbf{E}X_n^p = 1$ for all n. On the contrary,

$$Y_n \xrightarrow{\mathcal{L}^p} 0 \quad \text{but not} \quad Y_n \xrightarrow{a.s.} 0$$

Clearly, $\mathbf{E}Y_n^p = 1/a \to 0$, but $Y_n(\omega)$ does not converge for any $\omega \in [0,1]$.

Remarks

(a) It is also obvious from the above counter examples that $\xrightarrow{p}$ does not imply $\xrightarrow{a.s.}$ nor $\xrightarrow{\mathcal{L}^p}$. If $\xrightarrow{p}\Rightarrow \xrightarrow{a.s.}$, for instance, it falsely follows that $\xrightarrow{\mathcal{L}^p} \Rightarrow\xrightarrow{p}\Rightarrow \xrightarrow{a.s.}$, i.e., $\xrightarrow{\mathcal{L}^p} \Rightarrow \xrightarrow{a.s.}$. (b) It is clearly untrue that $\xrightarrow{d}\Rightarrow\xrightarrow{p}$ because the former does not even require that $\{X_n\}$ be defined on a common probability space. It can be also obvious if we notice that convergence in distribution is a condition only on the distribution functions of the X_n and therefore, it contains no reference to Ω and no information about X_n. If, however, they are defined on a common probability space, then the implication does holds, as laid out in Theorem 8.10 (b) below.

Theorem 8.10. *We have*
(a) If $X_n\xrightarrow{\mathcal{P}}X$, then there exists a subsequence $\{X_{n_k}\}$ such that $X_{n_k}\xrightarrow{a.s.}X$.
(b) Let $\{X_n\}$ be defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and c be a constant. If $X_n \xrightarrow{d} c$, then $X_n \xrightarrow{p} c$.

Proof For (a), let $X_n \xrightarrow{\mathcal{P}} X$. Since

$$\mathbf{P}\left\{|X_n - X| > \frac{1}{2^k}\right\} \;\to\; 0$$

for any k, we may choose n_k for each k such that

$$\mathbf{P}\left\{|X_{n_k} - X| > \frac{1}{2^k}\right\} \;\leq\; \frac{1}{2^k}$$

Now notice that

$$\sum_{k=1}^{\infty} \mathbf{P}\left\{|X_{n_k} - X| > \frac{1}{2^k}\right\} \;\leq\; \sum_{k=1}^{\infty} \frac{1}{2^k} < \infty$$

Then it follows from Borel-Cantelli(Theorem 8.2) that

$$\mathbf{P} \limsup_{k\to\infty} \left\{|X_{n_k} - X| > \frac{1}{2^k}\right\} = 0$$

i.e.

$$\mathbf{P}\left\{|X_{n_k} - X| > \frac{1}{2^k}\ i.o.\right\} = 0$$

as was to be shown.

To show part (b), let

$$f_c(x) = \mathsf{I}\{|x-c| > \varepsilon\}$$

for a given $\varepsilon > 0$. If $X_n \xrightarrow{d} c$, then $\mathbf{E}f_c(X_n) = \mathbf{P}\{|X_n - c| > \varepsilon\}$ converges to $\mathbf{E}f_c(c) = 0$, since f_c is continuous a.s., that is, $\mathbf{P}\{|X-c| = \varepsilon\} = 0$. The stated result thus follows easily. ■

8.4 Basics of Asymptotic Analysis

8.4.1 Stochastic Order

We now introduce set of notations, which is known as Mann and Wald's $O_p(1)$ and $o_p(1)$. In what follows, let $\{a_n\}$ and $\{b_n\}$ be sequences of real numbers. We define

Definition 8.11. We write $x_n = o(a_n)$ and $y_n = O(b_n)$, respectively, when

$$\frac{x_n}{a_n} \to 0 \quad \text{and} \quad \left|\frac{y_n}{b_n}\right| \leq M$$

for $M > 0$.

Remarks

(a) In particular, $x_n = o(1)$ if the sequence $\{x_n\}$ converges to zero, and $y_n = O(1)$ if the sequence $\{y_n\}$ is bounded.

(b) We may write

$$o(a_n) = a_n o(1) \quad \text{and} \quad O(b_n) = b_n O(1)$$

in general.

(c) The equality containing o and O is not really an equality. For instance, $o(1) = O(1)$, but $O(1) \neq o(1)$.

(d) For $y_n = O(1)$, it suffices to have $|y_n| \leq M$ for large n. If $|y_n| \leq M$ for all $n > N$, say, then it follows that $|y_n| \leq M_*$ for all n with $M_* = \max\{y_1, \ldots, y_N, M\}$.

Lemma 8.12. *We have*
(a) $O(o(1)) = o(1)$
(b) $o(O(1)) = o(1)$
(c) $o(1)O(1) = o(1)$

Proof For part (a), let $x_n = o(1)$ and $y_n = O(x_n)$. With M such that $|y_n/x_n| < M$, we have

$$|y_n| < M|x_n| \to 0$$

which shows that $y_n \to 0$, as required.

For (b), assume $x_n = O(1)$ and $y_n = o(x_n)$. Choose M such that $|x_n| < M$. Then it follows that

$$\frac{|y_n|}{M} < \left|\frac{y_n}{x_n}\right| \to 0$$

as was to be shown.

To prove (c), let $x_n = o(1)$ and $y_n = O(1)$. Then, for M such that $|y_n| < M$,

$$|x_n y_n| < |x_n|M \to 0$$

which yields the stated result. ■

Remark In general, we have

$$O(o(a_n)) = O(a_n o(1)) = a_n O(o(1)) = a_n o(1) = o(a_n)$$

for the result comparable to part (a) of Lemma 8.12.

Definition 8.13. We write $X_n = o_p(a_n)$ if X_n/a_n converges in probability to zero. Moreover, $Y_n = O_p(b_n)$ if for any $\varepsilon > 0$, there exists $M > 0$ such that

$$\mathbf{P}\left\{\left|\frac{Y_n}{b_n}\right| > M\right\} < \varepsilon.$$

Remarks

(a) In particular, $X_n = o_p(1)$ if $X_n \xrightarrow{p} 0$. When $Y_n = O_p(1)$, there exists M such that $\mathbf{P}\{|Y_n| > M\} < \varepsilon$ for any $\varepsilon > 0$. In this case, we say that Y_n is stochastically bounded.

(b) The remarks (b) - (d) for o and O also hold for o_p and O_p, with obvious modifications.

Lemma 8.14. *We have*
(a) $O(o_p(1)) = O_p(o(1)) = O_p(o_p(1)) = o_p(1)$
(b) $o(O_p(1)) = o_p(O(1)) = o_p(O_p(1)) = o_p(1)$
(c) $o_p(1)O_p(1) = o_p(1)O(1) = o_p(1)O_p(1) = o_p(1)$

8.4.2 Transformations

We introduce in this section some basic tools for asymptotic analysis.

Theorem 8.15. *Let h be a continuous function. Then we have*
(a) If $X_n \xrightarrow{a.s.} X$, then $h(X_n) \xrightarrow{a.s.} h(X)$.
(b) If $X_n \xrightarrow{p} X$, then $h(X_n) \xrightarrow{p} h(X)$.
(c) If $X_n \xrightarrow{d} X$, then $h(X_n) \xrightarrow{d} h(X)$.

Part (c), in particular, is called the *continuous mapping theorem*(CMT).

Proof Part (a) is obvious, since due to the continuity of h

$$X_n(\omega) \to X(\omega) \quad \text{implies} \quad h(X_n(\omega)) \to h(X(\omega))$$

for all ω.

For (b), let $X_n \xrightarrow{p} X$ and $h(X_{n_k})$ be an arbitrary subsequence of $h(X_n)$. Now it follows that

$$\begin{aligned}
X_n \xrightarrow{p} X &\Longrightarrow X_{n_k} \xrightarrow{p} X \\
&\Longrightarrow \exists\{X_{n_{k_i}}\} \text{ such that } X_{n_{k_i}} \xrightarrow{a.s.} X \\
&\Longrightarrow h\left(X_{n_{k_i}}\right) \xrightarrow{a.s.} h(X) \\
&\Longrightarrow h\left(X_{n_{k_i}}\right) \xrightarrow{p} h(X)
\end{aligned}$$

We have thus shown that any subsequence $h(X_{n_k})$ of $h(X_n)$ has a sub-subsequence $h(X_{n_{k_i}})$ such that $P\left\{\left|h\left(X_{n_{k_i}}\right) - h(X)\right| > \varepsilon\right\} \to 0$, which implies that $P\left\{\left|h\left(X_n\right) - h(X)\right| > \varepsilon\right\} \to$ 0 for any $\varepsilon > 0$, or equivalently $h(X_n) \xrightarrow{p} h(X)$.

For part (c), it suffices to show that

$$\mathbf{E}f(h(X_n)) \to \mathbf{E}f(h(X))$$

for $X_n \xrightarrow{d} X$ and f continuous and bounded. This however follows directly from the definition of the convergence in distribution of X_n to X, since $f \circ h$ is continuous and bounded. ■

8.4.3 Random Vectors

In general, the convergence of random vectors is not the same as that of each element in the vector. It is the case for the convergence in probability and the slutsky theorem, while the Cramer-Wold device is useful for the convergence in distribution for random vectors.

Theorem 8.16. *Let $X_n \xrightarrow{d} X$. Then we have*
(a) $X_n = O_p(1)$.
(b) $X_n + o_p(1) \xrightarrow{d} X$.

Proof For part (a), notice that we may choose sufficiently large M such that

$$\mathbf{P}\{|X| > M\} < \varepsilon \quad \text{and} \quad \mathbf{P}\{|X| = M\} = 0$$

since $\{|X| > M\} \downarrow \emptyset$ as $M \uparrow \infty$. Let $f(x) = \mathsf{I}\{|x| > M\}$. Since $X_n \xrightarrow{d} X$ and f bounded and continuous a.s., we have $\mathbf{E}f(X_n) = \mathbf{P}\{|X_n| > M\}$ converges to $\mathbf{E}f(X) = \mathbf{P}\{|X| > M\}$. Therefore, $\mathbf{P}\{|X_n| > M\} < \varepsilon$ for large n.

For (b), let $Y_n = o_p(1)$ and assume that f is uniformly continuous and bounded. It suffices to show that

$$\begin{aligned}|\mathbf{E}f(X_n + Y_n) - \mathbf{E}f(X)| &\le \mathbf{E}\,|f(X_n + Y_n) - f(X_n)| + |\mathbf{E}f(X_n) - \mathbf{E}f(X)| \\ &\to 0\end{aligned}$$

The second term can be made arbitrarily small, since $X_n \xrightarrow{d} X$. To show that the first term is also negligible, we note that

$$|f(X_n + Y_n) - f(X_n)| \le \varepsilon + 2M\,\mathsf{I}\{|Y_n| > \delta\}$$

where $M = \sup|f(x)|$ and ε and δ are chosen as in the proof of Theorem 8.9. The stated result now follows from that $\mathbf{P}\{|Y_n| > \delta\} \to 0$. ■

Corollary 10. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $X_n Y_n \xrightarrow{d} cX$.*

Proof Since $X_n = O_p(1)$ and $Y_n = c + o_p(1)$, we have

$$\begin{aligned} X_n Y_n &= X_n(c + o_p(1)) \\ &= cX_n + O_p(1)o_p(1) \\ &= cX_n + o_p(1) \end{aligned}$$

Apply CMT to get the stated result. ■

Remarks

(a) Let $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then, $X_n + Y_n \xrightarrow{p} X + Y$. However, even if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, it isn't necessarily true that $X_n + Y_n \xrightarrow{d} X + Y$ because we need to take a joint distribution into consideration.

(b) Let $\hat{\theta}_n$ be an estimator of the parameter θ with the true value θ_0. If $\hat{\theta}_n$ is consistent, then

$$\hat{\theta}_n = \theta_0 + o_p(1)$$

If it is asymptotically normal, then

$$\hat{\theta}_n = \theta_0 + O_p\left(\frac{1}{\sqrt{n}}\right)$$

in particular.

Now we introduce the Cramer-Wold device.

Theorem 8.17. *Let* $\{X_n\}$ *be a sequence of a vector of random variables. Then*

$$X_n \xrightarrow{d} X \quad iff \quad \lambda' X_n \xrightarrow{d} \lambda' X$$

for all $\lambda \neq 0$.

8.5 Delta Method

Now we introduce the so-called Δ-*method*, which is useful in obtaining asymptotic distribution of nonlinear functions of $\hat{\theta}_n$ whose distribution is asymptotically normal. Suppose

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathbf{N}(0, \Sigma)$$

and h is differentiable at θ_0. Let $H(\theta) = \partial h(\theta) / \partial \theta'$. It follows from the Taylor expansion of $h(\theta)$ around θ_0 that

$$\begin{aligned} h(\hat{\theta}_n) &= h(\theta_0) + H(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|) \\ &= h(\theta_0) + H(\theta_0)(\hat{\theta}_n - \theta_0) + o_p\left(O_p\left(\frac{1}{\sqrt{n}}\right)\right) \\ &= h(\theta_0) + H(\theta_0)(\hat{\theta}_n - \theta_0) + o_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

which implies

$$\begin{aligned} \sqrt{n}(h(\hat{\theta}_n) - h(\theta_0)) &= H(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \\ &\xrightarrow{d} \mathbf{N}(0, H(\theta_0)\Sigma H(\theta_0)'). \end{aligned}$$

If, for instance, we let $\theta = (\theta_1, \theta_2)'$ and $\hat{\theta}_n = (\hat{\theta}_{n1}, \hat{\theta}_{2n})'$, then the limiting distribution of $\hat{\theta_{n1}}/\hat{\theta_{n2}}$ is given as above with

$$H(\theta) = \frac{\partial h(\theta)}{\partial \theta'} = \left(\frac{\partial h}{\partial \theta_1}, \frac{\partial h}{\partial \theta_2}\right) = \left(\frac{1}{\theta_2}, -\frac{\theta_1}{\theta_2^2}\right)$$

8.6 Law of Large Numbers

Let $\{\xi_i\}$ be a sequence of random variables such that $\mathbf{E}\,\xi_i = 0$. Then under general regularity conditions

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \overset{a.s.\text{or}P}{\rightarrow} 0$$

which is called *law of large numbers* (LLN). It is often referred to as *strong law of large numbers* (SLLN) or *weak law of large numbers* (WLLN), depending upon whether the mode of convergence is a.s. or in $\mathbf{P}$.

Remarks If $\mathbf{E}\,\xi_i = \mu$, LLN can be applied to $\{\zeta_i\}$ with $\zeta_i = \xi_i - \mu$. This then yields

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \overset{a.s.\text{or}P}{\rightarrow} \mu$$

For more general case with $\mathbf{E}\,\xi_i = \mu_i$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i \overset{a.s.\text{or}P}{\rightarrow} 0$$

and if we assume further that $\frac{1}{n}\sum_{i=1}^{n}\mu_i \to \mu$, then $\frac{1}{n}\sum_{i=1}^{n}\xi_i \overset{a.s. \text{or} P}{\to} \mu$.

We now provide different sets of sufficient conditions for WLLN.

1. (Khinchine) $\{\xi_i\}$ is a sequence of $i.i.d\ r.v.'s$ with $\mathsf{E}|\xi_i| < \infty$;

2. (Chevychev)$\{\xi_i\}$ is a sequence of uncorrelated $r.v.'s$ such that var$\xi_i = \sigma_i^2$ and $n^{-2}\sum_i \sigma_i^2 \to 0$;

In fact, the Khinchine's condition implies the strong law (Kolmogorov Theorem). There are variations of these conditions. They in principle impose something like var $\left(\frac{1}{n}\sum_i \xi_i\right) \to 0$.

As an example, we consider the experiment of an infinite toss of a coin, and define a sequence of Bernoulli random variables $\{\xi_n\}$ by $\xi_n = 1$ if the n-th toss is a head. The LLN implies in this specific setting that the sample proportion converges in probability or a.s. to a number, which is perceived as the true probability of getting a head.

8.7 Central Limit Theorem

Let $\{\xi_i\}$ be a sequence of random variables such that $\mathbf{E}\xi_i = 0$. Under regularity conditions, we have

$$\frac{\sum_{i=1}^{n} \xi_i}{\sqrt{\mathrm{var}\left(\sum_{i=1}^{n} \xi_i\right)}} \xrightarrow{d} \mathsf{N}(0,1)$$

which is referred to as *central limit theorem* (CLT). If $\{X_i\}$ is $i.i.d \left(\mu, \sigma^2\right)$, we may write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} \mathbf{N}\left(0,1\right),$$

or

$$\frac{1}{n} \sum_{i=1}^{n} X_i \sim AN\left(\mu, \sigma^2/n\right).$$

Some regularity conditions for CLT for independent sequences with $\mathbf{E}\xi_i^2 = \sigma_i^2$ are:

1. *Lindeberg-Lèvy* condition: the sequence is identically distributed and $\mathsf{E}\left(\xi_i - \mu\right)^2 = \sigma^2 < \infty$.

2. *Lindeberg-Feller* condition: (i) $B_n = \sum_{i=1}^{n} \sigma_i^2 \to \infty$, (ii) $\max_i \frac{\sigma_i^2}{B_n} \to 0$ and (iii) for all $\delta > 0$

$$B_n^{-1} \sum_{i=1}^{n} \mathsf{E}\left[\left(\xi_i - \mu_i\right)^2 1\left\{\left|\xi_i - \mu_{ii}\right| > \delta B_n\right\}\right] \to 0. \tag{1}$$

   The condition (iii) , in particular, is called as the Lindeberg condition, which can be replaced with

3. *Lyapunov* condition: for some $v > 0$, $\mathsf{E}\left|\xi_i - \mu_i\right|^{2+v} < \Delta < \infty$ for all i and $n/B_n^{1+v} = o\left(1\right)$.

Remark The Lyapunov condition is easier to check, but stronger than the Lindeberg condition, as can be easily seen from

$$\sum_{i=1}^{n} \mathsf{E}\,\xi_{ni}^2\, \mathsf{I}\{|\xi_{ni}| > \varepsilon\} \;\leq\; \frac{\sum_{i=1}^{n} \mathsf{E}\left|\xi_{ni}\right|^{2+v}}{\varepsilon^v}.$$

8.8 Asymptotics of Maximum Likelihood Estimation

In this section, we show consistency and asymptotic normality of maximum likelihood estimator (MLE). The tests based on MLE, such as *likelihood ratio* (LR), *Wald* (W) and *Lagrange multiplier* (LM) *tests*, are also introduced. Throughout the section, we let $X_1, \ldots, X_n$ be i.i.d. random variables with the common distribution P. The parameter θ is treated here as a variable taking values from the parameter set Θ, and the true value of θ is denoted by θ_0. Expectation E is an integral operator on $\mathcal{R}$ with respect to P, whose density is given by $p(x, \theta_0)$. We let $\hat{\theta}_n$ be the MLE for θ_0.

Denote by p, ℓ, s, H and I, density, loglikelihood, score, Hessian (and expected Hessian), and (Fisher) information of P. They are defined by

$$\ell(x,\theta) = \log\, p(x,\theta), \quad s(x,\theta) = \frac{\partial}{\partial\theta}\ell(x,\theta), \quad H(x,\theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\ell(x,\theta)$$

and

$$H(\theta) = \mathsf{E}\, H(\cdot,\theta), \quad I(\theta) = \mathsf{E}\, s(\cdot,\theta)s(\cdot,\theta)'$$

Here we use this notation generically, not referring to specific functions. Since X_i's are i.i.d., we have

$$\ell(x_1, \ldots, x_n, \theta) = \sum_{i=1}^{n} \ell(x_i, \theta)$$

Similar results hold also for $s(x_1, \ldots, x_n, \theta)$ and $H(x_1, \ldots, x_n, \theta)$.

8.8.1 Consistency

Consistency often involves more technicality than asymptotic normality. Thus, you may skip this subsection. However, when you have an explicit expression for the MLE, you may proceed more easily. For example, the MLE $\left(\hat{\mu}, \hat{\sigma}^2\right)$ for $\mathbf{N}\left(\mu, \sigma^2\right)$ is $\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} (X_i - \hat{\mu})^2\right)$. We can easily see that $\hat{\mu} \xrightarrow{p} \mu$ by the WLLN. It is left as a homework to show the consistency of $\hat{\sigma}^2$.

First we consider

$$\mathsf{E}\ell(\cdot, \theta) = \int \ell(x, \theta) p(x, \theta_0)\, dx$$

as a function of θ. The following lemma shows that it is maximized at θ_0.

Lemma 8.18. *We have*

$$\mathsf{E}\ell(\cdot,\theta_0) \geq \mathsf{E}\ell(\cdot,\theta)$$

for all $\theta \in \Theta$.

Proof It follows from the Jensen's inequality that

$$\begin{aligned} \mathsf{E}\ell(\cdot,\theta) - \mathsf{E}\ell(\cdot,\theta_0) &= \mathsf{E}\log\frac{p(\cdot,\theta)}{p(\cdot,\theta_0)} \\ &\leq \log\ \mathsf{E}\frac{p(\cdot,\theta)}{p(\cdot,\theta_0)} \\ &= \log\int\frac{p(x,\theta)}{p(x,\theta_0)}p(x,\theta_0)dx = 0 \end{aligned}$$

which completes the the proof. ■

We define

Definition 8.19. For a sequence of random variables $X_n(\theta)$ indexed by $\theta \in \Theta$, if

$$\sup_{\theta \in \Theta} |X_n(\theta) - X(\theta)| = o_p(1),$$

then we say X_n converges in probability to X uniformly on Θ.

A theorem that yields such a convergence is called a uniform law of large numbers (ULLN). They not only require that the LLN holds for each fixed θ but also that the function $\ell(\cdot, \theta)$ is smooth.

Lemma 8.20. *If (i) the data are i.i.d.; (ii) Θ is compact; (iii) $m(w_i, \theta)$ is continuous at each $\theta \in \Theta$ with probability one; and (iv) there is $d(w)$ such that $|m(w, \theta)| \leq d(w)$ for all $\theta \in \Theta$ and $\mathsf{E}(d(w_i)) < \infty$, then $\mathsf{E}(m(w_i, \theta))$ is continuous and*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m(w_i, \theta) - \mathsf{E}(m(w_i, \theta)) \right| \xrightarrow{p} 0.$$

- Condition (iii) is more general than continuity at θ for all w_i and useful.

Theorem 8.21. *Under suitable regularity conditions, we have*

$$\hat{\theta}_n \overset{a.s. \text{or} \mathcal{P}}{\longrightarrow} \theta_0$$

Proof Since X_i's are i.i.d., so are $\ell(X_i, \theta)$'s for any $\theta \in \Theta$. We may thus invoke a LLN for i.i.d. random variables to deduce that

$$\frac{1}{n} \sum_{i=1}^{n} \ell(X_i, \theta) \overset{a.s. \text{or} \mathcal{P}}{\longrightarrow} \mathsf{E}\ell(\cdot, \theta)$$

for all $\theta \in \Theta$. Under certain regularity conditions, the convergence is uniform on Θ, i.e., *uniform law of large numbers* (ULLN) holds. Furthermore, it is known that the argmax operator is continuous. Therefore, we get

$$\text{argmax}_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, \theta) \overset{a.s. \text{or} \mathcal{P}}{\longrightarrow} \text{argmax}_\theta \, \mathsf{E}\ell(\cdot, \theta)$$

The stated result can now be easily deduced. ■

8.8.2 Asymptotic Normality

Theorem 8.22. *Under suitable regularity conditions, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1})$$

Proof Under certain regularity conditions, we may expect to have

1. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(X_i, \theta_0) \xrightarrow{d} \mathbb{N}(0, I(\theta_0))$, by CLT, Cramor-Wold device since for any $\lambda \neq 0, \mathsf{E}\lambda' s\left(X_i, \theta_0\right) = 0$

2. $\frac{1}{n}\sum_{i=1}^{n} H(X_i, \grave{\theta}) \xrightarrow{p} \mathsf{E}\, H(\cdot, \theta_0) = H(\theta_0) = -I(\theta_0)$ if ULLN holds and

$\mathsf{E}H\left(X_i,\theta\right)$ is continuous at $\theta=\theta_0$. That is,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^{n}H(X_i,\grave{\theta})-\mathsf{E}\,H(\cdot,\theta_0)\right|\\&\le\sup_{\theta}\left|\frac{1}{n}\sum_{i=1}^{n}H(X_i,\theta)-\mathsf{E}\,H(\cdot,\theta)\right|+\left|\mathsf{E}\,H(\cdot,\grave{\theta})-\mathsf{E}\,H(\cdot,\theta_0)\right|\\&=o_p\left(1\right)+o\left(1\right),\end{aligned}$$

by the ULLN and the continuity of $\mathsf{E}H\left(X_i,\theta\right)$.

Then,

$$0=\frac{1}{n}\sum_{i=1}^{n}s(X_i,\hat{\theta}_n)=\frac{1}{n}\sum_{i=1}^{n}s(X_i,\theta_0)+\left(\frac{1}{n}\sum_{i=1}^{n}H(X_i,\tilde{\theta})\right)(\hat{\theta}_n-\theta_0).$$

It is important for s to be evaluated at $\theta=\theta_0$ to apply the CLT. Otherwise

E$s(X_i,\theta)$ may not be zero.Therefore,

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n-\theta_0) &= -\left(\frac{1}{n}\sum_{i=1}^n H(X_i,\tilde{\theta})\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta_0)\\ &= -H(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n s(X_i,\theta_0)+o_p(1)\\ &\xrightarrow{d}\mathcal{N}(0,I(\theta_0)^{-1})\end{aligned}$$

by CMT. ■

Remark The MLE achieves the Cramer-Rao lower bound asymptotically, and is therefore often said to be *efficient*. The bound, however, is a finite sample result, and not necessarily hold in asymptotics. We can indeed construct an estimator which has asymptotic variance smaller than the bound. Such an estimator is called *hyper-efficient*. However, it is known that hyper-efficient estimators can be constructed only for certain parameter values, whose collection can be treated negligible.

8.9 Asymptotic Tests based on MLE

We now introduce the likelihood ratio (LR), Lagrange multiplier (LM), and Wald (W) tests, which are based on MLE. It will be shown that their limiting distributions are all chi-squre, and that they are asymptotically equivalent. For simplicity, we consider the hypothesis

$$H_0 : \theta = \theta_0$$

which is to be tested against $\theta \neq \theta_0$. Assume θ is m-dimensional.

Define

Definition 8.23. Let

$$\begin{aligned}
\mathsf{LR} &= 2\left(\sum_{i=1}^{n} \ell(x_i, \hat{\theta}_n) - \sum_{i=1}^{n} \ell(x_i, \theta_0)\right) \\
\mathsf{W} &= \sqrt{n}(\hat{\theta}_n - \theta_0)' I(\hat{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_0) \\
\mathsf{LM} &= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(x_i, \theta_0)\right)' I(\theta_0)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(x_i, \theta_0)\right).
\end{aligned}$$

Remarks

(a) For the LR statistic, notice that the likelihood ratio is given by

$$\lambda(x_1, \ldots, x_n) = \frac{\max\limits_{\theta \in \Theta} p(x_1, \ldots, x_n, \theta)}{p(x_1, \ldots, x_n, \theta_0)}$$

and, therefore, we may write LR $= 2 \log \lambda(x_1, \ldots, x_n)$.

(b) The LR, W and LR statistics are based, respectively, on the ratio of restricted and unrestricted maximum likelihoods, the difference between the estimated and hypothesized values of the parameter, and the first derivative of the likelihood function at the hypothesized value of the parameter. If the hypothesized value is equal to the true value, all three must be small.

We have

Theorem 8.24. *Under suitable regularity conditions,*

$$\text{LR}, \text{W}, \text{LM} \quad \xrightarrow{d} \chi^2_m.$$

Proof Assume the conditions introduced in Theorem 3 hold. Since

$$\ell(x,\theta) = \ell(x,\theta_0) + s(x,\theta_0)'(\theta-\theta_0) + \frac{1}{2}(\theta-\theta_0)'H(x,\theta_0)(\theta-\theta_0) + o(\|\theta-\theta_0\|^2)$$

$$s(x,\theta_0) = s(x,\theta) - H(x,\theta_0)(\theta-\theta_0) + o(\|\theta-\theta_0\|)$$

we have

$$\ell(x,\theta) = \ell(x,\theta_0) + s(x,\theta)'(\theta-\theta_0) - \frac{1}{2}(\theta-\theta_0)'H(x,\theta_0)(\theta-\theta_0) + o(\|\theta-\theta_0\|^2)$$

It follows that

$$\begin{aligned}
&\sum_{i=1}^{n}\ell(X_i,\hat{\theta}_n) - \sum_{i=1}^{n}\ell(X_i,\theta_0)\\
&= -\frac{1}{2}\sqrt{n}(\hat{\theta}_n-\theta_0)'\left(\frac{1}{n}\sum_{i=1}^{n}H(X_i,\theta_0)\right)\sqrt{n}(\hat{\theta}_n-\theta_0) + o_p(1)\\
&= \frac{1}{2}\sqrt{n}(\hat{\theta}_n-\theta_0)'I(\theta_0)\sqrt{n}(\hat{\theta}_n-\theta_0) + o_p(1)
\end{aligned}$$

which yields the result for the LR statistic.

To get the result for the Wald statistic, assume that I is continuous at θ_0 so that

$I(\hat{\theta}_n) = I(\theta_0) + o_p(1)$, and notice that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathbf{N}(0, I(\theta_0)^{-1})$$

For the LM test (or score test), we consider the limiting distribution of the score, i.e.,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i, \theta_0) \xrightarrow{d} \mathbf{N}(0, I(\theta_0))$$

from which the stated result follows directly. ■

Remark Different versions of W are possible with the replacement of $I(\hat{\theta}_n)$ by any one of the following:

(a) $\displaystyle\frac{1}{n} \sum_{i=1}^{n} s(X_i, \hat{\theta}_n) s(X_i, \hat{\theta}_n)'$

(b) $-H(\hat{\theta}_n)$

(c) $\dfrac{1}{n}\sum_{i=1}^{n} H(X_i, \hat{\theta}_n)$

We have

Corollary 11. *The tests based on the statistics LR, W and LM are asymptotically equivalent.*

Proof Observe that

$$\begin{aligned}
&2\left(\sum_{i=1}^{n} \ell(X_i, \hat{\theta}_n) - \sum_{i=1}^{n} \ell(X_i, \theta_0)\right) \\
&= \sqrt{n}(\hat{\theta}_n - \theta_0)' I(\hat{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \\
&= \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i, \theta_0)\right)' I(\theta_0)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s(X_i, \theta_0)\right) + o_p(1)
\end{aligned}$$

from which the stated result is immediate. ■

8.10 t-test and Asymptotic Confidence Interval

Consider a random sample $\{X_i\}$ from a common distribution F. Let $\mu = Ef(X_i) = \int f(x)\, dF$ denote the parameter of interest for a given function f. For simplicity, let f be the identity function $f(x) = x$.

Consider the hypothesis testing for the null hypothesis

$$H_0 : \mu = \mu_0$$

against its negation. A common approach is to use the t-test, which compares an estimator $\hat{\mu}$ with μ_0. Typically, $\hat{\mu} = \bar{X}$ and the closeness of $\bar{X}$ to μ_0 is measured in terms of the distance between the two after normalizing it by the standard deviation $\sigma/\sqrt{n}$ of $\bar{X}$. Thus, the test statistic may be constructed as

$$t = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

if σ is known, and the critical region is given by

$$\mathcal{C} = \{|t| \geq c\}.$$

It is usually the case that the variance σ^2 is not known. Accordingly, the test statistic is modified as

$$t = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$

which is asymptotically standard normal for any consistent estimator $\hat{\sigma}$.

Also, it motivates the common $100\,(1-\alpha)\,\%$ level confidence interval

$$CI = \left[\bar{X} - c_{1-\alpha/2}\hat{\sigma}/\sqrt{n}, \bar{X} + c_{1-\alpha/2}\hat{\sigma}/\sqrt{n}\right],$$

where c_a denotes the a-quantile of the standard normal $\Pr\{N(0,1) \le c_a\} = a$. A justification of the asymptotic confidence interval is that $\Pr\{\mu_0 \in CI\} \to 1-\alpha$ as $n \to \infty$.

8.11 Tests of the Equality of Two Means

Let independent random variables X and Y have normal distributions $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$, respectively. We would like to test the null hypotheses

$$H_0 : \mu_X - \mu_Y = 0$$

against the alternative hypothesis $H_1 : \mu_X - \mu_Y \neq 0$. If we assume respective random samples of sizes n and m, then we can find a test based on the statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)}}.$$

And when σ is unknown, we may consider

$$T = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where $\hat{\sigma}$ is a consistent estimator such as $(n+m)^{-1} \left(\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 + \sum_{j=1}^{m} \left(Y_j - \bar{Y}\right)^2 \right)$.

From the asymptotic perspective, we may assume $Z = DX + (1 - D) Y$ for a dummy variable D and we observe (D_i, Z_i), $i = 1, ..., n$. Also assume that both X and Y are independent and have finite second moments. Note that $E(Z|D) = \mu_X D + \mu_Y (1 - D)$. Then, the null hypothesis can be written as

$$H_0 : \mu_Y := E(Z|D = 0) = E(Z|D = 1) =: \mu_X.$$

Since $E(ZD) = E[E(ZD|D)] = E(Z|D = 1) P(D = 1)$ and $P(D = 1) = E(D)$, method of moment estimators can be cosidered.

$$\hat{\mu}_Y = \frac{\frac{1}{n} \sum_i Z_i (1 - D_i)}{\frac{1}{n} \sum_i (1 - D_i)} \quad \text{and} \quad \hat{\mu}_X = \frac{\frac{1}{n} \sum_i Z_i D_i}{\frac{1}{n} \sum_i D_i}.$$

The asymptotic distributions of the esitmators and that of the difference of the two can be obtained by the Δ-method and the asymptotic normality of

$$\left(\frac{1}{\sqrt{n}} \sum_i (Z_i D_i - ED_i Z_i), \frac{1}{\sqrt{n}} \sum_i (Z_i (1 - D_i) - E(1 - D_i) Z_i), \frac{1}{\sqrt{n}} \sum_i (D_i - ED_i) \right),$$

which follows from the CLT and the Cramer-Wold device.

9 Time Series

A MATRIX BACKGROUND

1. **Basic Definitions and Axioms**

   (a) $A = \underset{p\times q}{A} = (a_{ij})$ is a $p \times q$ matrix (p rows, q columns) with a_{ij} as its (real) element in the i-th row, j-th column, $i = 1, ..., p$; $j = 1, ..., q$.

   (b) $A' = (a_{ji})$ is the transpose of A.

   (c) $cA = (ca_{ij})$, if c is scalar.

   (d) $A + B = (a_{ij} + b_{ij})$ if B is $p \times q$ also.

   (e) $AB =$ if B has q rows.

   (f) $(AB)C = A(BC) = ABC$ (associative property). (But in general matrices don't commute, $AB \neq BA$ in general.)

   (g) $(AB)' = B'A'$.

(h) $r(A)$ is the rank of A, i.e. the number of linearly independent rows (columns).

(i) If $r(A) =$ number of columns (rows) we say A is of full column (row) rank.

(j) $r(AB) \leq$min$(r(A), r(B))$.

(k) If $a_{ij} = 0$, all i, j, we write $A = 0$.

2. Square A

(a) $tr(A) =$, the trace of A.

(b) If $a_{ij} = 0$, all $i \neq j$, we write $A = diag\,(a_{11}, ..., a_{pp})$.

(c) $tr(BC) = tr(CB)$ if BC (and thus CB) are square.

(d) $tr(A') = tr(A)$.

(e) $|A| =$ is the determinant of A, where A_{ij} is the (i, j)-th minor, i.e. the determinant of the $(p-1) \times (p-1)$ matrix formed by deleting the i-th row and j-th column of A.

(f) $|A'| = |A|$.

(g) $|AB| = |A|\,|B|$ if B is $p \times p$ also.

(h) $|cA| = c^p\,|A|$ if c is scalar.

(i) If $|A| = 0$ we say A is singular. Then $r(A) < p$.

(j) If $|A| \neq 0$ we say A is nonsingular. Then $r(A) = p$, and the inverse of A, denoted A^{-1}, exists, such that $AA^{-1} = I_p$, where $I_p = diag(1, ..., 1)$ is the $p \times p$ identity matrix.

(k) The elements of A^{-1} are continuous in A, except at $|A| = 0$.

(l) $(AB)^{-1} = B^{-1}A^{-1}$ if $|A| \neq 0$, $|B| \neq 0$.

(m) $\left|A^{-1}\right| = |A|^{-1}$ if $|A| \neq 0$.

(n) $r(AB) = r(A)$ if $|B| \neq 0$. (True also for rectangular A.)

(o) The zeros $\lambda_1, ..., \lambda_p$ of the polynomial $|A - \lambda I_p| =$ (λ scalar) are called <u>eigenvalues</u> of A.

(p) The zeros of $|A - B\lambda|$ are the eigenvalues of $B^{-1}A$ (and also of AB^{-1}) when $|B| \neq 0$.

(q) Because $A - \lambda_i I_p$ is singular, for λ_i an eigenvalue, there exists a $p \times 1$ vector x_i, called an <u>eigenvector</u> of A, such that $(A - \lambda_i I_p)\, x_i = 0$. In a matrix form, $AX = \Lambda X$, where X is the matrix collecting $x_i's$ and Λ is a diagonal matrix with $\lambda_i's$. The eigenvectors that correspond to the nonzero eigenvalues are linearly independent, which implies that if A is p.d. so is X.

(r) If $A'A = I_p$ we say A is <u>orthogonal</u>. Then $A^{-1} = A'$.

(s) If $A = A'$, we say A is <u>symmetric</u>.

(t) $tr(A) = \sum_{i=1}^{p} \lambda_i$

(u) $|A| = \prod_{i=1}^{p} \lambda_i$

(v) If $\lambda_i \geq 0$, for all i, we say A is non-negative definite and write $A \geq 0$. (This does not mean all elements of A are non-negative.)

(w) If $\lambda_i > 0$, for all i, we say A is positive definite and write $A > 0$.

3. Matrix Decomposition

(a) Eigendecomposition of p.d. matrix $A : A = X \Lambda X^{-1}$, where Λ is the diagonal matrix of the eigenvalues and X is the matrix of eigenvectors.

(b) Eigendecomposition of symmetrix matrix $A : A = X \Lambda X'$, where X is orthogonal, i.e., $X'X = I_m$ and m is the number of nonzero eigenvalues. Thus, if A is symmetric and p.d., $X^{-1} = X'$.

(c) Cholesky decomposition for symmetric p.d. matrix $A : A = U'U$ for an upper triangular matrix U.

4. **Non-Negative Definite A**

   (a) $x'Ax \geq 0$, for $x \neq 0$.

   (b) $B'AB \geq 0$ for all (possibly rectangular) B.

   (c) Thus $B'B \geq 0$, for all (possibly rectangular) B. Then call

   $$\|B\| = \left\{\bar{\lambda}(B'B)\right\}^{1/2}$$

   the <u>norm</u> of B. We have

   $$\|BC\| \leq \|B\| \, \|C\| \, , \quad \|B+C\| \leq \|B\| + \|C\| \, .$$

   (d) $\|A\| = \bar{\lambda}(A)$.

   (e) $\bar{\lambda}(A) \leq tr(A)$.

5. **Positive Definite A**

(a) $x'Ax > 0$, all $x \neq 0$.

(b) If B is of full column rank $B'AB > 0$.

(c) $A^{-1} > 0$.

(d) $r(A) = p$.

(e) $\bar{\lambda}(A) = \left(\underline{\lambda}(A^{-1})\right)^{-1}$.

(f) $\bar{\lambda}(A) < tr(A)$ if $p > 1$.

(g) Let Λ. Then writing $A = X\Lambda X'$,

$$B = X\Lambda^{1/2}X'$$

We call $B = A^{1/2}$ the unique positive definite square root of A.

6. **Partitioned A**

(a)

$$A = \begin{bmatrix} \begin{matrix} A_{11} \\ p_1 \times q_1 \end{matrix} & \vdots & \begin{matrix} A_{12} \\ p_1 \times q_2 \end{matrix} \\ \cdots & \vdots & \cdots \\ \begin{matrix} A_{21} \\ p_2 \times q_1 \end{matrix} & \vdots & \begin{matrix} A_{22} \\ p_2 \times q_2 \end{matrix} \end{bmatrix} \text{ is a } \underline{\textit{partitioned}} \text{ matrix,}$$

where $p_1 + p_2 = p$, $q_1 + q_2 = q$.

(b) If $p_1 = q_1$, $p_2 = q_2$ and $|A_{11}| \neq 0$,

(i) $|A| = |A_{11}|\,|A_{22\cdot 1}|$ where $A_{22\cdot 1} = A_{22} - A_{21}A_{11}^{-1}A_{12}$.

(ii) If also $|A_{22\cdot 1}| \neq 0$

$$A^{-1} = \begin{bmatrix} A_{11}^{-1}\left(I_{p_1} + A_{12}A_{22\cdot 1}A_{21}A_{11}^{-1}\right) & -A_{11}^{-1}A_{12}A_{22\cdot 1}^{-1} \\ -A_{22\cdot 1}^{-1}A_{21}A_{11}^{-1} & A_{22\cdot 1}^{-1} \end{bmatrix}$$

if also $|A_{22\cdot 1}| \neq 0$.

3. **Matrix Differentiation**

(a) $\frac{\partial}{\partial x}x'Ay = Ay$, and $\frac{\partial}{\partial x}y'Ax = A'y$

(b) $\frac{\partial}{\partial x}x'Ax = (A + A')x$.

(c) Let A be a function of a scalar θ. If $|A| \neq 0$ then

i. $\frac{\partial}{\partial \theta}\ell n\,|A| = tr\left(A^{'-1}\frac{\partial A}{\partial \theta}\right)$.

ii. $\frac{\partial}{\partial \theta}A^{-1} = -A^{-1}\frac{\partial A}{\partial \theta}A^{-1}$.

iii. $\frac{\partial^2}{\partial \theta^2}\ell n\,|A| = tr\left(A^{'-1}\frac{\partial^2 A}{\partial \theta^2} - A^{'-1}\frac{\partial A'}{\partial \theta}A^{'-1}\frac{\partial A}{\partial \theta}\right)$.

iv. $\frac{\partial^2}{\partial \theta^2}A^{-1} = A^{-1}\left(\frac{\partial A}{\partial \theta}A^{-1}\frac{\partial A}{\partial \theta} - \frac{\partial^2 A}{\partial \theta^2}\right)A^{-1}$.