Advanced Statistical Theory I

Lecturer: Lin Zhenhua

Semester I, AY2019/2020

Contents

Lecture 1: Review of Probability Theory

Lecturer: LIN Zhenhua ST5215 AY2019/2020 Semester I

1.1 Topological Spaces and Continuity

Topology

- It is well known that, the interval of the form $(a, b) \subset \mathcal{R}$ is called an *open* interval, while the interval$[a, b]$ is called an *closed* interval. We also learn that a function $f : \mathcal{R} \to \mathcal{R}$ is *continuous*, if for every $x \in \mathcal{R}$, for every $\epsilon > 0$, there exists a $\delta > 0$ such that, for all $y \in \mathcal{R}$ satisfying $|y-x| < \delta$, then $|f(x)-f(y)| < \epsilon$. All these concepts, open, closed and continuous, are topological concepts.
- In fact, in topoloyg, one concerns with the properties that are preserved/invariant under continuous deformations/functions.
- Sometimes, one needs to deal with objects other than real numbers or even Euclidean space $\mathcal{R}^d$. It is important to generalize these concepts to general spaces. The generalization in turn will deepen our understanding of the usual Euclidean spaces.
- We briefly review some basic topological concepts in their most general form. Further information about topology can be found in Munkres (2000).

Definition 1.1. A topology on a set S is a collection $\mathcal{T}$ of subsets of S such that

1. The empty set is in $\mathcal{T}$, i.e. $\emptyset \in \mathcal{T}$;
2. If $\mathcal{A} \subset \mathcal{T}$, then $\bigcup_{A \in \mathcal{A}} A \in \mathcal{T}$;
3. If $\mathcal{A} \subset \mathcal{T}$ and the cardinality of $\mathcal{A}$ is finite, then $\bigcap_{A \in \mathcal{A}} A \in \mathcal{T}$.

S is called a topological space if a topology on it has been specified. Elements in $\mathcal{T}$ (recall that these elements are subsets of S) are called *open sets*. If A is an open set, then its complement A^c is called a *closed set*.

- Note that, a topology has two components
  - a set of objects (S in the above defintion), and
  - a structure ($\mathcal{T}$ in the above defintion) about the set.
- This pattern is very common in mathematics: a space is often a set of objects endowed with certain structure.

- Notationally, the component $\mathcal{T}$ is often omitted if it is clear from the context.

- Given a set S, how to introduce a topology $\mathcal{T}$ on it?
  - approach 1: Enumerate all open sets, and make sure they satisfy the conditions listed in Definition (1.1).
  - approach 2: Declare some "seed" subsets of S as open sets and then specify a topology on S as the "smallest" topology containing the seed subsets.
  - These seed subsets are called *basis elements* and the collection of basis elements are call a *basis* for the topology it induces.

Definition 1.2. A *basis* for a set S is a collection $\mathcal{B}$ of subsets of S such that

1. If $x \in S$, then there is $B \in \mathcal{B}$ such that $x \in B$;
2. If $x \in B_1 \cap B_2$ and $B_1, B_2 \in \mathcal{B}$, then there exists $B_3 \in \mathcal{B}$ such that $B_3 \subset B_1 \cap B_2$.

The elements in $\mathcal{B}$ are called *basis elements*.

- Let $\mathcal{B}$ be a basis for S, and define $\mathcal{T}$ to be the collection of all unions of elements of $\mathcal{B}$. One can check that $\mathcal{T}$ is a topology on S.
- "Smallest": if $\mathcal{G}$ is a topology on S such that $\mathcal{B} \subset \mathcal{G}$, then $\mathcal{T} \subset \mathcal{G}$.
- We say that $\mathcal{T}$ is generated by the basis $\mathcal{B}$.

Example 1.3. Let $S = \mathcal{R}$ and $\mathcal{B} = \{(a, b) : -\infty < a < b < +\infty\}$. We can check that $\mathcal{B}$ is a basis. The topology generated by $\mathcal{B}$ is the standard/canonical topology on the real line $\mathcal{R}$.

Example 1.4. Let $S = \mathcal{R}^d$ and $B_x(\epsilon) = \{y \in \mathcal{R}^d : \|x - y\|_2 < \epsilon\}$. The topology generated by $\mathcal{B} = \{B_x(\epsilon) : x \in \mathcal{R}^d, \epsilon > 0\}$ is called the standard/canonical topology on $\mathcal{R}^d$.

- When we talk about $\mathcal{R}^d$, by default, we assume the standard topology on it.

Continuous Functions

- In multivariate calculus, we learned continuous functions $f : \mathcal{R}^d \to \mathcal{R}$ via the $\delta - \epsilon$ language
- From the perspective of topology, they are indeed functions between two topological spaces, namely, $\mathcal{R}^d$ and $\mathcal{R}$
- More generally, consider functions $f : S \to V$ between two general topological spaces $(S, \mathcal{T})$ and $(V, \mathcal{V})$.
- Define $f^{-1}(B) = \{x \in S : f(x) \in B\}$, called the preimage of B under f

Definition 1.5. A function $f : S \to V$ between topological spaces $(S, \mathcal{T})$ and $(V, \mathcal{V})$ is *continuous* if and only if for any open set $B \in \mathcal{V}$, the preimage $f^{-1}(B)$ belongs to $\mathcal{T}$, i.e. $f^{-1}(B) \in \mathcal{T}$. If f is bijective and both f and its inverse f^{-1} are continuous, then we say f is a *homeomorphism*.

- We say S is homeomorphic to V if there exists a homeomorphism between them.
- Homeomorphic spaces share the same topological properties.
- Topological properties are properties that are invariant under continuous deformation/functions.
- For example, compactness is a topological property, and a continuous function f preserves compactness.

Definition 1.6. A collection $\mathcal{A}$ of subsets of S is said to *cover* A, or to be a *covering* of A, if $A \subset \bigcup \mathcal{A}$. If $\mathcal{A}$ is a covering of A and all elements in $\mathcal{A}$ are open, then $\mathcal{A}$ is an *open covering* of A. A subset A of S is said to be *compact* if every open covering of A contains a finite subcollection that also covers A.

- Examples of compact subsets of $\mathcal{R}$ (endowed with the canonical topology) are closed intervals of the form $[a, b]$.
- In fact, all closed subsets of finite diameter of $\mathcal{R}^d$ are compact.
- We already know that if $f : [a, b] \to \mathcal{R}$ is continuous, then f has a maximum and a minimum value on $[a, b]$.
- This generalizes to any compact subset of a general topological space: If $f : A \to \mathcal{R}$ is continuous and $A \subset S$ is compact, then f attains its extreme value (either maximum or minimum) at some element of A.

1.2 Measure Spaces, Borel Sets and Probability Spaces

- Probability theory is essential for mathematical statistics, and is based on measure theory.
- We now briefly introduce some general concepts from measure theory and then specialize them to probability theory.
- Let Ω be a set of objects. In probability theory, this will be our sample space. For the moment, we treat it as a general set of objects.
- Now, we want to measure the "size" of subsets of Ω.
  - For example, if $\Omega = \mathcal{R}$ and $A = [a, b]$, then the size of A is naturally defined as its length $b - a$.
  - If $\Omega = \mathcal{R}^2$ and A is a polygon, we can measure its size by its area.

  - Similarly, if $\Omega = \mathcal{R}^3$ and A is a bounded subset, we might measure its size by its volume.

- In all of these examples, the measure is a set function ν that maps a subset of $\mathcal{R}$, $\mathcal{R}^2$ or $\mathcal{R}^3$ to a (nonnegative) real number
  - e.g. $\nu([a,b]) = b - a$
- Generalize this concept to a set of general objects? Infinite ways to do so!
- But, what properties we expect from such a generalization?
  - What do we expect from "size"?
- Intuition 1 (finite additivity): if $A \subset \Omega$ and $B \subset \Omega$ are disjoint, then the size of their union $A \cup B$ shall be equal to the sum of the size of A and the size of B.
  - If $A \cap B = \emptyset$, then $\nu(A) + \nu(B) = \nu(A \cup B)$
  - More generally, if $A_1, \ldots, A_k$ are disjoint, then $\sum_{i=1}^{k} \nu(A_i) = \nu\left(\bigcup_{i=1}^{k} A_i\right)$.
- Intuition 2 ("empty" has zero measure): $\nu(\emptyset) = 0$.
- Attempt: define measure ν on S as a set function satisfying the above intuitions.
- In addition, when $S = \mathcal{R}, \mathcal{R}^2, \mathcal{R}^3, \ldots$, the length/area/volume shall be a measure
- One quick question: can we define measure ν for all subsets of S? In other words, can we measure the "size" of each subset of S, while the above two intuitions still hold? The answer is
  - yes, if the set S is countable
  - no, if the set is uncountable
- Why?
  - important feature of length/area/volume: the congruence invariance. For example, for $\Omega = \mathcal{R}^3$, $m(A) = m(x + A)$ for all $x \in \mathcal{R}^3$, where m denotes the volume.
- Banach-Tarski Paradox: this provides an example that not every subset of $\mathcal{R}^3$ has a Lebesgue measure: A ball B in $\mathcal{R}^3$ can be partitioned into two disjoint subsets B_1 and B_2 such that, each of this subsets can be further divided into several pieces, and these pieces, after some translation, rotation and reflection operations, together form a new ball that is identical to the original ball. This implies that, $B = B_1 \cup B_2$ and $m(B) = m(B_1) = m(B_2)$!
  - This paradox implies that we cannot define volume for every subset of $\mathcal{R}^3$.
- We are forced to declare some subsets to have volume and some not. Those with volume are called measurable subsets.

- More generally, before we can measure the size of subsets of a general set Ω, we need to specify which subsets are measurable, and these measurable subsets shall allow us to define a "measure" on them, e.g. a measure that satisfies finite additivity.

Definition 1.7. A collection $\mathcal{F}$ of subsets of a set Ω is called a *σ-field* (or *σ-algebra*) if

1. $\emptyset \in \mathcal{F}$,
2. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$,
3. if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$,then $\bigcup A_i \in \mathcal{F}$.

A pair $(\Omega, \mathcal{F})$ of a set Ω and a σ-field $\mathcal{F}$ on it is called a *measurable space*.

Definition 1.8. A (positive) measure ν on a measurable space $(\Omega, \mathcal{F})$ is a nonnegative function $\nu : \mathcal{F} \to \mathcal{R}$ such that

1. (nonnegativity) $0 \leq \nu(A) \leq \infty$ for all $A \in \mathcal{F}$,
2. (empty is zero) $\nu(\emptyset) = 0$, and
3. (σ-additivity): $\sum_{i=1}^{\infty} \nu(A_i) = \nu\left(\bigcup_{i=1}^{\infty} A_i\right)$ if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$ and $A_1, A_2, \ldots$ are disjoint.

The triple $(\Omega, \mathcal{F}, \nu)$ is called a *measure space*.

- There are many ways to define a σ-field and a measure on a given set.
- For $\mathcal{R}$, we want open intervals (a, b) to be measurable and the measure is $b - a$.
- More generally, we want all open sets to be measurable. The "smallest" σ-field that contains all open sets of $\mathcal{R}$ is called the *Borel* σ-field.
- This generalizes to any topological space: For a topological space S, the smallest σ-field containing all open sets is called the Borel σ-field of S. The elements of a Borel σ-field are called *Borel sets*.

Exercise 1.9. Let $\mathcal{A}$ be a collection of subsets of Ω. Show that there exists a σ-field $\mathcal{F}$ such that $\mathcal{A} \subset \mathcal{F}$ and if $\mathcal{E}$ is a σ-field that also contains $\mathcal{A}$, then $\mathcal{F} \subset \mathcal{E}$. In this sense, such σ-field is the smallest one containing $\mathcal{A}$. It is often denoted by $\sigma(\mathcal{A})$ and said to be generated by $\mathcal{A}$.

Example 1.10 (Lebesgue measure on $\mathcal{R}$)**.** Let $\mathcal{B}$ be the Borel σ-field on $\mathcal{R}$. By definition, this is the smallest σ-field that contains all open sets of $\mathcal{R}$. There exists a unique measure m on $(\mathcal{R}, \mathcal{B})$ that satisfies $m([a, b]) = b - a$. This is called the *Lebesgue measure* on $\mathcal{R}$. It is the standard/canonical measure on $\mathcal{R}$. When $\mathcal{R}$ is mentioned, without otherwise explicitly mentioned, it is by default endowed with such Borel σ-field and Lebesgue measure.

- Note that $m(\{a\}) = 0$ for any $a \in \mathcal{R}$
- More generally, $m(A) = 0$ if A is countable (note that any countable set of $\mathcal{R}$ is measurable)

Example 1.11 (Counting measure)**.** Let $\mathcal{F}$ be the collection of all subsets of Ω, and $\nu(A) = |A|$ if $|A| < \infty$ and $\nu(A) = \infty$ if $|A| = \infty$. This measure is called the *counting measure* on $(\Omega, \mathcal{F})$.

Example 1.12 (Point mass)**.** Let $x \in \Omega$ be a fixed point. Define

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

- How to introduce a measure on a product space $\Omega_1 \times \cdots \times \Omega_d$, like, $\mathcal{R}^d = \mathcal{R} \times \cdots \times \mathcal{R}$?
- For a product space $\Omega_1 \times \cdots \times \Omega_d$,where each Ω_i is endowed with a σ-field $\mathcal{F}_i$, the σ-field generated by $\prod_{i=1}^d \mathcal{F}_i = \{A_1 \times \cdots A_d : A_i \in \mathcal{F}_i\}$ is called the *product σ-field*.
- For $\mathcal{R}^d$, the product σ-field is the same as its Borel σ-field.
- A measure ν on $(\Omega, \mathcal{F})$ is said to be *σ-finite* if there exists a countable number of measurable sets $A_1, A_2, \ldots$ such that $\bigcup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i.
- The Lebesgue measure is clearly σ-finite, since $\mathcal{R} = \bigcup A_i$ with $A_i = [-i, i]$ and $m(A_i) = 2i < \infty$.

Proposition 1.13. *Suppose $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, 2, \ldots, d$, are measure spaces and $\nu_1, \ldots, \nu_d$ are all σ-finite. There exists a unique σ-finite measure on the product σ-field, denoted by $\nu_1 \times \cdots \times \nu_d$, such that*

$$\nu_1 \times \cdots \times \nu_d(A_1 \times \cdots \times A_d) = \prod_{i=1}^d \nu(A_i)$$

for all $A_i \in \mathcal{F}_i$.

- σ-finite is required in some important theorems (Radon-Nikodym, Fubin's). So we only focus on σ-finite measures in this course. In particular, all finite measures are σ-finite.

Example 1.14 (Lebesgue measure on $\mathcal{R}^d$)**.** For $\mathcal{R}^d$, the unique product measure is called the Lebesgue measure on $\mathcal{R}^d$ on the Borel σ-field $\mathcal{B}^d$ on $\mathcal{R}^d$. It is the standard/canonical measure on $\mathcal{R}^d$. Again, without otherwise explicitly mentioned, $\mathcal{R}^d$ is endowed with such Borel σ-field and Lebesgue measure.

- A probability space is a special measure space

Definition 1.15. A measure space $(\Omega, \mathcal{F}, \nu)$ is called a *probability space* if $\nu(\Omega) = 1$. In this case, Ω is called a sample space, the elements of $\mathcal{F}$ are called events, and the measure ν is called a probability measure. The number $\nu(A)$ is interpreted as the probability of the event A to happen.

- In probability theory, the probability measure ν is often denoted by P or Pr.
- Like continuous functions between topological spaces, for two measurable spaces, we want to study functions between them that preserve measure properties, like measurability, etc.
- This is a quite common pattern: for a category of spaces of the same kind, there are functions between them that preserve the space structure
  - for the category of topological spaces, they are continuous funcitons
  - for the category of measurable spaces, they are measurable functions
  - for the category of linear spaces, they are linear transformations

Definition 1.16. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be two measurable spaces and $f : \Omega \to \Lambda$ a function. The function f is called a *measurable function* if and only if $f^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{G}$. When $\Lambda = \mathcal{R}$ and $\mathcal{G}$ is the Borel σ-field, then we say f is *Borel measurable* or a *Borel function* on $(\Omega, \mathcal{F})$.

- In probability theory, a measurable function is also called a *random element*, and often denoted by capital letters X, Y, Z, ... If X is real-valued, then it is called a *random variable*; if it is vector-valued, then it is called a *random vector*.

Exercise 1.17. Check that the indicator function I_A for a measurable set A is a Borel function. Here,

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

More generally, a simple function of the form

$$f(\omega) = \sum_{i=1}^{k} a_i I_{A_i}(\omega) \tag{1.1}$$

is also a Borel function for any real numbers $a_1, \ldots, a_k$ and measurable sets $A_1, \ldots, A_k$.

- Note: when we say $A_1, A_2, \ldots$ are measurable without explicitly mentioning a measurable space, we often assume a common measurable space, such as $(\Omega, \mathcal{F})$.

1.3 Integration and Expectation

- In calculus, the integral of a continuous function is defined as the limit of a Riemann sum. For example,
  - let f be a continuous function defined on the interval $[0, 1]$.
  - chop the interval into subintervals of equal length, say $D_{ni} = [(i-1)/n, i/n]$ for some n and $i = 1, 2, \ldots, n$.

  - Let $a_{ni} = \min\{f(x) : x \in D_{ni}\}$ and $b_{ni} = \max\{f(x) : x \in D_{ni}\}$.
  - Define $A_n = \sum_{i=1}^n a_{ni}/n = \sum_{i=1}^n a_{ni} m(D_{ni})$ and similarly $B_n = \sum_{i=1}^n b_{ni}/n = \sum_{i=1}^n a_{ni} m(D_{ni})$.
  - For a continuous function, one can show that $A_n \to c$ and $B_n \to c$, and this common c is defined as the Riemann integral of f on $[0,1]$ and denoted by $\int_0^1 f(x)\mathrm{d}x$ or simply $\int f$ when the domain is known from the context.

- We can see that the Riemann integral is a kind of average of the value of f over some domain/interval.
- In statistics, we also want to express the concept of average, but for all random variables which might not be continuous at all.
  - We need to generalize Riemann integral to the so-called Lebesgue integral, as follows.
- We do it in three steps:
  - step 1: define Lebesgue integral on "simple" functions – easy case
  - step 2: use integral of simple functions to approximate integral of nonnegative Borel functions
  - step 3: define integral for all Borel functions
- Let us fix a σ-finite measure space $(\Omega, \mathcal{F}, \nu)$.

Integral of a nonnegative simple function

- Suppose $f : \Omega \to \mathcal{R}$ is a simple nonnegative function: $f(x) = \sum_{i=1}^k a_i I_{A_i}(x)$ for $A_i \in \mathcal{F}$ and $a_i \geq 0$.
- It is quite intuitive and straightforward to define the integral (average) of f as $\int f\mathrm{d}\nu = \sum_{i=1}^k a_i \nu(A_i)$.
- This is well defined even when $\nu(A_i) = \infty$ for some A_i, since $a\infty = \infty$ when $a > 0$ and $a\infty = 0$ when $a = 0$.
- Note that $\int f\mathrm{d}\nu = \infty$ is possible and allowed.

Integral of a nonnegative Borel function

- For a general Borel function, it is difficult to define an integral directly.
- Note that, a Borel function can be approximated by simple functions to any arbitrary precision (in certain sense)
- Since we have integrals for simple functions, we shall use the integrals of these simple functions as proxy of the integral of the Borel function.

- Let $\mathcal{S}_f$ be the collection of all nonnegative simple functions of the form (1.1) such that $g(\omega) \leq f(\omega)$ for all $\omega \in \Omega$ if $g \in \mathcal{S}_f$. Intuitively, functions in $\mathcal{S}_f$ approximate f from below.
- Define the integral of f as

$$\int f \mathrm{d}\nu = \sup\{\int g \mathrm{d}\nu : g \in \mathcal{S}_f\}. \tag{1.2}$$

- compare to the definition of Riemann integral of a continuous function f:
  - chop the interval into subintervals of equal length, say $D_{ni} = [(i-1)/n, i/n)$ for some n and $i = 1, 2, \ldots, n$.
  - Let $a_{ni} = \min\{f(x) : x \in D_{ni}\}$ and $b_{ni} = \max\{f(x) : x \in D_{ni}\}$.
  - Define $A_n = \sum_{i=1}^n a_{ni}/n = \sum_{i=1}^n a_{ni} m(D_{ni})$ and similarly $B_n = \sum_{i=1}^n b_{ni}/n = \sum_{i=1}^n a_{ni} m(D_{ni})$.
  - Let $g_n(x) = a_{ni}$ if $x \in D_{ni}$, and $h_n(x) = b_{ni}$ if $x \in D_{ni}$.
  - These g_n and h_n are simple functions!
  - Also $g_n(x) \leq f(x) \leq h_n(x)$
  - $\int f \mathrm{d}\nu = \lim_{n\to\infty} \int g_n \mathrm{d}\nu = \lim_{n\to\infty} \int h_n \mathrm{d}\nu$ since f is continuous.
  - For a continuous f, the Riemann integral is equal to its Lebesgue integral

Integral of a arbitrary Borel function

- Divide f into two parts
  - positive part: $f_+(x) = \max\{f(x), 0\}$
  - negative part: $f_-(x) = -\min\{f(x), 0\} = \max\{-f(x), 0\}$.
  - note that the negative part is also a nonnegative function
- $f = f_+ - f_{-1}$
- Define $\int f \mathrm{d}\nu$ as

$$\int f \mathrm{d}\nu = \int f_+ \mathrm{d}\nu - \int f_- \mathrm{d}\nu$$

  if at least one of $\int f_+ \mathrm{d}\nu$ and $\int f_- \mathrm{d}\nu$ is finite.
  - if yes, we say the integral of f exists
  - if not, then we can the integral of f does not exist
- When both $\int f_+ \mathrm{d}\nu$ and $\int f_- \mathrm{d}\nu$ are finite, we say f is integrable.
- Sometimes, we only want to see the average of f over a subset A of Ω. The above definition is for the whole domain Ω. Then how?
- Note that I_A is measurable, and so is the product $I_A f$. Note that $(I_A f)(x) = I_A(x) f(x)$.

- If the integral of $I_A f$ exists, then we can define
$$\int_A f \mathrm{d}\nu = \int I_A f \mathrm{d}\nu.$$
- Notation: $\int f \mathrm{d}\nu = \int_\Omega f \mathrm{d}\nu = \int f(x) \mathrm{d}\nu(x) = \int f(x)\nu(\mathrm{d}x)$
- If ν is a probabiltiy measure, $\int X \mathrm{d}P = \mathbb{E}X = \mathbb{E}(X)$, and called the expectation of X

Change of variables

Let f be measurable from $(\Omega, \mathcal{F}, \nu)$ to $(\Lambda, \mathcal{G})$. Then f induces a measure on Λ, denoted by $\nu \circ f^{-1}$ and defined by
$$\nu \circ f^{-1}(B) = \nu(f^{-1}(B)) \qquad \forall B \in \mathcal{G}.$$

- when $\nu = P$ is a probability measure and $\Lambda = \mathcal{R}$ and $f = X$ is a random variable, then $P \circ X^{-1}$ is often denoted by P_X
- P_X is called the law or the distribution of X
- The CDF of X is denoted by F_X and defined by $F_X(x) = P(X \leq x)$.
- sometimes, we also use F_X in the place of P_X

Theorem 1.18 (Change of variables)**.** *The integral of Borel function $g \circ f$ is computed via*
$$\int_\Omega g \circ f \mathrm{d}\nu = \int_\Lambda g \mathrm{d}(\nu \circ f^{-1}).$$

- Application:
  - Ω : a general measurable space
  - $\Lambda = \mathcal{R}$
  - X: a random variable defined on Ω.
  - $\mathbb{E}X$ is not easy to computed, but $\int_\mathcal{R} x \mathrm{d}P_X$ might be easy. We then compute $\mathbb{E}X = \int_\Omega X \mathrm{d}P = \int_\mathcal{R} x \mathrm{d}P_X = \int_\mathcal{R} x \mathrm{d}F_X$.
  - In this example, g is the identity function $g(x) = x$.

Properties of expectation/integral

- Assume the expectation of random variables below exists
- Linearity: $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$ when $\mathbb{E}X$, $\mathbb{E}Y$ and $\mathbb{E}(aX + bY)$ exist.
- $\mathbb{E}X$ is finite if and only if $\mathbb{E}|X|$ is finite

- a.e. (almost everywhere) and a.s. statements: A statement holds ν-a.e. (or simple a.e.) if it holds for all ω in A^c with $\nu(A) = 0$ for some (measurable) A.
  - if ν is a probability, the a.e. is often written as a.s. (almost surely)
  - e.g. Let $f(x) = x^2$, then $f(x) > 0$ m-a.e. (recall: m denotes the Lebesgue measure on $\mathcal{R}$): $f(x) = 0$ iff $x = 0$, and $m(\{0\}) = 0$.
- if $X \leq Y$ a.s., then $\mathbb{E}X \leq \mathbb{E}Y$
  - if $X \geq 0$ a.s., then $\mathbb{E}X \geq 0$
  - $|\mathbb{E}X| \leq \mathbb{E}|X|$
- If $X \geq 0$ a.s., and $\mathbb{E}X = 0$, then $X = 0$ a.s.
  - If $X = Y$ a.s., then $\mathbb{E}X = \mathbb{E}Y$

Theorem 1.19. *Let $f_1, \ldots$ be a sequence of Borel functions on $(\Omega, \mathcal{F}, \nu)$.*

- *Fatou's lemma: If $f_n \geq 0$, then*
$$\int \liminf_n f_n \mathrm{d}\nu \leq \liminf_n \int f_n \mathrm{d}\nu.$$
- *Dominated convergence theorem: If $\lim_{n\to\infty} f_n = f$ a.e. and there exists an integrable function g such that $|f_n| \leq g$ a.e., then*
$$\int \lim_n f_n \mathrm{d}v = \lim_n \int f_n \mathrm{d}\nu.$$
- *Monotone convergence theorem: If $0 \leq f_1 \leq \cdots$ and $\lim_n f_n = f$ a.e., then*
$$\int \lim_n f_n \mathrm{d}\nu = \lim_n \int f_n \mathrm{d}\nu.$$

Theorem 1.20 (Fubini)**.** *Let ν_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$,$i = 1, 2$, and let f be a Borel function on $\prod_{i=1}^2 \Omega_i$ endowed with the product σ-field. Suppose that either $f \geq 0$ or $\int |f| \mathrm{d}(\nu_1 \times \nu_2) < \infty$. Then*
$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) \mathrm{d}\nu_1(\omega_1)$$
exists ν_2-a.e. and is a Borel function on Ω_2 whose integral exists, and
$$\begin{aligned}\int_{\Omega_1 \times \Omega_2} f \mathrm{d}(\nu_1 \times \nu_2) &= \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2) \mathrm{d}\nu_1(\omega_1) \right] \mathrm{d}\nu_2(\omega_2) \\ &= \int_{\Omega_1} \left[\int_{\Omega_2} f(\omega_1, \omega_2) \mathrm{d}\nu_2(\omega_2) \right] \mathrm{d}\nu_1(\omega_1).\end{aligned}$$

Example 1.21. Let $\Omega_1 = \Omega_2 = \{1, 2, \ldots\}$,and $\nu_1 = \nu_2$ be the counting measure. A function a on $\Omega_1 \times \Omega_2$ defines a double sequence, and $a(i, j)$ is often denoted by a_{ij}. If either $a_{ij} \geq 0$ for all i, j or $\int |a| \mathrm{d}(\nu_1 \times \nu_2) = \sum_{ij} |a_{ij}| < \infty$, then
$$\sum_{ij} a_{ij} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{ij}.$$

1.4 Radon-Nikodym derivative and probability density

- We learned PDF (probability density function) as the derivative of CDF (cummulative distribution function) of a random variable
- This is a special case of Radon-Nikodym derivative
- Let λ and ν be two measures on a measurable space $(\Omega, \mathcal{F})$
- We say λ is *absolutely continuous* w.r.t. ν and write $\lambda \ll \nu$ iff

$$\nu(A) = 0 \qquad \text{implies} \qquad \lambda(A) = 0.$$

Exercise 1.22. Check that the measure λ defined by

$$\lambda(A) := \int_A f \mathrm{d}\nu, A \in \mathcal{F}$$

for a nonegative Borel function f is absolutely continuous w.r.t. ν.

- Conversely, if $\lambda \ll \nu$, then there exists a Borel function f such that $\lambda(A) = \int_A f \mathrm{d}\nu$, $A \in \mathcal{F}$.

Theorem 1.23 (Radon-Nikodym)**.** *Let ν and λ be two measures on $(\Omega, \mathcal{F})$ and ν be σ-finite. If $\lambda \ll \nu$, then there exists a nonnegative Borel function f on Ω such that*

$$\lambda(A) = \int_A f \mathrm{d}\nu, \qquad A \in \mathcal{F}.$$

In addition, f is unique ν-a.e., i.e., if $\lambda(A) = \int_A g \mathrm{d}\nu$ for any $A \in \mathcal{F}$, then $f = g$ ν-a.e.

- The function f above is called the *Radon-Nikodym derivative* or density of λ w.r.t. ν, and denoted by $\frac{\mathrm{d}\lambda}{\mathrm{d}\nu}$.

Example 1.24 (Probability density)**.** Let $\lambda = F$ be a probability measure on $\mathcal{R}$, i.e., a probability distribution/law of a random variable, and $\nu = m$ the Lebesgue measure. If $F \ll m$, then it has a probability density f w.r.t. m. Such f is called PDF of F. In particular, when F has a derivative in the usual sense of calculus, then

$$F(x) = \int_{-\infty}^{x} f(y) \mathrm{d}m(y) = \int_{-\infty}^{x} f(y) \mathrm{d}y.$$

In this case, Radon-Nikodym derivative is the same as the usual derivative in calculus.

- A PDF w.r.t. Lebesgue measure is called a *Lebesgue PDF*.

Example 1.25 (Discrete CDF and PDF)**.** Let $a_1 < a_2 < \cdots$ be a sequence of real numbers and X a random variable that $X \in \Lambda = \{a_1, \ldots\}$. Let $p_n = P(X = a_n)$. Then the CDF of X is

$$F(x) = \begin{cases} \sum_{i=1}^{n} p_i & a_n \le x < a_{n+1}, \\ 0 & -\infty < x < a_1. \end{cases}$$

This CDF is a stepwise function, and called a discrete CDF. The corresponding probability measure on Λ is given by

$$P_X(A) = \sum_{i:a_i \in A} p_i, A \in \mathcal{F} = \{B : B \subset \Lambda\} \text{ (power set of}\Lambda).$$

Suppose ν is the counting measure on $\mathcal{F}$. Then

$$P_X(A) = \int_A f \mathrm{d}\nu$$

with $f(a_i) = p_i$. Here f is defined on Λ. This f is the PDF of P w.r.t. the counting measure ν.

- Any discrete CDF has a PDF w.r.t. the counting measure, and such PDF is called discrete PDF (or PMF, probability mass function).

- Properties of Radon-Nikodym derivatives: Let ν be a σ-finite measure on a measurable space $(\Omega, \mathcal{F})$. Suppose all other measures discussed below are also defined on $(\Omega, \mathcal{F})$
  - If $\lambda \ll \nu$ and $f \geq 0$, then
  $$\int f \mathrm{d}\lambda = \int f \frac{\mathrm{d}\lambda}{\mathrm{d}\nu} \mathrm{d}\nu.$$
  - If $\lambda_i \ll \nu$, then $\lambda_1 + \lambda_2 \ll \nu$ and
  $$\frac{\mathrm{d}(\lambda_1 + \lambda_2)}{\mathrm{d}\nu} = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu} + \frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu} \qquad \nu\text{-a.e.}$$
  - Chain rule: If λ is σ-finite and $\tau \ll \lambda \ll \nu$, then
  $$\frac{\mathrm{d}\tau}{\mathrm{d}\nu} = \frac{\mathrm{d}\tau}{\mathrm{d}\lambda} \frac{\mathrm{d}\lambda}{\mathrm{d}\nu} \qquad \nu\text{-a.e.}$$
  In particular, if $\lambda \ll \nu$ and $\nu \ll \lambda$,then let $\tau = \nu$ in the above, and we have
  $$\frac{\mathrm{d}\lambda}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right)^{-1} \qquad \nu \text{ or } \lambda\text{-a.e.}$$

1.5 Moment Inequalities

- In (mathematical) statistics, we often need to control the tail of the distribution of random variables
- e.g. no too heavy probability mass is placed on very "large" values of a random variable
- This intuition is sometimes expressed as a condition on the "moment" of a random variable

We have following definitions of moments of a random variable X:

- If $\mathbb{E}|X|^p < \infty$ for some real number p, then $\mathbb{E}|X|^p$ is called the pth *absolute moment* of X or its law P_X
- If $\mathbb{E}X^k$ is finite, where k is a positive integer, then $\mathbb{E}X^k$ is called the kth *moment* of X or P_X
  - when $k = 1$, it is the expectation of X we introduced previously
- If $\mu = \mathbb{E}X$ and $\mathbb{E}(X-\mu)^k$ are finite for a positive integer k, then $\mathbb{E}(X-\mu)^k$ is called the kth *central moment* of X or P_X
  - When $k = 2$, it is called the *variance* of X or P_X, and denoted by $\mathrm{Var}(X)$ or σ_X^2
  - The square-root of $\mathrm{Var}(X)$ is called the standard deviation of X, often denoted by σ_X

We have similar definitions for a random vector $X \in \mathcal{R}^d$ or a random matrix $X \in \mathcal{R}^{d_1 \times d_2}$

- Notation:
  - $(a_1, a_2, \ldots, a_d)$ denotes a row vector, and $(a_1, a_2, \ldots, a_d)^\top$ denotes its transport, which is a column vector
  - For a random vector $X = (X_1, \ldots, X_d)^\top$, we use $\mathbb{E}X$ to denote $(\mathbb{E}X_1, \ldots, \mathbb{E}X_d)^\top$
  - Similarly, for a random matrix

$$X = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{1d} \\ X_{21} & X_{22} & \cdots & X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ X_{d1} & X_{d2} & \cdots & X_{dd} \end{pmatrix}$$

    we denote

$$\mathbb{E}X = \begin{pmatrix} \mathbb{E}X_{11} & \mathbb{E}X_{21} & \cdots & \mathbb{E}X_{1d} \\ \mathbb{E}X_{21} & \mathbb{E}X_{22} & \cdots & \mathbb{E}X_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}X_{d1} & \mathbb{E}X_{d2} & \cdots & \mathbb{E}X_{dd} \end{pmatrix}$$

- For a random vector $X \in \mathcal{R}^d$, $\mathrm{Var}(X) = \mathbb{E}\{(X - \mathbb{E}X)(X - \mathbb{E}X)^\top$ is called the *covariance matrix* of X
  - note that, $\mathrm{Var}(X)$ is a matrix when $X = (X_1, \ldots, X_n)^\top$ is a random vector. Its (i,j)th element is $\mathbb{E}\{(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)\}$.
- For two random variables X and Y, the quantity $\mathbb{E}\{(X - \mathbb{E}Y)(X - \mathbb{E}Y)\}$, denoted by $\mathrm{Cov}(X, Y)$, is called the *covariance* of X and Y
  - If $\mathrm{Cov}(X, Y) = 0$, then we say X and Y are *uncorrelated*
  - The standardized covariance, $\mathrm{Cov}(X, Y)/(\sigma_X \sigma_Y)$, is called the *correlation* of X and Y

Some basic properties:

- Symmetry of covariance: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- If X is a random matrix, then $\mathbb{E}(AXB) = A\mathbb{E}(X)B$ for non-random matrices A and B
- For some non-random vector $a \in \mathcal{R}^d$, we have
  - $\mathbb{E}(a^\top X) = a^\top \mathbb{E}X$
  - $\mathrm{Var}(a^\top X) = a^\top \mathrm{Var}(X)a$
- For a random vector X, $\mathrm{Var}(X)$ is a *symmetric positive semi-definite* (SPSD) matrix
  - a matrix M is symmetric if $M = M^\top$
  - a $d \times d$ square matrix M is positive semi-definite (PSD) if for any $v \in \mathcal{R}^d$, $v^\top M v \geq 0$.
  - A simple proof: Let $M = \mathrm{Var}(X)$.
    - symmetry: $M_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i) = M_{ji}$.
    - positive semi-definite:

$$\begin{aligned} v^\top M v &= v^\top \mathbb{E}\{(X - \mathbb{E}X)(X - \mathbb{E}X)^\top\} v \\ &= \mathbb{E}\{v^\top (X - \mathbb{E}X)(X - \mathbb{E}X)^\top v\}. \end{aligned}$$

      Let $Y = v^\top (X - \mathbb{E}X)$. Not ethat Y is a scalar. Then $v^\top M v = \mathbb{E}(YY^\top) = \mathbb{E}(Y^2) \geq 0$.

Chebyshev's and Jensen's inequalities

Theorem 1.26 (Chebyshev). *Let X be a random variable and φ a nonnegative and nondecreasing function on $[0, \infty)$ and $\varphi(-t) = \varphi(t)$ for all real t. Then, for each constant $t \geq 0$,*

$$\varphi(t) P(|X| \geq t) \leq \int_{\{|X| \geq t\}} \varphi(X) \mathrm{d}P \leq \mathbb{E}\varphi(X).$$

- when $\varphi(t) > 0$, we have

$$P(|X| \geq t) \leq \int_{\{|X| \geq t\}} \frac{\varphi(X)}{\varphi(t)} \mathrm{d}P \leq \frac{\mathbb{E}\varphi(X)}{\varphi(t)}.$$

- when $\varphi(t) = t$, we have Markov's inequality

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

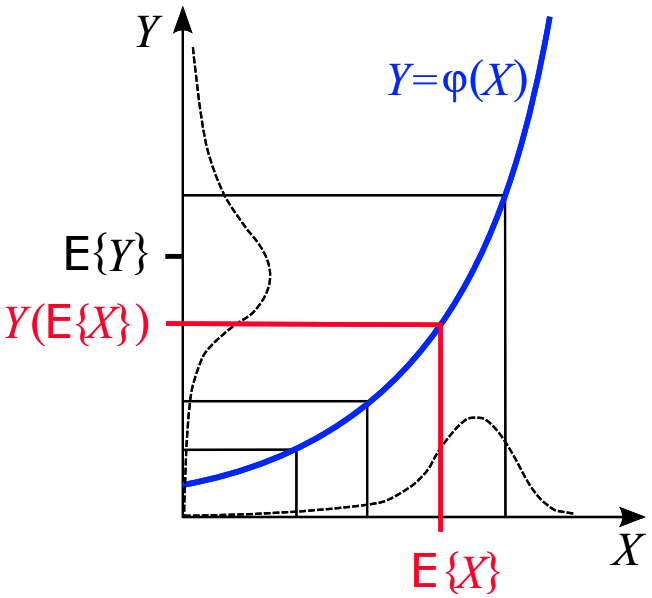
- when $\varphi(t) = t^2$ and X is replaced with $X - \mu$ where $\mu = \mathbb{E}X$, we obtain the classic Chebyshev' inequality:

$$P(|X - \mu| \geq t^2) \leq \frac{\sigma_X^2}{t^2}.$$

Theorem 1.27 (Jensen). *For a random vector and a convex function φ,*

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X).$$

- If φ is differentiable, then the convexity of φ is implied by the positive semi-definiteness of its Hessian (or second derivative if φ is univariate) φ''.
- Intutition illustrated graphically (from Wikipedia)
  - the dashed curve along the X axis is the hypothetical distribution of X
  - the dashed curve along Y axis is the corresponding distribution of $Y = \varphi(X)$
  - the convexity of φ increasingly "stretches" the distribution for increasing values of X
    - the distribution of Y is broader in the interval corresponding to $X > x_0$ and narrower in the region $X < x_0$ for any x_0
    - in particular, this is true for $x_0 = \mathbb{E}X$, so the expectation of $Y = \varphi(X)$ is shifted upwards and hence $\mathbb{E}\varphi(X) \geq \varphi(\mathbb{E}X)$
  - keep this graph in mind and you won't make mistake on the direction of the inequality



- Many well known elementary inequalities can be derived from Jensen's inequality
  - E.g: Let $X \in \{a_1, \ldots, a_n\}$ and $P(X = a_i) = 1/n$. Let $\varphi(x) = x^2$ which is clearly convex. Then
$$\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^2 \leq \frac{1}{n}\sum_{i=1}^{n} a_i^2.$$
- $(\mathbb{E}X)^{-1} < \mathbb{E}(X^{-1})$ for a nonconstant positive random variable X

L^p spaces

To prepare for the discussion of Hölder's inequality, we introduce L^p spaces.

Definition 1.28 (L^p spaces)**.** Fix a measure space $(\Omega, \mathcal{F}, \nu)$. For a real-valued measurable function f on Ω, for $p \in (0, \infty)$, define

$$\|f\|_p = \left(\int_\Omega |f|^p \mathrm{d}\nu \right)^{1/p}.$$

For $p = \infty$, define

$$\|f\|_\infty = \inf\{c \geq 0 : |f(x)| \leq c \text{ for almost every } x\}.$$

The $L^p(\Omega, \mathcal{F}, \nu)$ space is the collection of measurable functions f such that $\|f\|_p < \infty$.

- If $f = g$ ν-a.e., then $\|f - g\|_p = 0$: we can not distinguish functions that are identical almost everywhere in terms of $\|\cdot\|_p$
- In this course, we always identify f with g if $f = g$ a.e.
  - we essentially treat them as the same function
  - the relation $f = g$ a.e. is an equivalence relation:
    - if $f = g$ a.e. and $g = h$ a.e., then $f = h$ a.e.
  - therefore, we can treat L^p space as a space of such equivalance classes
- One can check that L^p spaces (note that $\Omega, \mathcal{F}, \nu$ are often omitted when there are clear from the context) are linear spaces:
  - $f, g \in L^p$, then $af + bg \in L^p$ for real numbers a and b
- $\|\cdot\|_p$ is a norm on L^p (of equivalance classes), and L^p is a Banach space [see Chapter 5 of Rudin (1986) for more information about Banach spaces]
  - a norm $\|\cdot\|$ on L^p must satisfy the following three conditions
    - triangle inequality: $\|f + g\| \leq \|f\| + \|g\|$
    - absolutely scalable: $\|af\| = |a| \cdot \|f\|$ for all real a
    - $\|f\| = 0$ if and only if $f = 0$ a.e.
  - we can check that $\|\cdot\|_p$ satisfies the above conditions.
- See Chapter 3 of Rudin (1986) for more about L^p spaces
- For L^2 spaces, we have additional structure:
  - Define

$$\langle f, g \rangle = \int f g \mathrm{d}\nu.$$

- This is called the inner product or scalar product of L^2 space, and turn L^2 into a Hilbert space [Rudin (1986) for more about Hilbert spaces]
- It is seen that $\|f\|_2^2 = \langle f, f\rangle$.
- We say f is orthogonal to g if $\langle f, g\rangle = 0$.

Hölder's inequality

Theorem 1.29 (Hölder)**.** *Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and $p, q \in [1, \infty]$ satisfying $1/p + 1/q = 1$. Then $\|fg\|_1 \le \|f\|_p \|g\|_q$.*

- if $1/p + 1/q = 1$, then we say p and q are *Hölder conjugate* of each other.
- in a probability space, it is written as

$$\mathbb{E}|XY| \le (\mathbb{E}|X|^p)^{1/p} (\mathbb{E}|Y|^q)^{1/q}$$

- *Cauchy-Schwarz inequality*: when $p = q = 2$, we have $\|fg\|_1 \le \|f\|_2 \|g\|_2$, or more explicitly,

$$\int |fg| \mathrm{d}\nu \le \sqrt{\int |f| \mathrm{d}\nu} \sqrt{\int |g| \mathrm{d}\nu},$$

  or in probability theory,

$$\mathbb{E}|XY| \le \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}.$$

  - this also implies that $|\mathrm{Cov}(XY)| \le \sigma_X \sigma_Y$ and hence the correlation between X and Y are between -1 and 1

- *Minkowski's inequality*: $\|f+g\|_p \le \|f\|_p + \|g\|_p$. Proof: let $q = p/(p-1)$ so that $1/p + 1/q = 1$.

$$\begin{aligned}
\|f+g\|_p^p &= \int |f+g|^p \mathrm{d}\nu \\
&= \int |f+g| \cdot |f+g|^{p-1} \mathrm{d}\nu \\
&\le \int (|f| + |g|) |f+g|^{p-1} \mathrm{d}\nu \\
&= \int |f| \cdot |f+g|^{p-1} \mathrm{d}\nu + \int |g| \cdot |f+g|^{p-1} \mathrm{d}\nu \\
&\le \left(\int |f|^p \mathrm{d}\nu \right)^{1/p} \left(\int |f+g|^{(p-1)\frac{p}{p-1}} \mathrm{d}\nu \right)^{(p-1)/p} \\
&\quad + \left(\int |g|^p \mathrm{d}\nu \right)^{1/p} \left(\int |f+g|^{(p-1)\frac{p}{p-1}} \mathrm{d}\nu \right)^{(p-1)/p} \\
&= (\|f\|_p + \|g\|_p) \|f+g\|_p^{p-1}.
\end{aligned}$$

  Then divide both sides by $\|f+g\|_p^{p-1}$.

- *Lyapunov's inequality*: for a random variable X, for $0 < s \leq t$,

$$(\mathbb{E}|X|^s)^{1/s} \leq (\mathbb{E}|X|^t)^{1/t}.$$

- Proof: for $1 \leq s \leq t$, use Hölder's inequality $\mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$:
  - Let $Y \equiv 1$ and $p = t/s \geq 1$, then $\mathbb{E}|X| \leq (\mathbb{E}|X|^{t/s})^{s/t}$. Replace $|X|$ by $|X|^s$, and we have $\mathbb{E}|X|^s \leq (\mathbb{E}|X|^t)^{s/t}$ and raise the power of both sides to $1/s$ to get the Lyapunov's inequality.
- for the case $0 < s \leq t < 1$, use Jensen's inequality: since $p = t/s \geq 1$, $\varphi(x) = x^p$ is convex on $[0, \infty)$. By Jensen's inequality, with $Y = |X|^s$,

$$\begin{aligned} &\varphi(\mathbb{E}Y) \leq \mathbb{E}\varphi(Y) \\ \Longrightarrow &(\mathbb{E}|X|^s)^{t/s} \leq \mathbb{E}(|X|^{s\cdot t/s}) = \mathbb{E}(|X|^t) \\ \Longrightarrow &(\mathbb{E}|X|^s)^{1/s} = (\mathbb{E}|X|^t)^{1/t}. \end{aligned}$$

1.6 Independence and conditioning

- We want to study relations between two or more random variables $X_1, \ldots, X_n$.
  - e.g. are they correlated?
  - e.g. are they "dependent": does knowing some of them give us information about the others?
- The last one is captured by the concepts "independence" and condititioning in probability theory

L^2 conditional expectation

Let $(\Omega, \mathcal{F}, P)$ be a measure space, X a random variable defined on $(\Omega, \mathcal{F})$, i.e., X is $\mathcal{F}$-$\mathcal{B}$ measurable. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$.

- Here, recall that $\mathcal{B}$ denotes the standard Borel σ-field on the real line $\mathcal{R}$.
- Suppose $X \in L^2(\mathcal{F}) = L^2(\Omega, \mathcal{F}, P)$, i.e., $\mathbb{E}X^2 < \infty$.
- Now we interpret $\mathcal{G}$ as a kind of information available (observable) to us, i.e., we know that events to happen fall into $\mathcal{G}$.
- Given the information $\mathcal{G}$, we want to construct a random variable Y that approximates X
- This Y must be $\mathcal{G}$-measurable, since we can only based on the information we know

- We want the approximation to be optimal in the sense that the "mean squared error" $\mathbb{E}(X - Y)^2$ sis minimized (among all $L^2(\mathcal{G})$ random variables)
- Note that $L^2(\mathcal{G}) \subset L^2(\mathcal{F})$, i.e., a linear subspace of $L^2(\mathcal{F})$ endowed with the scalar product $\langle X, Y \rangle = \mathbb{E}(XY)$.
- The "best" approximation of X from $L^2(\mathcal{G})$ is the orthogonal projection of X on to $L^2(\mathcal{G})$
  - recall that L^2 spaces are Hilbert spaces and thus orthogonal projection is defined.

Definition 1.30 (Conditional expection in L^2 sense)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. For any real random variable $X \in L^2(\Omega, \mathcal{F}, P)$, the conditional expectation of X given $\mathcal{G}$, denoted by $\mathbb{E}(X \mid \mathcal{G})$, is defined as the orthogonal projection of X onto the closed subspace $L^2(\Omega, \mathcal{G}, P)$.

- Orthogonal projection means: $\langle X - \mathbb{E}(X \mid \mathcal{G}), Z \rangle = 0$ for all $Z \in L^2(\mathcal{G})$
- Covariance matching: $\mathbb{E}(X \mid \mathcal{G})$ is the unique random variable $Y \in L^2(\mathcal{G})$ such that for every $Z \in L^2(\mathcal{G})$,
$$\mathbb{E}(XZ) = \mathbb{E}(YZ)$$
  - this is a re-statement of orthogonal projection: $\langle X - \mathbb{E}(X \mid \mathcal{G}), Z \rangle = 0$ implies $\langle X, Z \rangle = \langle \mathbb{E}(X \mid \mathcal{G}), Z \rangle$.

L^1 conditional expectation

- The previous definition of conditional expectation requires square-integrability
- However, we also want conditional expectation for integrable random variables which might not be square-integrable.
- In short, we want conditional expectation in L^1 sense, but L^1 is not a Hilbert space (and no orthogonality)
- We use the covariance matching as a basis for definition of conditional expectation for L^1 random variables

Definition 1.31 (Conditional expectation)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. The *conditional expectation* of a random variable $X \in L^1(\mathcal{F})$, denoted by $\mathbb{E}(X \mid \mathcal{G})$, is defined to be the unique random variable $Y \in L^1(\mathcal{G})$ such that, for every bounded $\mathcal{G}$-measurable random variable Z,
$$\mathbb{E}(XZ) = \mathbb{E}(YZ).$$

- Such random variable Y exists and is unique (a.s.)
- By definition, $\mathbb{E}(X \mid \mathcal{G})$ is measurable from $(\Omega, \mathcal{G})$ to $(\mathcal{R}, \mathcal{B})$

- If $Z = I_A$ for $A \in \mathcal{G}$, then $\mathbb{E}(XZ) = \mathbb{E}(YZ)$ becomes $\int_A \mathbb{E}(X \mid \mathcal{G})\mathrm{d}P = \int_A X\mathrm{d}P$.
- The last properties can also be used as a definition of conditional expectation

Definition 1.32 (Conditional expectation)**.** Let X be an integrable random variable on a measure space $(\Omega, \mathcal{F}, P)$. The *conditional expectation* of X given a sub-σ-field $\mathcal{G}$ of $\mathcal{F}$,denoted by $\mathbb{E}(X \mid \mathcal{G})$ is the a.s.-unique random variable satisfying the following two conditions:

1. $\mathbb{E}(X \mid \mathcal{G})$ is measurable from $(\Omega, \mathcal{G})$ to $(\mathcal{R}, \mathcal{B})$;
2. $\int_A \mathbb{E}(X \mid \mathcal{G})\mathrm{d}P = \int_A X\mathrm{d}P$ for any $A \in \mathcal{G}$.

- Note: $\mathbb{E}(X \mid \mathcal{G})$ is a $\mathcal{G} - \mathcal{B}$ measurable function, and thus a random variable!
- The conditional expectation of X given Y is defined to be $\mathbb{E}(X \mid Y) = \mathbb{E}\{X \mid \sigma(Y)\}$, where
  - $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\}$ where $\mathcal{B}$ is the Borel σ-field of $\mathcal{R}$.
  - We call $\sigma(X)$ the σ-field generated by X. It is a sub σ-field of $\mathcal{E}$, i.e. $\sigma(X) \subset \mathcal{E}$.
- $\mathbb{E}(X \mid Y)$ is a function of Y.
- Conditional probability: $P(A \mid \mathcal{G}) = \mathbb{E}(I_A \mid \mathcal{G})$
- When X is a L^2 random variable, then all these three definitions coincide.

Properties of conditional expectation

- linearity: $\mathbb{E}(aX + bY \mid \mathcal{G}) = a\mathbb{E}(X \mid \mathcal{G}) + b\mathbb{E}(X \mid \mathcal{G})$ a.s.
- If $X = c$ a.s. for a constant c, then $\mathbb{E}(X \mid \mathcal{G}) = c$ a.s.
- monotonicity: if $X \leq Y$, then $\mathbb{E}(X \mid \mathcal{G}) \leq \mathbb{E}(Y \mid \mathcal{G})$ a.s.
- if $\mathcal{G} = \{\emptyset, \Omega\}$ (a trivial σ-field), then $\mathbb{E}(X \mid \mathcal{G}) = \mathbb{E}(X)$
- tower property: if $\mathcal{H} \subset \mathcal{G}$ is a σ-field, (so that $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$), then

$$\mathbb{E}(X \mid \mathcal{H}) = \mathbb{E}\{\mathbb{E}(X \mid \mathcal{G}) \mid \mathcal{H}\}.$$

  - if $\mathcal{H} = \{\emptyset, \Omega\}$, then $\mathbb{E}(X) = \mathbb{E}\{\mathbb{E}(X \mid \mathcal{G})\}$.
- if $\sigma(Y) \subset \mathcal{G}$ and $\mathbb{E}|XY| < \infty$, then $\mathbb{E}(XY \mid \mathcal{G}) = Y\mathbb{E}(X \mid \mathcal{G})$
  - since $\sigma(Y) \subset \mathcal{G}$, information about Y is contained in $\mathcal{G}$, and thus, Y is kind of "known" given the information $\mathcal{G}$.
- if $\mathbb{E}X^2 < \infty$, then $\{\mathbb{E}(X \mid \mathcal{G})\}^2 \leq \mathbb{E}(X^2 \mid \mathcal{G})$ a.s.

Independence

Definition 1.33. Let $(\Omega, \mathcal{E}, P)$ be a probability space.

- (Independent events) The events in a subset $\mathcal{C} \subset \mathcal{E}$ are said to be *independent* iff for any positive n and distinct events $A_1, \ldots, A_n \in \mathcal{C}$,

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n).$$

- (Independent collections) Collections $\mathcal{C}_i \subset \mathcal{E}$, $i \in \mathcal{I}$ (the index set $\mathcal{I}$ could be uncountable) are independent if events in a collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent.
- (Independent random variables): random variables $X_1, \ldots, X_n$ are said to be independent iff $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.
- If $X \perp Y$ (that denotes X and Y are independent), then $\mathbb{E}(X \mid Y) = \mathbb{E}X$ and $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$

1.7 Convergence modes

- In statistics, we often need to assess the quality of an estimator for some unknown quantity
  - e.g. how good is $\overline{X}$ as an estimator for the mean $\mu = \mathbb{E}X$, where $\overline{X}$ is the sample mean of a sample $X_1, \ldots, X_n$?
- There are many way to quantify the estimation quality, one of them is asymptotic convergence rate
  - intuitively, for a good estimator, it becomes closer to the true quantity if we collect more and more data
  - e.g., $\overline{X}$ gets closer to μ if n is large
  - in math language, $\overline{X}$ converges to μ "in some sense"
  - how to define "convergence" properly?
- There are at least four popular definitions of "convergence" in statistics
  - almost sure convergence (or convergence with probability 1)
  - convergence in probablity
  - convergence in L^p
  - convergence in distribution (also called weak convergence)

Almost sure convergence

Definition 1.34. We say a sequence of random elements $X_1, X_2, \ldots$ converges almost surely to a random element X, denoted by $X_n \overset{a.s.}{\to} X$ if

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

- Notation: $P\left(\lim_{n\to\infty} X_n = X\right)$ is a shorthand of the following

$$P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right)$$

- Note that is a type of pointwise convergence, but allow an exceptional set of probability zero
- Note that we assume a common probability space $(\Omega, \mathcal{F}, P)$ for X, $X_1, \ldots$
- How to show almost sure convergence in practice?
  - one way is to do it via Borel-Cantelli lemma
- Infinitely often:
  - Let $\{A_n\}_{n=1}^{\infty}$ be an infinite sequence of events
  - For an outcome $\omega \in \Omega$, we say A holds true or A happens if $x \in A$
  - For an outcome $\omega \in \Omega$, we say the events in the sequence $\{A_n\}_{n=1}^{\infty}$ happen "infinitely often" if A_i happens for an infinite number of indices i.
  - $\{A_i \ i.o.\} = \{\omega \in \Omega : \omega \in A_i \text{ for an infinite number of indices } i\}$ is the collection of outcomes that make the events in the sequence $\{A_n\}_{n=1}^{\infty}$ happen infinitely often.
  - If $\{A_i \ i.o.\}$ happens, then infinitely many of $\{A_n\}_{n=1}^{\infty}$ happen
  - mathematically,

$$\{A_i \quad i.o.\} = \bigcap_{n\geq 1} \bigcup_{j\geq n} A_j \equiv \limsup_{n\to\infty} A_n$$

  - this also shows that $\{A_i \quad i.o.\}$ is measurable

Lemma 1.35 (First Borel-Cantelli)**.** *For a sequence of events* $\{A_n\}_{n=1}^{\infty}$*, if* $\sum_{n=1}^{\infty} P(A_n) < \infty$*, then* $P(A_n \quad i.o.) = 0$.

- Intuition: because $\sum_{n=1}^{\infty} P(A_n) < \infty$, $P(A_n)$ must be very small for large n, and we cannot find a sufficiently number of ω that make infinitely many A_n happen
- In fact, $\sum_{j\geq n} P(A_j) \to 0$ as $n \to \infty$. So $P(\bigcup_{j\geq n} A_j) \leq \sum_{j\geq n} P(A_j) \to 0$ and $\bigcup_{j\geq n} A_j$ becomes too small for sufficiently large n.

Lemma 1.36 (Second Borel-Cantelli)**.** *For a sequence of pairwisely* independent *events* $\{A_n\}_{n=1}^{\infty}$*, if* $\sum_{n=1}^{\infty} P(A_n) = \infty$*, then* $P(A_n \quad i.o.) = 1$.

Theorem 1.37. *Let $X, X_1, X_2, \ldots$ be a sequence of random variables. For a constant $\epsilon > 0$, define the sequence of events $\{A_n(\epsilon)\}_{n=1}^{\infty}$ to be $A_n(\epsilon) = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}$. If $\sum_{n=1}^{\infty} P\{A_n(\epsilon)\} < \infty$ for all $\epsilon > 0$, then $X_n \overset{a.s.}{\to} X$.*

- According to the first Borel-Cantelli lemma, if $A(\epsilon)$ denotes the collection of ω that makes $\{A_n(\epsilon)\}_{n=1}^{\infty}$ happen finite times, then $P\{A(\epsilon)\} = 1$.
- This implies that, for all $\omega \in A(\epsilon)$, $|X_n(\omega) - X(\omega)| < \epsilon$ for sufficiently large n
- This holds for all $\epsilon > 0$, so X_n converges to X on a set of probablity 1

Convergence in L^p

- In statistics, mean squared error (MSE) is a popular measure for esitmation quality
  - e.g. $\mathbb{E}(\overline{X} - \mu)^2$ becomes small if n is large
  - This is indeed the convergence in L^2 for random variables (treat μ as a degenerate random variable)
- more generally, we can consider convergence in L^p for $p > 0$

Definition 1.38. A sequene $\{X_n\}_{n=1}^{\infty}$ of random variables converges to a random variable X in the L^p sense (or L^p-norm when $p \geq 1$) for some $p > 0$ if $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|X_n|^p < \infty$, and

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

- denoted by $X_n \overset{L^p}{\to} X$
- For L^2, it is also called convergence in mean square.
- By Lyapunov's inequality, convergence in L^p sense implies convergence in L^q sense if $q \leq p$.
- This is not a pointwise convergence.

Convergence in probability

- We might say that, an estimator behaves well at ω if it converges to its target, and say that it behaves badly if it does not converge to the target.
- Likely almost sure convergence, we might allow the estimator to behave badly at some outcomes ω
- However, the collection of such outcomes shall be "small" in some sense
  - in almost sure convergence, such set has zero probablity

  - we wan to relax it a little bit, for example, the probability of such set shall decrease to zero as we get more and more samples

Definition 1.39. A sequene $\{X_n\}_{n=1}^{\infty}$ of random variables converges to a random variable X in probability if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

- denoted by $X_n \xrightarrow{P} X$

Convergence in distribution

- Under certain conditions, CLT implies that $\sqrt{n}\overline{X}$ converges to $N(\mu, \sigma^2)$ where $\sigma^2 = \mathbb{E}X_n$
- This is indeed convergence in distribution
- There are several definitions that agree wit each other

Definition 1.40. A sequene $\{X_n\}_{n=1}^{\infty}$ of random variables converges to a random variable X in distribution (or in law or weakly), if

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

for every $x \in \mathcal{R}$ at which F is continuous, where F_n and F are CDF of X_n and X, respectively.

- denoted by $X_n \xrightarrow{D} X$ or $F_n \Rightarrow F$
- the requirement that only the continuity points of F should be considered is to make the definition agree with other definitions of weak convergence
- this is the elementary version learned perhaps in undergraduate courses
- observation: it is about the convergence of CDFs, not really about random variables (they are dummy variables, indeed)
- CDFs are indeed probability measures
- Convergence of probability measures: A sequence of probability measures ν_n converges weakly to ν if $\int f \mathrm{d}\nu_n \to \int f \mathrm{d}\nu$ for every bounded and continuous real function f
  - if you know Riesz representation theorem for Borel measures, then this defintion justifies the term "weak convergence".
- It can be shown that the above weak convergence of probability measures is equivalent to the convergence in distribution for $\Omega = \mathcal{R}$.
- More about weak convergence of measures can be bounded in:

  - Chapter 5 of Billingsley (2012)
  - Chapter 1 & 2 of Billingsley (1999) (quite advanced topics on weak convergence)
  - Billingsley (1971) (quite advanced topics on weak convergence)

- convergence in distribution can be characterized by characteristic functions.
- characteristic function φ of a random vector X is $\phi_X(t) = \mathbb{E}e^{it^\top X}$, where $i = \sqrt{-1}$ and $e^{it^\top X} = \cos(t^\top X) + \sqrt{-1}\sin(t^\top X)$.
- a similar but different concept is moment generating function (MGF): $\psi_X(t) = \mathbb{E}e^{t^\top X}$.
- these functions determine distributions uniquely in the sense that
  - if $\phi_X(t) = \phi_Y(t)$ for all t, then $P_X = P_Y$
  - if $\psi_X(t) = \psi_Y(t) < \infty$ for all t in a neighborhood of 0, then $P_X = P_Y$.

Theorem 1.41 (Lévy continuity). *$\{X_n\}$ converges in distribution to X iff the corresponding characteristic functions $\{\phi_n\}$ converges pointwise to ϕ_X.*

- If the CDF F_n of X_n, then there is another way to check convergence in distribution

Theorem 1.42 (Scheffé). *Let $\{f_n\}$ be a sequence of PDF on $\mathcal{R}^k$ with respect to a measure ν. Suppose that $\lim_{n\to\infty} f_n(x) = f(x)$ ν-a.e. and f is a PDF with respect to ν. Then*

$$\lim_{n\to\infty} \int |f_n(x) - f(x)| \mathrm{d}\nu = 0.$$

- If f_n is the PDF/PMF of X_n and f is the PDF/PMF of X, and if $f_n(x) \to f(x)$ a.e., then $X_n \xrightarrow{D} X$.
  - proof: for any Borel $A \subset \mathcal{R}$,

$$\left| \int_A f_n \mathrm{d}\nu - \int_A f \mathrm{d}\nu \right| \le \int |f_n - f| \mathrm{d}\nu \to 0.$$

    This is in particular true for $A = (-\infty, x]$, and the above implies that $|F_n(x) - F(x)| \to 0$.
  - ν could be the Lebesgue measure or counting measure
  - e.g. $X_n \sim \text{binom}(n, p)$ and $np \to \lambda$, then $X_n \xrightarrow{D} X \sim \text{Poisson}(\lambda)$

Properties and relations

Theorem 1.43 (Continuous mapping). *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random k-vectors and X is a random vector in the same probability space. Let $g : \mathcal{R}^k \to \mathcal{R}$ be continuous.*

- *If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.*

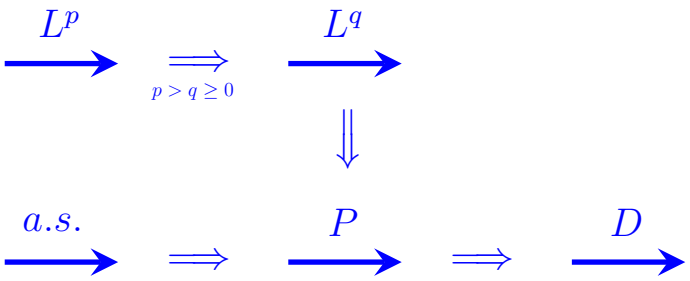
- *If* $X_n \xrightarrow{P} X$, *then* $g(X_n) \xrightarrow{P} g(X)$.
- *If* $X_n \xrightarrow{D} X$, *then* $g(X_n) \xrightarrow{D} g(X)$.

- Uniqueness of the limit
  - If $X_n \xrightarrow{*} X$ and $X_n \xrightarrow{*} Y$, then $X = Y$ a.s., where $*$ could be $a.s.$, P or L^p
  - If $F_n \Rightarrow F$ and $F_n \Rightarrow P$, and $F = P$

Remark. For those who know topology, it can be shown that there is a topology on the space of all probability distributions on a common metric space, and the convergence in distribution is the same as the convergence in such topology. Similarly, there is a topology on the space of all random variables residing on the sample probability space, and the convergence in probability is the same as the convergence in such topoloyg. However, there is no topology for almost sure convergence.

- Concatenation:
  - If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c$, then $(X_n, Y_n) \xrightarrow{D} (X, c)$ for a constant c
  - If $X_n \xrightarrow{*} X$ and $Y_n \xrightarrow{*} Y$, then $(X_n, Y_n) \xrightarrow{*} (X, Y)$ where $*$ is either P or $a.s.$
- Linearity
  - If $X_n \xrightarrow{*} X$ and $Y_n \xrightarrow{*} Y$, then $aX_n + bY_n \xrightarrow{*} aX + bY$, where $*$ could be $a.s.$, P or L^p, and a and b are real numbers
  - When $*$ is P or $a.s.$, it is the consequence of continuous mapping theorem and concatenation property
  - Note that the above statements are NOT true for convergence in distribution
- Cramér-Wold device $X_n \xrightarrow{D} X$ iff $c^\top X_n \xrightarrow{D} c^\top X$ for every $c \in \mathcal{R}^k$

We have the following relations between different modes of convergence



- If $X_n \xrightarrow{D} c$ for a constant c, then $X_n \xrightarrow{P} c$. In general, convergence in distribution does not imply convergence in probability
- Slutsky's theorem: if $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c$ for a constant c, then
  - $X_n + Y_n \xrightarrow{D} X + c$
  - $X_n Y_n \xrightarrow{D} cX$
  - $X_n / Y_n \xrightarrow{D} X/c$ if $c \neq 0$
- Slutsky's theorem is a consequence of continuous mapping theorem and concatenation property

Stochastic order

- In calculus, for two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$
  - $a_n = O(b_n)$ iff $|a_n| \leq c|b_n|$ for a constant c and all n
  - $a_n = o(b_n)$ iff $a_n/b_n \to 0$ as $n \to \infty$
- For two sequences of random variables, $\{X_n\}$ and $\{Y_n\}$, we have similar notations
  - $X_n = O_{a.s.}(Y_n)$ iff $P\{|X_n| = O(|Y_n|)\} = 1$
    - in other words, there is a subset $A \subset \Omega$ such that $P(A) = 1$, and for each $\omega \in A$, there exists a constant c (depending on ω), and for all n, $|X_n(\omega)| \leq c|Y_n(\omega)|$
  - $X_n = o_{a.s.}(Y_n)$ iff $X_n/Y_n \xrightarrow{a.s.} 0$
  - $X_n = O_P(Y_n)$ iff, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that the events $A_n(\epsilon) = \{\omega \in \Omega : |X_n(\omega)| \geq C_\epsilon |Y_n(\omega)|\}$ satisfies $\limsup_n P\{A_n(\epsilon)\} < \epsilon$
    - in the textbook, $\limsup_n$ is replaced with $\sup_n$. I believe it is a typo: stochast order is an asymptotic relation
      - Let $X_n = 1$ and $X_n = -1$ with probability $1/2$, and $Y_1 = 0$ and $Y_n = 1$, then we still think of $X_n = O_P(Y_n)$, but $\sup_n P\{A_n(\epsilon)\} = 1$ for any $\epsilon > 0$.
    - in most cases, $Y_n = a_n$ for a sequence of real numbers, e.g., $\overline{X} - \mu = O_P(1/\sqrt{n})$
    - If $X_n = O_P(1)$, we say $\{X_n\}$ is bounded in probability
  - $X_n = o_P(Y_n)$ iff $X_n/Y_n \xrightarrow{P} 0$
- Some properties
  - if $X_n = O_P(Y_n)$ and $Y_n = O_P(Z_n)$, then $X_n = O_P(Z_n)$
  - if $X_n = O_P(Z_n)$, then $X_n Y_n = O_P(Y_n Z_n)$
  - if $X_n = O_P(Z_n)$ and $Y_n = O_P(Z_n)$, then $X_n + Y_n = O_P(Z_n)$
  - same conclusion for $O_{a.s.}$

- For weak convergence:
  - If $X_n \xrightarrow{D} X$ for a random variable, then $X_n = O_P(1)$
  - If $\mathbb{E}|X_n| = O(a_n)$, then $X_n = O_P(a_n)$, according to Chebyshev's inequality
  - If $X_n \xrightarrow{a.s.} X$, then $\sup_n |X_n| = O_P(1)$

1.8 Law of large numbers and CLT

- In statistics, we often need to quantify the stochastic order of a random variable that is the sum/average of a sequence of other random variables, or study its stochastic limit if we push n to ∞
  - e.g. $\hat{\mu}_n = \overline{X} = n^{-1}\sum_{i=1}^n X_i$
- This often involves law of large numbers

WLLN

- weak law of large numbers concerns the limiting behavior in probability

Theorem 1.44 (WLLN). *Let $\{X_n\}$ be IID random variables. If $nP(|X_1| > n) \to 0$, then*

$$\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}(X_1 I_{\{|X_1| \le n\}}) \xrightarrow{P} 0.$$

- a more familiar condition is $\mathbb{E}|X_1| < \infty$, in this case
  - $nP(|X_1| > n) \le \int_n^\infty |x| \mathrm{d}F_{|X_1|}(x) \le \mathbb{E}\{I_{[n,\infty)}(|X_1|)|X_1|\} \to 0$ (by DCT)
  - $\mathbb{E}(X_1 I_{\{|X_i| \le n\}}) \to \mathbb{E}X_1$ (again by DCT)
  - so, $n^{-1}\sum_{i=1}^n X_i \xrightarrow{P} \mathbb{E}X_1$
- conditions in terms of finiteness of certain order of moments are quite common in statistics
- for independent but not identically distributed random variables, we have

Theorem 1.45 (WLLN). *If there is a constant $p \in [1, 2]$ such that $\lim_{n\to\infty} n^{-p}\sum_{i=1}^n \mathbb{E}|X_i|^p = 0$, then*

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) \xrightarrow{P} 0.$$

SLLN

- strong law of large numbers concerns the limiting behavior in "almost sure" sense

Theorem 1.46 (SLLN)**.** *Let $\{X_i\}$ be IID random variables. If $\mathbb{E}|X_1| < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mathbb{E}X_1$$

and

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - \mathbb{E}X_i) \overset{a.s.}{\to} 0$$

for any bounded sequence of real numbers $\{c_i\}$.

- under the IID assumption and the condition $\mathbb{E}|X_1| < \infty$, we can indeed show that the average converges almost surely, not just in probability
- for independent but not identically distributed case, we have

Theorem 1.47 (SLLN)**.** *Let $\{X_i\}$ be independent random variables with finite expections, i.e., $\mathbb{E}X_i < \infty$ for all i. If there is a constant $p \in [1, 2]$ such that $\sum_{i=1}^{\infty} i^{-p}\mathbb{E}|X_i|^p < \infty$, then*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \overset{a.s.}{\to} 0.$$

- note that this condition is stronger than the condition for WLLN (Kronecker's Lemma)

Example 1.48. Let $S_n = \sum_{i=1}^{n} X_i$, where $\{X_i\}$ are independent and $P(X_i = \pm i^\theta) = 1/2$ and $\theta > 0$ is a constant. We claim that $S_n/n \overset{a.s.}{\to} 0$ when $\theta < 1/2$. This because, when $\theta < 1/2$,

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}X_i^2}{i^2} = \sum_{i=1}^{\infty} \frac{i^{2\theta}}{i^2} < \infty.$$

Then the claim follows from SLLN.

CLT

- The limits in WLLN and SLLN are constants
- Sometimes, we also want asymptotic distributions of the (normalized) sum/average
  - e.g. asymptotic hypothesis test, confidence intervals

Theorem 1.49 (Lindeberg's CLT). *Let $\{X_{nj}, j = 1, \ldots, k_n\}$ be row independent array of random variables with $k_n \to \infty$ as $n \to \infty$ and*

$$0 < \sigma_n^2 = \operatorname{Var}\left(\sum_{j=1}^{k_n} X_{nj}\right) < \infty, \qquad n = 1, 2, \ldots.$$

If

$$\frac{1}{\sigma_n^2} \sum_{j=1}^{k_n} \mathbb{E}\left\{(X_{nj} - \mathbb{E}_{nj})^2 I_{\{|X_{nj} - \mathbb{E}X_{nj}| > \epsilon\sigma_n}\right\} \to 0 \tag{1.3}$$

for any $\epsilon > 0$, then

$$\frac{1}{\sigma_n} \sum_{j=1}^{k_n} (X_{nj} - \mathbb{E}X_{nj}) \xrightarrow{D} N(0, 1).$$

- (1.3) controls the tails of X_{nj}
- The condition (1.3) is implied by the Lyapunov condition:

$$\frac{1}{\sigma_n^{2+\delta}} \sum_{j=1}^{k_n} \mathbb{E}|X_{nj} - \mathbb{E}X_{nj}|^{2+\delta} \to 0 \text{ for some } \delta > 0.$$

- It is also implied by the condition of uniform boundedness: if $|X_{nj}| \leq M$ for all n and j and $\sigma_n^2 = \sum_{j=1}^{k_n} \operatorname{Var}(X_{nj}) \to \infty$.
- IID case: $k_n = n$ and $X_{nj} = X_j$ and $\{X_j\}$ are IID with $\operatorname{Var}(X_j) > 0$. In this case, Lindeberg's condition holds.

Theorem 1.50 (Multivariate CLT). *For IID random k-vectors $\{X_i\}$ with $\Sigma = \operatorname{Var}(X_1)$, we have*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}X_1) \xrightarrow{D} N(0, \Sigma).$$

- This is a consequence of Lindeberg's CLT and Cramér-Wold device

1.9 δ-Method

- Motivating example
  - Suppose $X_1, \ldots, X_n \sim X \sim \text{Exp}(\lambda)$ IID with PDF

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \in [0, \infty)$$

  - the parameter $\lambda > 0$ is called the rate
  - one can check that $\mu = \mathbb{E}X = 1/\lambda$. Or $\lambda = \mu^{-1}$.

  - by CLT, we know $\hat{\mu} = \overline{X} \approx N(\mu, \sigma^2/n)$ where $\sigma^2 = \mathrm{Var}(X_1)$.
  - what's the approximate distribution of $\hat{\lambda} = \hat{\mu}^{-1}$?

- More generally, if we know the approximate distribution of $\hat{\theta}$ (often by CLT), what is the approximate distribution of $g(\hat{\theta})$ for a "well behaved" function g?

- Suppose $a_n(\hat{\theta}_n - \theta) \xrightarrow{D} Z$
  - when $\hat{\theta}_n \approx \theta$, and g is differentiable, then by Taylor expansion

$$\frac{g(\hat{\theta}_n) - g(\theta)}{\hat{\theta}_n - \theta} \approx g'(\theta)$$

or

$$\frac{g(\hat{\theta}_n) - g(\theta)}{g'(\theta)} \approx \hat{\theta}_n - \theta$$

and further

$$a_n \frac{g(\hat{\theta}_n) - g(\theta)}{g'(\theta)} \approx a_n(\hat{\theta}_n - \theta) \xrightarrow{D} Z.$$

Theorem 1.51 (δ-method, univariate)**.** *Let $Y_1, \ldots$ and Z be random variable such that $a_n(Y_n - c) \xrightarrow{D} Z$ for a constant c and a sequence of positive numbers $\{a_n\}$ satisfying $\lim_{n\to\infty} a_n = \infty$. For a function g that is differentiable at c, we have*

$$a_n\{g(Y_n) - g(c)\} \xrightarrow{D} g'(c)Z. \tag{1.4}$$

More generally, if g has continuous derivatives of order $m > 1$ in a neighborhood of c, $g^{(j)}(c) = 0$ for $1 \le j \le m-1$, and $g^{(m)}(c) \neq 0$. Then

$$a_n^m\{g(Y_n) - g(c)\} \xrightarrow{D} \frac{1}{m!} g^{(m)}(c) Z^m.$$

- e.g. $X_1, \ldots, X_n$ IID with $\mathrm{Var}(X_1) = 1$, $Y_n = \overline{X} = n^{-1}\sum_{i=1}^n X_i$, $c = \mathbb{E}X_1$, $a_n = \sqrt{n}$, and $Z \sim N(0,1)$
  - if $g(x) = x^{-1}$ and $c \neq 0$, then $\sqrt{n}(Y_n^{-1} - c^{-1}) \xrightarrow{D} N(0, 1/c^4)$, since $g'(c) = -c^{-2}$.
  - if $g(x) = x^2$, then $\sqrt{n}(Y_n^2 - c^2) \xrightarrow{D} N(0, 4c^2)$ since $g'(c) = 2c$.
    - if $c = 0$, then $g'(c) = 0$ but $g''(c) = 2 \neq 0$, so we have $(\sqrt{n})^2(Y_n^2 - 0) \xrightarrow{D} Z^2 \sim \chi_1^2$

- go back to our example on exponential distribution
  - $c = \mu$, $g(\mu) = \mu^{-1} = \lambda$, $g'(\mu) = -\mu^{-2} = -\lambda^2$
  - $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{D} Z \sim N(0, \sigma^2) = N(0, \mu^2) = N(0, \lambda^{-2})$
  - $\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{D} -\lambda^2 Z \sim N(0, \lambda^2)$.

Proof of (1.4). By Taylor expansion

$$g(Y_n) - g(c) = g'(c)(Y_n - c) + Q_n,$$

or

$$a_n\{g(Y_n) - g(c)\} = a_n g'(c)(Y_n - c) + R_n$$

where R_n is the residual and is random. By assumption, $a_n(Y_n - c) \xrightarrow{D} Z$. Further, by Cramér-Wold device, $a_n g'(c)(Y_n - c) \xrightarrow{D} g'(c)Z$. If we can show that $R_n = o_P(1)$, then by Slutsky's lemma, we have $a_n\{g(Y_n) - g(c)\} \xrightarrow{D} g'(c)Z$.

- how to show $R_n = o_P(1)$, or equivalently, $R_n \xrightarrow{P} 0$? By defintion, for any $\eta > 0$, $\lim_n P(|R_n| > \eta) = 0$.

Note that $R_n = a_n\{g(Y_n) - g(c)\} - a_n g'(c)(Y_n - c)$. The differentiability of g at c implies that, for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$|g(x) - g(c) - g'(c)(x - c)| \le \epsilon|x - c|$$

whenever $|x - c| < \delta_\epsilon$. Then for a fixed $\eta > 0$, we have

$$P(|R_n| > \eta) \le P(|Y_n - c| \ge \delta_\epsilon) + P(a_n|Y_n - c| \ge \eta/\epsilon).$$

- $P(|R_n| > \eta) = P(|R_n| > \eta, |Y_n - c| \ge \delta_\epsilon) + P(|R_n| > \eta, |Y_n - c| < \delta_\epsilon)$
- for $\omega \in A_\epsilon \equiv \{\omega \in \Omega : |R_n(\omega)| > \eta, |Y_n(\omega) - c| < \delta_\epsilon\}$, we have

  $$|g(Y_n(\omega)) - g(c) - g'(c)(Y_n(\omega) - c)| \le \epsilon|Y_n(\omega) - c|$$

  and

  $$|a_n\{g(Y_n(\omega)) - g(c)\} - a_n g'(c)(Y_n(\omega) - c)| > \eta.$$

  This implies that, if $\omega \in A_\epsilon$, then

  $$\eta \le \epsilon a_n|Y_n(\omega) - c|, \qquad \text{or equivalently,} \quad a_n|Y_n(\omega) - c| \ge \eta/\epsilon.$$

- Let $B_\epsilon = \{\omega \in \Omega : a_n|Y_n(\omega) - c| \ge \eta/\epsilon\}$
- We just show that if $\omega \in A_\epsilon$, then $\omega \in B_\epsilon$. What does this mean? $A_\epsilon \subset B_\epsilon$! and therefore $P(A_\epsilon) \le P(B_\epsilon)$
- So

$$\begin{aligned} P(|R_n| > \eta) &= P(|Y_n - c| \ge \delta_\epsilon) + P(A_\epsilon) \\ &\le P(|Y_n - c| \ge \delta_\epsilon) + P(B_\epsilon) \\ &= P(|Y_n - c| \ge \delta_\epsilon) + P(a_n|Y_n - c| \ge \eta/\epsilon) \end{aligned}$$

- now we only need to show that
  - $P(|Y_n - c| \geq \delta_\epsilon) \to 0$: Intuitively, this must be true: by assumption, $a_n(Y_n - c) \xrightarrow{D} Z$ implies that $a_n(Y_n - c) = O_P(1)$. Then (multiply both sides by a_n^{-1}) $Y_n - c = O_P(a_n^{-1}) = o_P(1)$.
  - $P(a_n|Y_n - c| \geq \eta/\epsilon) \to 0$: This is also intuitively true: $a_n(Y_n - c) \approx O_P(1)$ and ϵ could be arbitrarily small.

Formal argument: By continuous mapping theorem, $a_n|Y_n - c| \xrightarrow{D} |Z|$. Then

$$\lim_n P(a_n|Y_n - c| \geq \eta/\epsilon) = 1 - F_{|Z|}(\eta/\epsilon) \to 0$$

since ϵ is arbitrary. □

Theorem 1.52 (δ-method, multivariate)**.** *Let $Y_1, \ldots$ and Z be random k-vectors such that $a_n(Y_n - c) \xrightarrow{D} Z$ for a constant k-vector c and a sequence of positive numbers $\{a_n\}$ satisfying $\lim_{n\to\infty} a_n = \infty$. For a function g that is differentiable at c, we have*

$$a_n\{g(Y_n) - g(c)\} \xrightarrow{D} \{\nabla g(c)\}^\top Z.$$

Example 1.53. Let $X_1, \ldots, X_n$ be IID such that $\mathbb{E}X_1^4 < \infty$. Let $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Denote $\sigma^2 = \mathrm{Var}(X_1)$, $\mu = \mathbb{E}X_1$, and $m_2 = \mathbb{E}X_1^2$. Now we derive the asymptotic distribution of $\sqrt{n}(\hat{\sigma}^2 - \sigma^2)$.

We first note that $\hat{\sigma}^2 = \hat{m}_2 - \hat{\mu}^2$, where $\hat{m}_2 = n^{-1}\sum_{i=1}^n X_i^2$. This motivates us to define $g(y_1, y_2) = y_2 - y_1^2$. We observe that $\nabla g(y_1, y_2) = (-2y_1, 1)^\top \neq 0$. To apply the δ-method, we also need to observe that, by multivariate CLT, for $Y_n = (\overline{X}, \hat{m}_2)^\top$, we have $\sqrt{n}(Y_n - c) \xrightarrow{D} N(0, \Sigma)$, where $c = (\mu, m_2)$ and $\Sigma = \mathrm{Cov}(X_1, X_1^2)$.

By δ-method, we then have

$$\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{D} N\left(0, (-2\mu, 1)\Sigma(-2\mu, 1)^\top\right).$$

- the asymptotic distribution of $\hat{\sigma}^2$ depends on μ! this is because μ is known
- if μ is known, we shall estimate $\hat{\sigma}^2$ by $n^{-1}\sum_{i=1}^n (X_i - \mu)^2$. What is the asymptotic distribution for this estimator?

References

Billingsley, P. (1971). *Weak Convergence of Measures: Applications in Probability.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM.

Billingsley, P. (1999). *Convergence of Probability Measures.* Wiley-Interscience, 2nd edition.

Billingsley, P. (2012). *Probability and Measure.* Wiley, anniversary edition edition.

Munkres, J. (2000). *Topology.* Pearson, 2nd edition.

Rudin, W. (1986). *Real and Complex Analysis.* McGraw-Hill Education, 3rd edition.

Lecture 2: Basic concepts, Exponential families and Sufficient statistics

Lecturer: LIN Zhenhua ST5215 AY2019/2020 Semester I

2.1 Populations, samples, and models

- A typical statistical problem
  - one or a series of random experiments is performed
  - some data are generated and collected from the experiments
  - extract information from the data
  - interpret results and draw conclusions

Example 2.1 (Measurement problems)**.** Suppose we want to measure an unknown quantity θ, e.g., weight of some object.

- n measurements $x_1, \ldots, x_n$ are taken in an experiment of measuring θ.
- data are $(x_1, \ldots, x_n)$
- information to extract: some estimator for θ
- draw conclusion: what is the possible range of θ (confidence interval)?

- In mathematical statistics, we only focus on statistical analysis of data; we assume data are given.
- In order to analyze data (mathematically/statistically), we need a model for the data
  - in physics, one requires a mathematical model to describe what are observed
    - $F = ma$, for example
  - models are (mathematicl) approximation of our reality
    - only approximation
  - good models approximate the reality well
    - Newton's physics is good for low-speed motion
    - For high-speed motion, we needs speical relativity or even general relativity
- In statistics, we use models to approximate the mechanism that generates the observed data
  - "All models are wrong." – George Box. But some are useful.
- In the measurement example: $x_i = \theta + \varepsilon_i$ for $\varepsilon_i \sim N(0, \sigma^2)$ IID

  - in reality: $\varepsilon \sim N(0, \sigma^2)$ might not be true
  - IID might not be true: in the scale of subatomic particle, measurement performed on the particle might change its status – by quantum physics
  - linearity might not be true
  - But this model might provide a good approximation to the measurement experiment well in some (most) cases
    * i.e. measure the weight of a baby

- Let us fix some terminology
  - the data set is viewed as a realization or observation of a random element defined on a probability space $(\Omega, \mathcal{E}, P)$ related to the random experiment (or observational studies).
  - The probability measure P is called the *population*
  - The random element that produces the data is called a *sample* from P
  - The data set is also often called a sample from P
  - The size of the data set is called the *sample size*
  - A population P is known iff $P(A)$ is known for very event $A \in \mathcal{E}$.
- In statistics, P is at least partially unknown. Otherwise, no statistical analysis is required.
- Statisticians are to deduce some properties of P based on the available sample based on some statistical models for the data
- A *statistical model* is a set of assumptions on the population P
  - a statistical model = $\{P$: P satisfies a set of assumptions$\}$
- A statistical model is also a set of probability measures on the space $(\Omega, \mathcal{E})$

Example 2.2 (Measurement problems)**.** Suppose we want to measure an unknown quantity θ, e.g., weight of some object.

- n measurements $x_1, \ldots, x_n$ are taken in an experiment of measuring θ.
- If no measurement error, then $x_1 = \cdots = x_n = \theta$.
- Otherwise, x_i are not the same due to measurement errors
- the data set $(x_1, \ldots, x_n)$, is viewed as an outcome of the experiment
  - the sample space in this case is $\Omega = \mathcal{R}^n$
  - $\mathcal{E} = \mathcal{B}^n$, the Borel σ-field of $\mathcal{R}^n$
  - and P is a probability measure on $\mathcal{R}^n$
  - sample size is n

  - the random element $X = (X_1, \ldots, X_n)$ is a random n-vector defined on $\mathcal{R}^n$, i.e., $X : \mathcal{R}^n \to \mathcal{R}^n$

- a statistical model here is a set of *joint* distribution of $X_1, \ldots, X_n$
  - not just the marginal distributions, since marginal distributions do not specify relations among $X_1, \ldots, X_n$ and thus can not fully specify the probability measure P
  - when $X_1, \ldots, X_n$ are IID, then $P = P_0^n$.
    * still, rigorously speaking, the model is a set of P_0^n, not simply P_0, although in this case, sometimes we say the model is a set of P_0.
    * $X_i = \theta + \varepsilon_i \sim N(\theta, \sigma^2)$, so $P_0 = N(\theta, \sigma^2)$ in this case (with IID assumption)
  - P is partially unknown, since θ (and perhaps σ^2) are unknown. But we know it is a multivariate Guassian distribution (by assumption, of course)
  - Statisticians are to deduce θ (and also σ^2)
  - a statistical model: $\{N(\theta \mathbf{1}_n, \sigma^2 \mathbf{I}_n) : \theta, \sigma^2 \in (0, \infty)\}$ – a set of probability distributions
    * well, we assume the weight is positive
    * we can consider a larger model like $\{N(\theta \mathbf{1}_n, \sigma^2 \mathbf{I}_n) : \theta \in \mathcal{R}, \sigma^2 \in (0, \infty)\}$, but this is not as good as the previous one, since we know the weight is positive

Definition 2.3. A set of probability measures P_θ on $(\Omega, \mathcal{E})$ indexed by a parameter $\theta \in \Theta$ is said to be a *parametric* family if and only if $\Theta \subset \mathcal{R}^d$ for some fixed $d > 0$ and each P_θ is a known probability measure when θ is known. The set Θ is called the *parameter space* and d is called the *dimension*.

- In a statistical model, if the set of probability measures is a parametric familiy, then we say the model is a *parametric model*.
- Otherwise, we say the model is a *nonparametric model*
- A parametric family $\{P_\theta : \theta \in \Theta\}$ is said to be identifiable if and only if $\theta_1 \neq \theta_2$ and $\theta_1, \theta_2 \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$
  - identifiable within the family
- Let $\mathcal{P}$ be a family of populations (they are probability measures) and ν a σ-finite measure on $(\Omega, \mathcal{E})$
  - if $P \ll \nu$ for all $P \in \mathcal{P}$,then we say $\mathcal{P}$ is dominated by ν
  - in this case, $\mathcal{P}$ can be identified by the family of densities $\{\frac{\mathrm{d}P}{\mathrm{d}\nu} : P \in \mathcal{P}\}$.
  - in statistics, the measure ν is often the Lebesgue measure (for continuous random variables) or the counting measure (for discrete random variables)

2.2 Statistics

- Let X be a sample (a random vector) from an unknown population P on a probability space
- A measurable function $T(X)$ of X is called a *statistic* if $T(X)$ is a known value whenever X is known
- Statistical analyses are based on various statistics, for various purposes
- Examples:
  - $T(X) = X$: this is a trivial statistic
  - $T(X) = n^{-1} \sum_{i=1}^{n} X_i$ for $X = (X_1, \ldots, X_n)$
  - Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then $T(X) = \theta X$, for example, is not a statistic
- One can easily show that $\sigma(T(X)) \subset \sigma(X)$, and these two σ-fields are the same iff T is one-to-one
  - the information contained in $T(X)$ is often less than X
  - $T(X)$ compresses the information provided by X (often in a good way if T is well chosen for the problem of interest)
  - Sometimes, $T(X)$ is simplier than X but still contains all information we need: sufficiency and completeness

2.3 Exponential families

- These are important parametric families in statistical applications, like GLM

Definition 2.4. A parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathcal{E})$ is called an *exponential family* iff

$$f_\theta(\omega) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}(\omega) = \exp\left\{[\eta(\theta)]^\top T(\omega) - \xi(\theta)\right\} h(\omega), \qquad \omega \in \Omega,$$

where T is a random p-vector, η is a function from Θ to $\mathcal{R}^p$, h is a nonnegative Borel function on $(\Omega, \mathcal{E})$, and

$$\xi(\theta) = \log\left\{\int_\Omega \exp\{[\eta(\theta)]^\top T(\omega)\} h(\omega) \mathrm{d}\nu(\omega)\right\}.$$

- T and h are functions of ω only
- ξ and η are functions of θ only
- These functions are not identifiable

  - $\tilde{\eta} = D\eta(\theta)$ for a $p \times p$ nonsingular matrix D and $\tilde{T} = D^{-\top}T$ give another representation for the same family
  - another measure that dominates the family also changes the representation

- we can reparametrize the family by $\eta = \eta(\theta)$, so that

$$f_\eta(\omega) = \exp\{\eta^\top T(\omega) - \zeta(\eta)\}h(\omega)$$

  where $\zeta(\eta) = \log\left\{\int_\Omega \exp\{\eta^\top T(\omega)\}h(\omega)\mathrm{d}\nu(\omega)\right\}$.

- This is the *canonical form* for the family – still not unique
- η is called the *natural parameter*
- the new parameter space is $\Xi = \{\eta(\theta) : \theta \in \Theta\} \subset \mathcal{R}^p$: *natural parameter space*
- An exponential family in its canonical form is called a *natural exponential family*
- *Full rank*: if Ξ contains an open set

Example 2.5 (Binomial distribution)**.** The Binomial distribution $\mathrm{Binom}(\theta, n)$ is an exponential family.

- $\Omega = \{0, 1, \ldots, n\}$ and $\nu =$ counting measure
- the density is for $x \in \Omega$,

$$\begin{aligned} f_\theta(x) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-\theta} \\ &= \exp\left\{x\log\frac{\theta}{1-\theta} + n\log(1-\theta)\right\}\binom{n}{x} \end{aligned}$$

- $T(x) = x$
- $\eta(\theta) = \log\frac{\theta}{1-\theta}$
- $\xi(\theta) = -n\log(1-\theta)$
- $h(x) = \binom{n}{x}$
- $\Theta = (0, 1)$
- This is not in its canonical form. Now make it a natural exponential family
  - $\eta = \log\frac{\theta}{1-\theta}$
  - $\Xi = \mathcal{R}$ and

$$f_\eta(x) = \exp\{\eta x - n\log(1+e^\eta)\}\binom{n}{x}, \qquad \forall x \in \Omega = \{0, 1, \ldots, n\}$$

Example 2.6 (Exponential distribution)**.** The exponential distribution with the density

$$f_\theta(x) = \theta^{-1} \exp\{-(x-a)/\theta\}, \quad \text{for } x > a$$

for a fixed $a \in \mathcal{R}$ is an exponential family.

- We can write it in the form that

$$f_\theta(x) = \exp\{-x/\theta + a/\theta - \log\theta\} I_{(a,\infty)}(x)$$

- $T(x) = x$
- $\eta(\theta) = -1/\theta$
- $\xi(\theta) = -a/\theta + \log\theta$
- $h(x) = I_{(a,\infty)}(x)$.
- Note: if a is not fixed, then it is not an exponential family: h depends on a
- To turn it into a natural family, reparametrize $\eta = -1/\theta$ and $\Xi = (-\infty, 0)$.

Properties

- For an exponential family P_θ, there is a nonzero measure λ such that $\frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda}(\omega) > 0$ for all ω (λ-a.e.) and θ.
- Use this property to show that some families of distributions are not exponential families.

Example 2.7 (Uniform distribution)**.** Let $U(0,\theta)$ denote the uniform distribution on $(0,\theta)$. Let $\mathcal{P} = \{U(0,\theta) : \theta \in \mathcal{R}\}$. Show that this family is not an exponential family.

- Note that $\Omega = \mathcal{R}$ (or $[0,\infty)$)
- If this is an exponential family, then $\frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda}(\omega) > 0$ for all θ, all $\omega \in \mathcal{R}$ for some measure λ.
- For any $t > 0$, there is a $\theta < t$ such that $P_\theta([t,\infty)) = 0$
- Then $\lambda([\epsilon,\infty)) = 0$ for any $\epsilon > 0$, or further $\lambda((0,\infty)) = 0$
- Also, for any $t \le 0$, $P_\theta((-\infty, t]) = 0$, which implies $\lambda((-\infty, 0]) = 0$
- Then $\lambda(\mathcal{R}) = 0$.

Suppose $X_i \sim f_i$ independently and each f_i is in an exponential family, what can we say about the joint distribution of $X_1, \ldots, X_n$?

- it is still in an exponential family

Here are some other properties of exponential families

Theorem 2.8. *Let $\mathcal{P}$ be a natural exponential family with PDF*

$$f_\eta(x) = \exp\{\eta^\top T(x) - \zeta(\eta)\}h(x)$$

- *Let $T = (Y, U)$ and $\eta = (\vartheta, \varphi)$, where Y and ϑ have the same dimension. Then Y has the PDF*

  $$f_\eta(y) = \exp\{\vartheta^\top y - \zeta(\eta)\}$$

  w.r.t. a σ-finite measure depending on φ.

- *If η_0 is an interior point of the natural parameter space, then the MGF ψ_{η_0} of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0 and is given by*

  $$\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}.$$

Example 2.9 (MGF of binomial distribution)**.** Recall that

- the canonical form is given by

  $$f_\eta(x) = \exp\{\eta x - n\log(1 + e^\eta)\}\binom{n}{x}, \qquad \forall x \in \Omega = \{0, 1, \ldots, n\}$$

- $\zeta(\eta) = n\log(1 + e^\eta)$
- $T(x) = x$

  $$\begin{aligned}\psi_{\eta_0}(t) &= \exp\{n\log(1 + e^{\eta_0 + t}) - n\log(1 + e^{\eta_0})\} \\ &= \left(\frac{1 + e^{\eta_0}e^t}{1 + e^{\eta_0}}\right)^n = (1 - \theta + \theta e^t)^n\end{aligned}$$

  since $\theta = e^\eta/(1 + e^\eta)$.

2.4 Location-scale families

- sometimes, we want a model that is invariant to translation and scaling

Definition 2.10 (Location-scale families)**.** Let P be a known probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$, $\mathcal{V} \subset \mathcal{R}^k$, and $\mathcal{M}_k$ be a collection of $k \times k$ symmetric positive defintie matrices. The family

$$\{P_{(\mu,\Sigma)} : \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k\}$$

is called a location-scale family on $\mathcal{R}^k$, where

$$P_{(\mu,\Sigma)}(B) = P\left(\Sigma^{-1/2}(B - \mu)\right), B \in \mathcal{B}^k.$$

The parameter μ is called the location parameter, and Σ is called the scale parameter.

- location family: $\{P_{(\mu,I_k)} : \mu \in \mathcal{R}^k\}$, where I_k is the $k \times k$ identity matrix
- scale family: $\{P_{(0,\Sigma)} : \Sigma \in \mathcal{M}_k\}$
- location with homogeneous scaling: $\{P_{(\mu,\sigma^2 I_k)} : \mu \in \mathcal{V}, \sigma \in \mathcal{R}_{++}\}$, where $\mathcal{R}_{++} = \{x \in \mathcal{R} : x > 0\}$.
- If F is the CDF of P, then $F(\Sigma^{-1/2}(x-\mu))$ is the CDF of $P_{(\mu,\Sigma)}$
- Examples
  - exponential distributions $\text{Exp}(a, \theta)$
  - uniform distributions $U(0, \theta)$
  - k-dimensional normal distributions

2.5 Sufficiency

Recall that

- A sample is a random vector on a probability space
- A measurable function $T(X)$ of X is called a statistic if $T(X)$ is a known value whenever X is known
- A statistic often compresses the information contained in a sample $\sigma(T(X)) \subset \sigma(X)$
- Compression might lead to loss of information

Definition 2.11 (Sufficiency)**.** Let X be sample from an unknown population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. A statistic $T(X)$ is said to be sufficient for $P \in \mathcal{P}$ if and only if the conditional distribution of X given T is known (does not depend on P)

- When $\mathcal{P}$ is a parametric family indexed by $\theta \in \Theta$, we say $T(X)$ is sufficient for θ if the conditional distribution of X given T does not depend on θ.
- Interpretation: once we observe X and compute $T(X)$, the original data X do not contain further information about the unknown population P or parameter θ
- No loss of informaiton due to compression by $T(X)$ if θ is of concern
- This concept depends on the family $\mathcal{P}$.
- Sufficiency passes over to smaller classes, but not larger
  - If $T(X)$ is sufficient for $P \in \mathcal{P}$, then it is also sufficient for $P \in \mathcal{P}_0 \subset \mathcal{P}$, but not necessarily for $P \in \mathcal{P}_1 \supset \mathcal{P}$

Example 2.12 (Sum of Bernoulli trials)**.** Let $X = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ be IID from the Bernoulli distribution with PDF (w.r.t. the counting measure)

$$f_\theta(z) = \theta^z (1-\theta)^{1-z} I_{\{0,1\}}(z), \qquad z \in \mathcal{R}, \qquad \theta \in (0,1).$$

- $\mathcal{P} = \{\prod_{i=1}^n f_\theta(x_i) : \theta \in (0,1)\}$
- Take $T(X) = \sum_{i=1}^n X_i$: the number of ones in X
- Once we know $T(X)$, other information in X is about the positions of these ones
  - but they are not useful for estimating θ which is the probability of getting a one
  - they are redundant for θ
- Formally, we compute the conditional distribution of X given T. Note that

$$P(X = x \mid T = t) = \frac{P(X = x, T = t)}{P(T = t)}$$

  and $P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t} I_{\{0,1,\ldots,n\}}(t)$, for $x = (x_1, \ldots, x_n)$.
  - If $t \neq \sum_{i=1}^n x_i$, then $P(X = x, T = t) = 0$.
  - If $t = \sum_{i=1}^n x_i$, then

$$\begin{aligned} P(X = x, T = t) &= P(X = x) \\ &= \prod_{i=1}^n P(X_i = x_i) \\ &= \theta^t (1-\theta)^{n-t} \prod_{i=1}^n I_{\{0,1\}}(x_i) \end{aligned}$$

    Then

$$P(X = x \mid T = t) = \frac{1}{\binom{n}{t}} \frac{\prod_{i=1}^n I_{\{0,1\}}(x_i)}{I_{\{0,1,\ldots,n\}}(t)}$$

    does not depend on θ.

Theorem 2.13 (Factorization)**.** *Suppose that X is a sample from $P \in \mathcal{P}$ and $\mathcal{P}$ is a family of probability measures on $(\mathcal{R}^n, \mathcal{B}^n)$ dominated by a σ-finite measure ν. Then $T(X)$ is sufficient for $P \in \mathcal{P}$ if and only if there are nonnegative Borel functions h (which does not depend on P) on $(\mathcal{R}^n, \mathcal{B}^n)$ and g_P (which depends on P) on the range of T such that*

$$\frac{\mathrm{d}P}{\mathrm{d}\nu}(x) = g_P(T(x)) h(x). \tag{2.5}$$

- Intuition: the unknown part g_P involve T only
- Application: the T in an exponential family

$$f_\theta(\omega) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}(\omega) = \exp\left\{[\eta(\theta)]^\top T(\omega) - \xi(\theta)\right\} h(\omega), \omega \in \Omega,$$

  is sufficient for θ

- Application: in the example of sum of Bernoulli trials, the joint distribution of $X_1, \ldots, X_n$ is in an exponential family with $T(X) = \sum_{i=1}^n X_i$

Example 2.14 (Truncation families)**.** Let $\phi(x)$ be a positive Borel function on $(\mathcal{R}, \mathcal{B})$ such that $\int_a^b \phi(x)\mathrm{d}x < \infty$ for any a and b, $-\infty < a < b < \infty$. Let $\theta = (a, b)$, $\Theta = \{(a, b) \in \mathcal{R}^2 : a < b\}$, and

$$f_\theta(x) = c(\theta)\phi(x)I_{(a,b)}(x),$$

where $c(\theta) = [\int_a^b \phi(x)\mathrm{d}x]^{-1}$.

- $\{f_\theta : \theta \in \Theta\}$ is called a truncation family. This is a parametric family.
- It is dominated by Lebesgue measure
- Suppose $X_1, \ldots, X_n$ are IID sampled from f_θ
- The joint PDF of $X = (X_1, \ldots, X_n)$ is

$$\prod_{i=1}^n f_\theta(x_i) = [c(\theta)]^n \left[\prod_{i=1}^n \phi(x_i)\right] \left[\prod_{i=1}^n I_{(a,b)}(x_i)\right]$$

- $\prod_{i=1}^n I_{(a,b)}(x_i) = I_{(a,\infty)}(x_{(1)})I_{(-\infty,b)}(x_{(n)})$.
- So $T(X) = (X_{(1)}, X_{(n)})$ is sufficient for $\theta = (a, b)$.

Proof of Factorization Theorem. We require the following lemma.

Lemma 2.15. *Let $\{c_i\}$ be a sequence of nonnegative numbers satisfying $\sum_{i=1}^\infty c_i = 1$ and let $\{P_i\}$ be a sequence of probability measures on a common measurable space. Define $P = \sum_{i=1}^\infty c_i P_i$.*

1. *P is a probability measure;*
2. *Let ν be a σ-finite measure. Then $P_i \ll \nu$ for all i if and only if $P \ll \nu$. When $P \ll \nu$,*

$$\frac{\mathrm{d}P}{\mathrm{d}\nu} = \sum_{i=1}^\infty c_i \frac{\mathrm{d}P_i}{\mathrm{d}\nu}.$$

3. *If a family $\mathcal{P}$ is dominated by a σ-finite measure, then $\mathcal{P}$ is dominated by a probability measure $Q = \sum_{i=1}^\infty c_i P_i$ where $P_i \in \mathcal{P}$.*

Now we begin the proof of Factorization Theorem. Let us do the "only if" part first. To this end, we shall assume T is sufficient and then try to establish Eq 2.5, which amounts to prove that

$$P(A) = \int_A g_P\big(T(x)\big)h(x)dx \qquad \forall A \in \mathcal{B}^n.$$

Before we dive into the details, let us take a look at the overview of the proof. We have

$$P(A) = \int P(A|T)dP = \int E_P(I_A|T)dP.$$

Since $\mathbb{E}_P(I_A|T) = P(A|T)$ does not depend on P, we hope that a common measure Q can be found such that $\mathbb{E}_P(I_A|T) = \mathbb{E}_Q(I_A|T)$ where Q does not depend on P but dominates P for all P. Then we will have

$$P(A) = \int P(A|T)dP = \int \mathbb{E}_Q(I_A|T)\frac{dP}{dQ}dQ.$$

If, further, $\frac{dP}{dQ}$ is measurable in T and $Q << \nu$, we will have

$$P(A) = \int \mathbb{E}_Q[I_A\frac{dP}{dQ}|T]dQ = \int I_A\frac{dP}{dQ}dQ = \int_A \frac{dP}{dQ}\frac{dQ}{d\nu}d\nu$$

Then, we can take $g_{_P}(T) = \frac{dP}{dQ}$ and $h = \frac{dQ}{d\nu}$.

Now here are the details. Let Q be the probability measure in Lemma 2.15. By Fubini's theorem and Lemma 2.15, for any $B \in \sigma(T)$,

$$Q(A \cap B) = \sum_{j=1}^{\infty} c_j P_j(A \cap B) = \sum_{j=1}^{\infty} c_j \int_B P(A|T)dP_j$$

$$= \int_B \sum_{j=1}^{\infty} c_j P(A|T)dP_j = \int_B P(A|T)dQ,$$

where the second equality holds since $P(A|T)$ does not depend on $P \in \mathcal{P}$. Hence, $P(A|T) = \mathbb{E}_Q(I_A|T)$ a.s. Q, where $\mathbb{E}_Q(I_A|T)$ denotes the conditional expectation of I_A given T w.r.t. Q. Let dP/dQ be the Radon-Nikodym derivative of P with respect to Q on the space $(\mathcal{R}^n, \sigma(T), Q)$. Then dP/dQ is measurable on $(\mathcal{R}^n, \sigma(T), Q)$ and hence there is a measurable function $g_{_P}(T)$ of T such that $g_{_P}(T) = dP/dQ$. Then

$$P(A) = \int P(A|T)dP = \int E_Q(I_A|T)dP = \int \mathbb{E}_Q(I_A|T)g_{_P}(T)dQ$$

$$= \int \mathbb{E}_Q[I_A g_{_P}(T)|T]dQ = \int I_A g_{_P}(T)dQ = \int_A g_{_P}(T)\frac{dQ}{d\nu}d\nu$$

for any $A \in \mathcal{B}^n$. Hence,

$$\frac{dP}{d\nu}(x) = g_{_P}\big(T(x)\big)h(x) \tag{2.6}$$

holds with $h = dQ/d\nu$. This establishes the "only if" part of the theorem.

Now we move to the "if" part. Suppose that (2.6) holds. By Chain rule, $\frac{dP}{d\nu} = \frac{dQ}{d\nu}\frac{dP}{dQ}$. Hence

$$\frac{dP}{dQ} = \frac{dP}{d\nu}\Big/\frac{dQ}{d\nu} = \frac{dP}{d\nu}\Big/\sum_{i=1}^{\infty} c_i\frac{dP_i}{d\nu} = g_{_P}(T)\Big/\sum_{i=1}^{\infty} c_i g_{P_i}(T), \tag{2.7}$$

a.s. Q, where the second equality follows from Lemma 2.15.

Let $A \in \sigma(X)$ and $P \in \mathcal{P}$. It suffices to show

$$P(A|T) = \mathbb{E}_Q(I_A|T) \quad a.s.\ P, \tag{2.8}$$

where $\mathbb{E}_Q(I_A|T)$ denotes the conditional expectation of I_A given T w.r.t. Q. This is because $\mathbb{E}_Q(I_A|T)$ does not vary with $P \in \mathcal{P}$, and result (2.8) and Theorem 1.7 (of the textbook) imply that the conditional distribution of X given T is determined by $\mathbb{E}_Q(I_A|T)$, $A \in \sigma(X)$.

By (2.7), dP/dQ is a Borel function of T. For any $B \in \sigma(T)$,

$$\int_B \mathbb{E}_Q(I_A|T)dP = \int_B \mathbb{E}_Q(I_A|T)\frac{dP}{dQ}dQ$$

$$= \int_B \mathbb{E}_Q\left(I_A\frac{dP}{dQ}\middle|T\right)dQ = \int_B I_A\frac{dP}{dQ}dQ = \int_B I_A dP.$$

This proves (2.8) and completes the proof.

□

Definition 2.16 (Minimal sufficiency)**.** Let T be a sufficient statistic for $P \in \mathcal{P}$. T is called a minimal sufficient statistic if and only if, for any other statistic S sufficient for $P \in \mathcal{P}$, there is a measurable function ψ such that $T = \psi(S)$ $\mathcal{P}$-a.s. (P-a.s. for all $P \in \mathcal{P}$).

- Minimal sufficient statistics are unique (almost surely): If both T and S are minimal sufficient statistics (for a family $\mathcal{P}$), then there is a one-to-one measurable function ψ such that $T = \psi(S)$ $\mathcal{P}$-a.s.
- Minimal sufficient statistics exist under weak conditions, e.g., $\mathcal{P}$ contains distributions on $\mathcal{R}^k$ dominated by a σ-finite measure.

Example 2.17. Let $X_1, \ldots, X_n \sim P_\theta = U(\theta, \theta+1)$ for $\theta \in \mathcal{R}$. Suppose $n > 1$.

- This is a location family, with the location parameter θ
- The joint PDF is

$$f_\theta(x) = \prod_{i=1}^n I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta), \qquad x = (x_1, \ldots, x_n) \in \mathcal{R}^n.$$

  - note that $\theta < x_{(1)} \le \cdots \le x_{(n)} < \theta + 1$ iff $I_{(x_{(n)}-1,x_{(1)})}(\theta) = 1$

- By Factorization Theorem, $T = (X_{(1)}, X_{(n)})$ is sufficient for θ
- To see T is minimal, we note that for any joint density f_θ,

$$x_{(1)} = \sup\{\theta : f_\theta(x) > 0\}$$
$$x_{(n)} = 1 + \inf\{\theta : f_\theta(x) > 0\}$$

- Suppose $S(X)$ is another sufficient statistic for θ.

- By Factorization Theorem, there are Borel functions h and g_θ such that

$$f_\theta(x) = g_\theta(S(x))h(x)$$

- For x with $h(x) > 0$,

$$x_{(1)} = \sup\{\theta : g_\theta(S(x)) > 0\}$$
$$x_{(n)} = 1 + \inf\{\theta : g_\theta(S(x)) > 0\}$$

  which are functions of S!

- So, $T(x) = \psi(S(x))$ when $h(x)$ for a measurable function ψ.
- Note that $h > 0$ $\mathcal{P}$-a.s.
- Therefore, T is minimal.

Theorem 2.18. *Let $\mathcal{P}$ be a family of distributions on $\mathcal{R}^k$.*

1. *Suppose $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_0$-a.s. implies $\mathcal{P}$-a.s. If T is sufficient for $P \in \mathcal{P}$ and minimal sufficient for $P \in \mathcal{P}_0$, then T is minimal sufficient for $P \in \mathcal{P}$.*

   Proof: If S is sufficient for $\mathcal{P}$, then it is also sufficient for $\mathcal{P}_0$. Thus, $T = \psi(S)$ $\mathcal{P}_0$-a.s. for a measurable function ψ. Then $T = \psi(S)$ $\mathcal{P}$-a.s. since $\mathcal{P}_0$-a.s. implies $\mathcal{P}$-a.s by assumption.

2. *Suppose that $\mathcal{P}$ contains only PDF $f_0, f_1, \ldots$ w.r.t. a σ-finite measure. Define $f_\infty(x) = \sum_{i=0}^\infty c_i f_i(x)$, where $c_i > 0$ and $\sum_{i=0}^\infty c_i = 1$. Let $T_i(x) = f_i(x)/f_\infty(x)$ when $f_\infty(x) > 0$. Then $T(X) = (T_0(X), T_1(X), \ldots)$ is minimal sufficient for $\mathcal{P}$. If $\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\}$ for all i, then f_∞ can be replace with f_0, in which case $T(X) = (T_1, T_2, \ldots)$ is minimal sufficient for $\mathcal{P}$.*

   Proof: The construction of f_∞ assures that $f_\infty > 0$ $\mathcal{P}$-a.s. Let $g_i(T) = T_i$. Then $f_i(x) = g_i(T(x))f_\infty(x)$. By Factorization theorem, T is sufficient for $\mathcal{P}$. Suppose $S(X)$ is another sufficient statistic. By Factorization theorem, $f_i(x) = \tilde{g}_i(S(x))h(x)$ for all i and some $\tilde{g}$ and h. Then

$$T_i(x) = \tilde{g}_i(S(x))/\sum_{j=0}^\infty c_j \tilde{g}_j(S(x))$$

   when $f_\infty(x) > 0$. Thus, T is minimal sufficient for $\mathcal{P}$.

3. *Suppose that $\mathcal{P}$ contains PDF f_P w.r.t. a σ-finite measure and that there exists a sufficient statistic $T(X)$ such that, for any possible values x and y of X, $f_P(x) = f_P(x)\phi(x, y)$ for all P implies $T(x) = T(y)$, where ϕ is a measurable function. Then $T(X)$ is minimal sufficient for $\mathcal{P}$.*

   Proof: See the textbook.

Example 2.19. Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be a p-dimensional exponential family with PDFs

$$f_\theta(x) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\}h(x).$$

By Factorization Theorem, $T(X)$ is sufficient for $\theta \in \Theta$. Suppose that there exists $\Theta_0 = \{\theta_0, \theta_1, \ldots, \theta_p\} \subset \Theta$ such that the vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0)$, $i = 1, \ldots, p$, are linearly independent in $\mathcal{R}^p$. (This is true if the exponential family is of full rank). Then T is also minimal sufficient.

- Method A: To use Theorem 2.18(1). Let $\mathcal{P}_0 = \{f_\theta : \theta \in \Theta_0\}$. Note that the set $\{x : f_\theta(x) > 0\}$ does not depend on θ. It follows from Theorem 2.3(ii) with $f_\infty = f_{\theta_0}$ that

  $$S(X) = \Big(\exp\{\eta_1^\tau T(x) - \xi_1\}, \ldots, \exp\{\eta_p^\tau T(x) - \xi_p\}\Big)$$

  is minimal sufficient for $\theta \in \Theta_0$. Since η_i's are linearly independent, there is a one-to-one measurable function ψ such that $T(X) = \psi(S(X))$ a.s. $\mathcal{P}_0$. Hence, T is minimal sufficient for $\theta \in \Theta_0$. It is easy to see that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. Thus, by Theorem 2.18(1), T is minimal sufficient for $\theta \in \Theta$.

- Method B: To use Theorem 2.18(3). Let $\phi(x, y) = h(x)/h(y)$. Then

  $$\begin{aligned} & f_\theta(x) = f_\theta(y)\phi(x, y) \\ \Rightarrow \quad & \exp\{[\eta(\theta)]^\tau [T(x) - T(y)\} = 1 \\ \Rightarrow \quad & T(x) = T(y). \end{aligned}$$

  Since T is sufficient, by Theorem 2.18 (3), T is also minimal sufficient.

Example 2.20 (revisited)**.** Let $X_1, \ldots, X_n \sim P_\theta = U(\theta, \theta + 1)$ for $\theta \in \mathcal{R}$. Suppose $n > 1$.

- This is a location family, with the location parameter θ
- The joint PDF is

  $$f_\theta(x) = \prod_{i=1}^n I_{(\theta, \theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta), \qquad x = (x_1, \ldots, x_n) \in \mathcal{R}^n.$$

- Another way to show that $T = (X_{(1)}, X_{(n)})$ is minimal sufficient:

  Let $\phi(x, y) = 1$. Then

  $$\begin{aligned} & f_\theta(x) = f_\theta(y), \text{ for all } \theta \\ \Rightarrow \quad & I_{(x_{(n)}-1, x_{(1)})}(\theta) = I_{(y_{(n)}-1, y_{(1)})}(\theta) \text{ for all } \theta \\ \Rightarrow \quad & (x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)}). \end{aligned}$$

  By Theorem 2.18 (3), $T = (X_{(1)}, X_{(n)})$ is minimal sufficient.

- Note that sufficiency depends on the family $\mathcal{P}$, or equivalently, our statistical model, which could (likely) be wrong.

- In this case, the concept of sufficiency might not be useful
- However, some statistics, like sample mean, sample variance, minimum and maximum statistics, are sufficient for many models
- They are still useful even we don't know the correct model

2.6 Completeness

- A minimal sufficient statistic might not be "minimal" in some sense – not always the "simplest sufficient statistic"
- May still contain redundant information
- e.g. if $\overline{X}$ is minimal sufficient, then so is $(\overline{X}, \exp(\overline{X}))$
- A statistic $V(X)$ is said to be ancillary if its distribution does not depend on the population P
- $V(X)$ is called first-order ancillary if $\mathbb{E}_P V(X)$ is independent of P
- e.g. trivial ancillary statistic: $V(X) = c$
- Note that $\sigma(V(X)) \subset \sigma(X)$.
  - If $V(X)$ is a nontrivial ancillary statistic, then $\sigma(V(X))$ is a nontrivial σ-field thap.d.f.'st does not contain any information about P.
- Similary, $\sigma(V(S(X))) \subset \sigma(S(X))$
  - if $V(S(X))$ is ancillary, then $\sigma(S(X))$ contains a nontrivial σ-field that does not contain any information about P
  - "data" $S(X)$ may be further compressed
  - Some sufficient statistics can not be further compressed in this sense

Definition 2.21 (Completeness)**.** A statistic $T(X)$ is said to be complete for $P \in \mathcal{P}$ if and only if, for any Borel function f, $\mathbb{E}_P f(T) = 0$ for all $P \in \mathcal{P}$ implies that $f(T) = 0$ $\mathcal{P}$-a.s. T is said to be boundedly complete if and only if the previous statement holds for any bounded Borel functions f.

- A complete statistic contains "completely" useful information about P
  - no redundance
- Clearly, a complete statistic is boundedly complete.
- If T is complete and $S = \psi(T)$ for a measurable function, then S is complete

- Similar result holds for bounded completeness
- A complete and sufficient statistic is minimial sufficient

Proposition 2.22. *If $\mathcal{P}$ is in an exponential family of full rank with PDFs given by*

$$f_\eta(x) = \exp\{\eta^\tau T(x) - \zeta(\eta)\}h(x),$$

then $T(X)$ is complete and sufficient for $\eta \in \Xi$.

Proof. We have shown that T is sufficient. We now show that T is complete. Suppose that there is a function f such that $\mathbb{E}[f(T)] = 0$ for all $\eta \in \Xi$.

$$\int f(t)\exp\{\eta^\tau t - \zeta(\eta)\}d\lambda = 0 \quad \text{for all } \eta \in \Xi,$$

where λ is a measure on $(\mathcal{R}^p, \mathcal{B}^p)$. Let η_0 be an interior point of Ξ. Then

$$\int f_+(t)e^{\eta^\tau t}d\lambda = \int f_-(t)e^{\eta^\tau t}d\lambda \quad \text{for all } \eta \in N(\eta_0), \tag{2.9}$$

where $N(\eta_0) = \{\eta \in \mathcal{R}^p : \|\eta - \eta_0\| < \epsilon\}$ for some $\epsilon > 0$. In particular,

$$\int f_+(t)e^{\eta_0^\tau t}d\lambda = \int f_-(t)e^{\eta_0^\tau t}d\lambda = c.$$

If $c = 0$, then $f = 0$ a.e. λ. If $c > 0$, then $c^{-1}f_+(t)e^{\eta_0^\tau t}$ and $c^{-1}f_-(t)e^{\eta_0^\tau t}$ are p.d.f.'s w.r.t. λ and result (2.9) implies that their m.g.f.'s are the same in a neighborhood of 0. By Theorem 1.6(ii) (of the textbook), $c^{-1}f_+(t)e^{\eta_0^\tau t} = c^{-1}f_-(t)e^{\eta_0^\tau t}$, i.e., $f = f_+ - f_- = 0$ a.e. λ. Hence T is complete. □

Example 2.23. Suppose that $X_1, \ldots, X_n$ are IID random variables having the $N(\mu, \sigma^2)$ distribution, $\mu \in \mathcal{R}$, $\sigma > 0$.

- It is easy to check that the joint PDF is

  $$(2\pi)^{-n/2}\exp\left\{\eta_1 T_1 + \eta_2 T_2 - n\zeta(n)\right\},$$

  where
  - $T_1 = \sum_{i=1}^n X_i$
  - $T_2 = -\sum_{i=1}^n X_i^2$
  - $\eta = (\eta_1, \eta_2) = (\mu/\sigma^2, 1/(2\sigma^2))$
- This is a natural exponential familyof full rank: $\Xi = \mathcal{R} \times (0, \infty)$ is an open set of $\mathcal{R}^2$
- So $T(X) = (T_1(X), T_2(X))$ is complete and sufficient for η
- There is a one-to-one correspondence between η and $\theta = (\mu, \sigma^2)$
  - T is also complete and sufficient for θ

- There is a one-to-one correspondence between $(\overline{X}, S^2)$ and (T_1, T_2)
  - $(\overline{X}, S^2)$ is complete and sufficient for θ

Example 2.24. Let $X_1, \ldots, X_n \sim P_\theta = U(0, \theta)$ be IID for $\theta > 0$. The largest order statistic, $X_{(n)}$, is complete and sufficient for θ.

- The sufficiency follows from Factorization theorem: the joint PDF is $\theta^{-n} I_{(0,\theta)}(x_{(n)})$
- The CDF of $X_{(n)}$ is

$$\begin{aligned} F_n(x) &= P(X_{(n)} \le x) = P(X_1 \le x, \ldots, X_n \le x) \\ &= \prod_{i=1}^{n} P(X_i \le x) = \frac{x^n}{\theta^n} I_{(0,\theta)}(x). \end{aligned}$$

- The PDF is then

$$f(x) = \frac{n x^{n-1}}{\theta^n} I_{(0,\theta)}(x).$$

- Let g be a Borel function on $[0, \infty)$ s.t. $\mathbb{E}[g(X_{(n)})] = 0$ for all $\theta > 0$. Then

$$\int_0^\theta g(x) x^{n-1} \mathrm{d}x = 0$$

  for all $\theta > 0$.

- Differentiate the above w.r.t. θ

$$g(\theta)\theta^{n-1} = 0$$

- Thus, $g(\theta) = 0$ for all $\theta > 0$.
- By definition, $X_{(n)}$ is complete for θ

Theorem 2.25 (Basu)**.** *Let V and T be two statistics of X from a population $P \in \mathcal{P}$. If V is ancillary and T is boundedly complete and sufficient for $P \in \mathcal{P}$, then V and T are independent w.r.t. any $P \in \mathcal{P}$.*

- Intuition: V does not contain information about P, while T carries non-redundant and sufficient information about P. This suggests that $\sigma(V)$ and $\sigma(T)$ are independent.

Proof: Let B be an event on the range of V. Since V is ancillary, $P(V^{-1}(B))$ is a constant. As T is sufficient, $\mathbb{E}[I_B(V)|T]$ is a function of T (not dependent on P). Because

$$\mathbb{E}\{\mathbb{E}[I_B(V)|T] - P(V^{-1}(B))\} = 0 \quad \text{for all } P \in \mathcal{P},$$

by the bounded completeness of T,

$$P(V^{-1}(B)|T) = \mathbb{E}[I_B(V)|T] = P(V^{-1}(B)) \quad \text{a.s. } \mathcal{P}$$

Let A be an event on the range of T. Then

$$P(T^{-1}(A) \cap V^{-1}(B)) = \mathbb{E}\{\mathbb{E}[I_A(T) I_B(V)|T]\} = \mathbb{E}\{I_A(T)\mathbb{E}[I_B(V)|T]\}$$

$$= \mathbb{E}\{I_A(T)P(V^{-1}(B))\} = P(T^{-1}(A))P(V^{-1}(B)).$$

Hence T and V are independent w.r.t. any $P \in \mathcal{P}$.

Example 2.26. Show that the sample mean and sample variance of an IID sample of normally distributed random variables are independent.

- Let $X_1, \ldots, X_n$ be IID $\sim N(\mu, \sigma^2)$
- Suppose σ^2 is known
- We can easily show that the family $\{N(\mu, \sigma^2) : \mu \in \mathcal{R}\}$ is an exponential family of full rank
- Natural parameter: $\eta = \mu/\sigma^2$
- Then $\overline{X}$ is complete and sufficient for η and μ, according to Proposition 2.22
- Note that $S^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \overline{Z})^2$, where $Z_i = X_i - \mu \sim N(0, \sigma^2)$
- S^2 is ancillary w.r.t. μ, since its distribution does not depend on μ
- By Basu's theorem, $\overline{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ for any $\mu \in \mathcal{R}$
- Since σ^2 is arbitrary, $\overline{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ for any μ and $\sigma^2 > 0$

Lecture 3: Statistical Decision Theory

Lecturer: LIN Zhenhua ST5215 AY2019/2020 Semester I

3.1 Decision rules, loss functions and risks

Consider the estimation problem. Let $X = (X_1, \ldots, X_n)$ be a sample from a population $P_\theta \in \mathcal{P}$, where $\theta \in \Theta$. Suppose $\mathbb{E}_\theta X_i \equiv \theta$. We estimate θ by $\hat{\theta}(X) = \overline{X} = n^{-1}\sum_{i=1}^n X_i$. This means, if we observe $X = x$, we estimate θ by $\hat{\theta}(x)$.

View this problem in the following way:

- After we observe $X = x$, we take an *action*: $\hat{\theta}(x) = \overline{x}$
- The set of allowable actions is Θ: *action space*, denoted by $\mathbb{A}$ and endowed with a σ-field $\mathcal{F}_\mathbb{A}$
- $\hat{\theta}$ is a *decision rule*: a measurable function from the range of X to $(\mathbb{A}, \mathcal{F}_\mathbb{A})$
- A *statistical decision* is an action that we take after we observe X
- For a problem, there are many decision rules: $\hat{\theta}(X) = \sum_{i=2}^{n-1} X_{(i)}$
- How to measure quality of decision rules?
  - loss function: a function $L : \mathcal{P} \times \mathbb{A} \to [0, \infty)$, Borel for each fixed $P \in \mathcal{P}$
  - When $\mathcal{P}$ is parametric and θ is the parameter, $L : \Theta \times \mathbb{A} \to [0, \infty)$
  - For a decision rule, $L(P, T(x))$ is the "loss" if we take the action $T(x)$ after observe $X = x$
  - The "average" loss, called *risk*, for a rule is defined by
$$R_T(P) = \mathbb{E}_P L(P, T(X)) = \int L(P, T(X)) \mathrm{d}P$$
- Note that the risk depends on both T and P (also the loss function which is often predetermined and fixed)

Now we can quantify the quality of a decision rule

- T_1 is as good as T_2 if
$$R_{T_1}(P) \le R_{T_2}(P), \qquad \forall P \in \mathcal{P}$$
- We say T_1 is better than T_2, if T_1 is as good as T_2 and $R_{T_1}(P) < R_{T_2}(P)$ for some $P \in \mathcal{P}$
  - we also say T_2 is dominated by T_1 in this case

- T_1 and T_2 are equivalent (equivalently good) if and only $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$
- Let $\mathfrak{J}$ be the collection of decision rules under consideration
- T_* is called an $\mathfrak{J}$-optimal rule if T_* is as good as any other rule in $\mathfrak{J}$
  - T_* is optimal if $\mathfrak{J}$ contains all possible rules

Example 3.1 (Measurement problem revisited)**.** Recall that we are to measure a quantity θ of an object. We take multiple measurements of the object and record the results $X_1, \ldots, X_n$.

- Action space $\mathbb{A} = \Theta$ the set of all possible values of θ
- $(\mathbb{A}, \mathcal{F}_{\mathbb{A}}) = (\Theta, \mathcal{B}_\Theta)$
- A simple decision rule: $T(X) = \overline{X}$
- A common loss function: squared error loss $L(P_\theta, a) = (\theta - a)^2$ for $\theta \in \Theta$ and $a \in \mathbb{A}$
  - risk function with the squared error loss is called the mean squared error (MSE)
- Suppose $X_1, \ldots, X_n$ are IID with mean θ and known variance σ^2
- The risk of T is

$$\begin{aligned} R_T(\theta) &= \mathbb{E}_\theta(\theta - \overline{X}) \\ &= (\theta - \mathbb{E}_\theta \overline{X})^2 + \mathbb{E}_\theta(\mathbb{E}_\theta \overline{X} - \overline{X})^2 \\ &= (\theta - \theta)^2 + \mathrm{Var}(\overline{X})^2 \\ &= \sigma^2/n \end{aligned}$$

  - the risk decrease as sample size n increases (variability of the estimator decreases as sample becomes larger)
  - it increases with σ^2 (problem with large variance is harder)
- This kind of problem is called *estimation*
- The decision rule here is called an *estimator*

Example 3.2 (Hypothesis test)**.** Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_1 = \{P \in \mathcal{P} : P \in \mathcal{P}_0\}$. A general hypothesis testing problem can be formulated as deciding which of the following statements is true:

$$H_0 : P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1 : P \in \mathcal{P}_1.$$

- H_0 is called the null hypothesis
- H_1 is called the alternative hypothesis
- The action space $\mathbb{A} = \{0, 1\}$

- A decision rule in this case is called a *test*
- $T : \mathcal{X} \to \{0, 1\}$, so mut be in the form $I_C(X)$ for some $C \subset \mathcal{X}$
- C is called the *rejection region* or *critical region* for testing H_0 versus H_1
- A common loss function: 0-1 loss, $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $L(P, j) = 1$ otherwise, $j = 0, 1$.
- The risk is

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & \text{when } P \in \mathcal{P}_0, \quad \text{[Type I error]} \\ P(T(X) = 0) = P(X \notin C) & \text{when } P \in \mathcal{P}_1, \quad \text{[Type II error]} \end{cases}$$

3.2 Admissibility and optimality

Definition 3.3 (Admissibility)**.** Let $\mathfrak{J}$ be a class of decision rules. A decision rule $T \in \mathfrak{J}$ is called $\mathfrak{J}$-admissible (or admissible if $\mathfrak{J}$ contains all possible rules) if and only if there does not exist any $S \in \mathfrak{J}$ that is better than T (in terms of the risk with respect to a loss function).

- In principle, inadmissible rule shall not be used
- Relationship between admissibility and optimality
  - If T_* is $\mathfrak{J}$-optimal, then it is $\mathfrak{J}$-admissible
  - If T_* is $\mathfrak{J}$-optimal and T_0 is $\mathfrak{J}$-admissible, then T_0 is also $\mathfrak{J}$-optimal and is equivalent to T_*
  - If there are two $\mathfrak{J}$-admissible rules that are not equivalent, then there does not exist any $\mathfrak{J}$-optimal rule

For convex loss function, admissible rules are functions of sufficient statistics

Theorem 3.4 (Rao-Blackwell)**.** *Let T be a sufficient statistic for $P \in \mathcal{P}$, $T_0 \in \mathcal{R}^k$ be a decision rule satisfying $\mathbb{E}_P\|T_0\| < \infty$ for all $P \in \mathcal{P}$. Let $T_1 = \mathbb{E}\{T_0(X) \mid T\}$. Then $R_{T_1}(P) \leq R_{T_0}(P)$ when the loss function $L(P, a)$ is convex in a. If L is strictly convex in a and T_0 is not a function of T, then T_0 is inadmissible.*

Example 3.5 (Poisson process)**.** Phone calls arrive at a switchboard according to a Poisson process at an average rate of λ per minute. This rate is not observable, but the numbers $X_1, \ldots, X_n$ of phone calls that arrived during n successive one-minute periods are observed. It is desired to estimate the probability $e^{-\lambda}$ that the next one-minute period passes with no phone calls.

- Start with the following (extremely) naive estimator

$$T_0 = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{otherwise} \end{cases}$$

- Possion distributions form an exponential family
- By Factorization theorem, $T = T_n = \sum_{i=1}^n X_i$ is a sufficient statistic
- Let $T_1(t) = \mathbb{E}\{T_0 \mid T = t\}$

$$\begin{aligned} T_1(t) &= \mathbb{E}\{I_{X_1=0} \mid T = t\} \\ &= P\left(X_1 = 0 \mid \sum_{i=1}^n X_i = t\right) \\ &= P\left(X_1 = 0, \sum_{i=2}^n X_i = t\right) / P\left(\sum_{i=1}^n X_i = t\right) \\ &= P\left(X_1 = 0\right) P\left(\sum_{i=2}^n X_i = t\right) / P\left(\sum_{i=1}^n X_i = t\right) \\ &= e^{-\lambda} \frac{((n-1)\lambda)^t e^{-(n-1)\lambda}}{t!} \frac{t!}{(n\lambda)^t e^{-n\lambda}} \\ &= \left(1 - \frac{1}{n}\right)^t. \end{aligned}$$

- For big n, $T \approx n\lambda$ in high probability, and thus

$$T_1 \approx \left(1 - \frac{1}{n}\right)^{n\lambda} \approx e^{-\lambda}.$$

3.3 Unbiasedness

- Optimal rule often does not exist
- We often restrict us to a certain class of decision rules and try to find the best among the class

Definition 3.6 (Unbiasedness)**.** In an estimation problem, the bias of an estimator $T(X)$ of a real-valued parameter θ of the unknown population is defined to be $b_T(P) = \mathbb{E}_P(T(X)) - \theta$ (denoted by $b_T(\theta)$ for a parametric family indexed by θ). An estimator $T(X)$ is said to be unbiased for θ if and only if $b_T(P) = 0$ for all $P \in \mathcal{P}$.

Example 3.7 (Measurement problem revisited)**.** Recall that we are to measure a quantity θ of an object. We take multiple measurements of the object and record the results $X_1, \ldots, X_n$.

- Suppose $X_1, \ldots, X_n$ are IID with mean θ and known variance σ^2
- Consider the class $\mathfrak{J} = \{T(X) = \sum_{i=1}^n c_i X_i : \sum_{i=1}^n c_i = 1\}$
- Then $\mathbb{E}_\theta(T(X)) = \theta$ so all estimators in $\mathfrak{J}$ is unbiased.

- The risk under squared error loss is

$$\begin{aligned} R_T(\theta) &= \mathbb{E}_\theta (T(X) - \theta)^2 \\ &= \mathbb{E}_\theta \left(\sum_{i=1}^n c_i X_i - \theta \right)^2 \\ &= \mathrm{Var}_\theta \left(\sum_{i=1}^n c_i X_i \right) \\ &= \sigma^2 \sum_{i=1}^n c_i^2 \end{aligned}$$

  This is minimized when $c_i = 1/n$ for all i

- So $T(X) = \overline{X}$ is the $\mathfrak{J}$-optimal estimator

- The MSE of an estimator can always be divided into two compnents: (square of) bias and variance

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &= \mathbb{E}_\theta (\hat{\theta} - \theta)^2 \\ &= (\mathbb{E}_\theta \hat{\theta} - \theta)^2 + \mathbb{E}_\theta (\hat{\theta} - \mathbb{E}_\theta \hat{\theta}) \\ &= \mathrm{bias}^2 + \mathrm{Var}(\hat{\theta}). \end{aligned}$$

- Bias and variance ofte compete with each other

- Trade-off often must be made

3.4 Consistency

- It is hard to compute the exact risk in practice, in particular when the model is nonparametric

- When the sample size is large (well, not too small, I mean), we shall resort to asymptotic criteria

  - by using CLT, SLLN, WLLN, etc

- We treat a sample $X = (X_1, \ldots, X_n)$ as a member of a sequence corresponding to $n = 1, 2, \ldots$

- Similarly, a statistic $T(X)$, often denoted by T_n to emphasize its dependence on the sample size n, is viewed as a member of a sequence $T_1, T_2, \ldots$

- Weakness: it is hard to deterimine whether the sample size n is large enough...

  - so we complement asymptotic analysis by numerical studies

- Recently, nonasymptotic bounds are also popular to characterize a statistic/estimator (maybe to be covered in the second part of this course...)

- Intuitively, a good estimator shall be close to its target parameter when n is large in some sense

Definition 3.8 (Consistency of point estimators)**.** Let $X = (X_1, \ldots, X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be an estimator of θ for every n.

1. $T_n(X)$ is called consistent for θ if and only if $T_n(X) \xrightarrow{P} \theta$ w.r.t. any $P \in \mathcal{P}$.
2. Let $\{a_n\}$ be a sequence of positive constants diverging to ∞. $T_n(X)$ is called a_n-consistent for θ if and only if $a_n\{T_n(X) - \theta\} = O_P(1)$ w.r.t. any $P \in \mathcal{P}$.
3. $T_n(X)$ is called strongly consistent for θ if and only if $T_n(X) \xrightarrow{a.s.} \theta$ w.r.t. any $P \in \mathcal{P}$
4. $T_n(X)$ is called L_r-consistent for θ if and only if $T_n(X) \xrightarrow{L^r} \theta$ w.r.t. any $P \in \mathcal{P}$ for some fixed $r > 0$.

Note that,

- "consistent" in (1) is the weakest one among these concepts of consistency
  - but the most common one in statistics
  - also the most basic requirement for an estimator
- In a_n-consistency, $a_n = \sqrt{n}$ is often the case
- Example: the sample mean is strongly consistent for the population mean by SLLN
- A more interesting example: $\overline{X}^2$ is $\sqrt{n}$-consistent for μ^2 under the assumption that P has a finite variance.
  - note that $\sqrt{n}(\overline{X}^2 - \mu^2) = \sqrt{n}(\overline{X} - \mu)(\overline{X} + \mu)$
  - $\overline{X}$ is $\sqrt{n}$-consistent for μ by CLT
  - $\overline{X} + \mu = O_P(1)$
  - If $Y_n = O_P(b_n)$ and $Z_n = O_P(c_n)$, then $Y_n Z_n = O_P(b_n c_n)$

3.5 Asymptotic unbiasedness

- Unbiasedness is a good property
- However, in some cases, it is impossible to have an unbiased estimator
- Moreover, a slight bias might not be a bad thing
  - trade for significantly reduced variability

- Nevertheless, asymptotically, the bias shall be small

Definition 3.9 (Asymptotic unbiasedness)**.** An estimator $T_n(X)$ for θ is called asymptotically unbiased if $b_{T_n}(\theta) \equiv \mathbb{E}_\theta T_n(X) - \theta \to 0$ as $n \to \infty$.

- Note that the definition in the textbook is more general than here
- Any consistent estimator is asymptotically unbiased
- If T_n is consistent for θ, then $g(T_n)$ is asymptotically unbiased for $g(\theta)$ for any continuous g

Lecture 4: UMVUE

Lecturer: LIN Zhenhua | ST5215 | AY2019/2020 Semester I

4.1 UMVUE

- For squred error loss, the risk of an unbiased estimator is equal to its variance
- We can compare unbiased estimators by their variance

Definition 4.1 (UMVUE)**.** An unbiased estimator $T(X)$ of θ is called the uniformly minimum variance unbiased estimator (UMVUE) if and only if $\mathrm{Var}(T(X)) \leq \mathrm{Var}(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of θ.

- "uniformly" refers to "for any $P \in \mathcal{P}$"
- $R_T(\theta) \leq R_U(\theta)$ under squared error loss
- a UMVUE estimator is $\mathfrak{J}$-optimal in MSE with $\mathfrak{J}$ being the class of all unbiased estimators.

Theorem 4.2 (Lehmann-Scheffé)**.** *Suppose that there exists a sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$. If there exists an unbiased estimator for θ, then there is a unique unbiased estimator of θ that is of the form $h(T)$ with a Borel function h. Furthermore, $h(T)$ is the unique UMVUE of θ.*

Proof:

- By assumption, there is an unbiased estimator $\hat{\theta}$ for θ.
- Let $h(T) = \mathbb{E}(\hat{\theta} \mid T)$. Then $\mathbb{E}h(T) = \mathbb{E}\hat{\theta} = \theta$.
  - $h(T)$ is unbiased for θ
- Suppose $g(T)$ is another unbiased estimator of θ
- $\mathbb{E}\{h(T) - g(T)\} = 0$
- The completeness of T implies that $h - g = 0$ $\mathcal{P}$-a.s.
- The squared error loss is strictly convex. So any admissible estimator must be of the form $h(T)$, by Rao-Blackwell theorem
- $h(T)$ is the only (possible) admissible unbiased estimator, and thus UMVUE.

Example 4.3. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$. In previous lectures, we have shown that the order statistic $X_{(n)}$ is sufficient and complete with Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$.

- We observe that
$$\mathbb{E}X_{(n)} = n\theta^{-n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$
- $\mathbb{E}\{(n+1)X_{(n)}/n\} = \theta$
- By Lehmann-Scheffé theorem, $\hat{\theta} = (n+1)X_{(n)}/n$ is the UMVUE for θ

Example 4.4. Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathcal{P}$.

- In many cases, the vector of order statistics, $T = (X_{(1)}, ..., X_{(n)})$, is sufficient and complete for $P \in \mathcal{P}$. (For example, $\mathcal{P}$ is the collection of all Lebesgue p.d.f.'s.)
- An estimator $\varphi(X_1, ..., X_n)$ is a function of T iff the function φ is symmetric in its n arguments.
- A symmetric unbiased estimator $h(T(X))$ of any estimable θ is the UMVUE. This is because, due to symmetry, $h(T(X))$ is a function of the order statistic T, and T is sufficient and complete for many nonparametric families.

The following are some examples:

- $\bar{X}$ is the UMVUE of $\theta = \mathbb{E}X_1$;
- S^2 is the UMVUE of $\text{Var}(X_1)$;
- $n^{-1}\sum_{i=1}^n X_i^2 - S^2$ is the UMVUE of $(\mathbb{E}X_1)^2$;
- $F_n(t)$ is the UMVUE of $P(X_1 \le t)$ for any fixed t.

The previous conclusions are not true if T is *not* sufficient and complete for $P \in \mathcal{P}$. For example, if $n > 2$ and $\mathcal{P}$ contains all symmetric distributions having Lebesgue p.d.f.'s and finite means, then below we show that there is no UMVUE for $\mu = EX_1$.

Proof.

- Suppose that T is a UMVUE of μ.
- Let $\mathcal{P}_1 = \{N(\mu, 1) : \mu \in \mathcal{R}\}$. Since the sample mean $\bar{X}$ is UMVUE when $\mathcal{P}_1$ is considered, and the Lebesgue measure is dominated by any $P \in \mathcal{P}_1$, we conclude that $T = \bar{X}$ a.e. Lebesgue measure.

- Let $\mathcal{P}_2$ be the family of uniform distributions on $(\theta_1 - \theta_2, \theta_1 + \theta_2)$, $\theta_1 \in \mathcal{R}$, $\theta_2 > 0$. Then $(X_{(1)} + X_{(n)})/2$ is the UMVUE when $\mathcal{P}_2$ is considered, where $X_{(j)}$ is the jth order statistic.
- Then $\bar{X} = (X_{(1)} + X_{(n)})/2$ a.s. P for any $P \in \mathcal{P}_2$, which is impossible if $n > 2$. Hence, there is no UMVUE of μ.

□

4.2 How to Find UMVUE?

- First method: solving equations for h
  - Find a sufficient and complete statistic T and its distribution.
  - Try some function h to see if $\mathbb{E}h(T)$ is related to θ.
  - Solve for h such that $\mathbb{E}h(T) = \theta$ for all P

Example 4.5. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$. The order statistic $X(n)$ is sufficient and complete with Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$. Consider $\eta = g(\theta)$, where g is a differentiable function on $(0, \infty)$.

- An unbiased estimator $h(X_{(n)})$ of η must satisfy

$$\theta^n g(\theta) = n \int_0^\theta h(x) x^{n-1} dx \qquad \text{for all } \theta > 0.$$

- Differentiating both sizes of the previous equation and applying the result of differentiation of an integral lead to

$$n\theta^{n-1} g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

- Hence, the UMVUE of η is

$$h(X_{(n)}) = g(X_{(n)}) + n^{-1} X_{(n)} g'(X_{(n)}).$$

- In particular, if $\eta = \theta$, then the UMVUE of θ is $(1 + n^{-1})X_{(n)}$.

Example 4.6. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$.

- $T(X) = \sum_{i=1}^n X_i$ is sufficient and complete for $\theta > 0$ and has the Poisson distribution $P(n\theta)$.
- Suppose that $\eta = g(\theta)$, where g is a smooth function such that $g(x) = \sum_{j=0}^\infty a_j x^j$, $x > 0$.

- An unbiased estimator $h(T)$ of η must satisfy (for any $\theta > 0$):

$$\begin{aligned}\sum_{t=0}^{\infty} \frac{h(t)n^t}{t!}\theta^t &= e^{n\theta}g(\theta) \\ &= \sum_{k=0}^{\infty} \frac{n^k}{k!}\theta^k \sum_{j=0}^{\infty} a_j\theta^j \\ &= \sum_{t=0}^{\infty} \left(\sum_{j,k:j+k=t} \frac{n^k a_j}{k!} \right) \theta^t.\end{aligned}$$

- A comparison of coefficients in front of θ^t leads to

$$h(t) = \frac{t!}{n^t} \sum_{j,k:j+k=t} \frac{n^k a_j}{k!},$$

  i.e., $h(T)$ is the UMVUE of η.

- In particular, if $\eta = \theta^r$ for some fixed integer $r \geq 1$, then $a_r = 1$ and $a_k = 0$ if $k \neq r$ and

$$h(t) = \begin{cases} 0 & t < r \\ \frac{t!}{n^r(t-r)!} & t \geq r \end{cases}$$

- Second approach
  - Find an unbiased estimator of θ, say $U(X)$.
  - Conditioning on a sufficient and complete statistic $T(X)$: $\mathbb{E}(U \mid T)$ is the UMVUE of θ.
  - The distribution of T is not needed. We only need to work out the conditional expectation $\mathbb{E}(U \mid T)$.
  - From the uniqueness of the UMVUE, it does not matter which $U(X)$ is used.
  - Thus, $U(X)$ should be chosen so as to make the calculation of $\mathbb{E}(U \mid T)$ as easy as possible.

Example 4.7. Let $X_1, \ldots, X_n$ be IID sampled from Exp(θ). Let F_θ be the CDF and $t > 0$. Find a UMVUE for the tail probability $p = 1 - F_\theta(t)$.

- $\overline{X} = n^{-1}\sum_{i=1}^n$ is sufficient and complete, since Exp(θ) is an exponential family of full rank
- $I_{(t,\infty)}(X_1)$ is unbiased: $\mathbb{E}I_{(t,\infty)}(X_1) = P_\theta(X_1 > t) = 1 - F_\theta(t)$.
- $\mathbb{E}(I_{(t,\infty)}(X_1) \mid \overline{X})$ is a UMVUE
- Note that the distribution of $X_1/\overline{X}$ does not depend on θ – ancillary for θ
- By Basu's theorem, $X_1/\overline{X}$ and $\overline{X}$ are independent

$$\begin{aligned}P_\theta(X_1 > t \mid \overline{X} = \bar{x}) &= P_\theta(X_1/\overline{X} > t/\overline{X} \mid \overline{X} = \bar{x}) \\ &= P(X_1/\overline{X} > t/\bar{x}).\end{aligned}$$

- We need the unconditional distribution of $X_1/\overline{X}$.

$$P(X_1/\overline{X} > t/\bar{x}) = P\left(\frac{X_1}{\sum_{i=1}^n X_i} > \frac{t}{n\bar{x}}\right)$$

- $X_1 \sim \text{gamma}(k = 1, \theta)$
- $\sum_{i=2}^n X_i \sim \text{gamma}(k = n-1, \theta)$
- $\frac{X_1}{X_1 + \sum_{i=2}^n X_i} \sim \text{beta}(1, n-1)$, the PDF is $(n-1)(1-x)^{n-2} I_{(0,1)}(x)$
  - the PDF of beta(a, b) is $x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)/B(a,b)$, where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(a)$ is the gamma function
  - $\Gamma(n) = (n-1)!$
- Thus

$$P\left\{\frac{X_1}{\sum_{i=1}^n X_i} > \frac{t}{n\bar{x}}\right\} = (n-1)\int_{t/(n\bar{x})}^1 (1-x)^{n-2}\mathrm{d}x = \left(1 - \frac{t}{n\bar{x}}\right)^{n-1}$$

- So the UMVUE is

$$T(X) = \left(1 - \frac{t}{n\overline{X}}\right)^{n-1}$$

Example 4.8. Let $X_1, ..., X_n$ be i.i.d. with Lebesgue p.d.f. $f_\theta(x) = \theta x^{-2} I_{(\theta,\infty)}(x)$, where $\theta > 0$ is unknown. Suppose that $\eta = P(X_1 > t)$ for a constant $t > 0$. Find a UMVUE of η

- The smallest order statistic $X_{(1)}$ is sufficient and complete for θ.
- Hence, the UMVUE of η is

$$\begin{aligned} P(X_1 > t|X_{(1)}) &= P(X_1 > t|X_{(1)} = x_{(1)}) \\ &= P\left(\frac{X_1}{X_{(1)}} > \frac{t}{X_{(1)}}\middle| X_{(1)} = x_{(1)}\right) \\ &= P\left(\frac{X_1}{X_{(1)}} > \frac{t}{x_{(1)}}\middle| X_{(1)} = x_{(1)}\right) \\ &= P\left(\frac{X_1}{X_{(1)}} > s\right) \end{aligned}$$

  (Basu's theorem), where $s = t/x_{(1)}$.
- If $s \leq 1$, this probability is 1.

- Consider $s > 1$ and assume $\theta = 1$ in the calculation:

$$\begin{aligned}
P\left(\frac{X_1}{X_{(1)}} > s\right) &= \sum_{i=1}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= \sum_{i=2}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= (n-1)P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_n\right) \\
&= (n-1)P\left(X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n\right) \\
&= (n-1)\int_{x_1 > sx_n, x_2 > x_n, \ldots, x_{n-1} > x_n} \prod_{i=1}^{n} \frac{1}{x_i^2} dx_1 \cdots dx_n \\
&= (n-1)\int_1^\infty \left[\int_{sx_n}^\infty \prod_{i=2}^{n-1} \left(\int_{x_n}^\infty \frac{1}{x_i^2} dx_i\right) \frac{1}{x_1^2} dx_1\right] \frac{1}{x_n^2} dx_n \\
&= (n-1)\int_1^\infty \frac{1}{s x_n^{n+1}} dx_n = \frac{(n-1)x_{(1)}}{nt}
\end{aligned}$$

- This shows that the UMVUE of $P(X_1 > t)$ is

$$h(X_{(1)}) = \begin{cases} \frac{(n-1)X_{(1)}}{nt} & X_{(1)} < t \\ 1 & X_{(1)} \geq t \end{cases}$$

4.3 A Necessary and Sufficient Condition for UMVUE

Theorem 4.9. *Let $\mathcal{U}$ be the set of all unbiased estimators of 0 with finite variances and T be an unbiased estimator of η with $\mathbb{E}(T^2) < \infty$.*

(i) A necessary and sufficient condition for $T(X)$ to be a UMVUE of η is that $\mathbb{E}[T(X)U(X)] = 0$ for any $U \in \mathcal{U}$ and any $P \in \mathcal{P}$.

(ii) Suppose that $T = h(\tilde{T})$, where $\tilde{T}$ is a sufficient statistic for $P \in \mathcal{P}$ and h is a Borel function. Let $\mathcal{U}_{\tilde{T}}$ be the subset of $\mathcal{U}$ consisting of Borel functions of $\tilde{T}$. Then a necessary and sufficient condition for T to be a UMVUE of η is that $\mathbb{E}[T(X)U(X)] = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and any $P \in \mathcal{P}$.

- Use of this theorem:
  - find a UMVUE
  - check whether a particular estimator is a UMVUE
  - show the nonexistence of any UMVUE

Proof: (i) Suppose that T is a UMVUE of η. Then $T_c = T + cU$, where $U \in \mathcal{U}$ and c is a fixed constant, is also unbiased for η and, thus,

$$\mathrm{Var}(T_c) \geq \mathrm{Var}(T) \qquad c \in \mathcal{R},\ P \in \mathcal{P},$$

which is the same as

$$c^2\mathrm{Var}(U) + 2c\mathrm{Cov}(T,U) \geq 0 \qquad c \in \mathcal{R},\ P \in \mathcal{P}.$$

This is impossible unless $\mathrm{Cov}(T,U) = \mathbb{E}(TU) = 0$ for any $P \in \mathcal{P}$.

Suppose now $\mathbb{E}(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$. Let T_0 be another unbiased estimator of η with $\mathrm{Var}(T_0) < \infty$. Then $T - T_0 \in \mathcal{U}$ and, hence,

$$\mathbb{E}[T(T - T_0)] = 0 \qquad P \in \mathcal{P},$$

which with the fact that $\mathbb{E}T = \mathbb{E}T_0$ implies that

$$\mathrm{Var}(T) = \mathrm{Cov}(T, T_0) \qquad P \in \mathcal{P}.$$

Note that $[\mathrm{Cov}(T,T_0)]^2 \leq \mathrm{Var}(T)\mathrm{Var}(T_0)$. Hence $\mathrm{Var}(T) \leq \mathrm{Var}(T_0)$ for any $P \in \mathcal{P}$.

(ii) It suffices to show that $\mathbb{E}(TU) = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and $P \in \mathcal{P}$ implies that $\mathbb{E}(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$ Let $U \in \mathcal{U}$. Then $\mathbb{E}(U|\tilde{T}) \in \mathcal{U}_{\tilde{T}}$ and the result follows from the fact that $T = h(\tilde{T})$ and

$$\mathbb{E}(TU) = \mathbb{E}[\mathbb{E}(TU|\tilde{T})] = \mathbb{E}[\mathbb{E}(h(\tilde{T})U|\tilde{T})] = \mathbb{E}[h(\tilde{T})\mathbb{E}(U|\tilde{T})].$$

Corollary 4.10. *(i) Let T_j be a UMVUE of η_j, $j = 1, ..., k$, where k is a fixed positive integer. Then $\sum_{j=1}^k c_j T_j$ is a UMVUE of $\eta = \sum_{j=1}^k c_j \eta_j$ for any constants $c_1, ..., c_k$.*

(ii) Let T_1 and T_2 be two UMVUE's of η. Then $T_1 = T_2$ a.s. P for any $P \in \mathcal{P}$.

Example 4.11. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on the interval $(0, \theta)$. In Example 4.5 we have shown that $(1 + n^{-1})X_{(n)}$ is the UMVUE for θ when the parameter space is $\Theta = (0, \infty)$. Suppose now that $\Theta = [1, \infty)$. Then $X_{(n)}$ is not complete, although it is still sufficient for θ. Thus, Lehmann-Scheffé theorem does not apply to $X_{(n)}$.

We now use Theorem 4.9(ii) to find a UMVUE of θ. Let $U(X_{(n)})$ be an unbiased estimator of 0. Since $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$,

$$0 = \int_0^1 U(x)x^{n-1}dx + \int_1^\theta U(x)x^{n-1}dx \quad \text{for all } \theta \geq 1.$$

This implies that $U(x) = 0$ a.e. Lebesgue measure on $[1, \infty)$ and

$$\int_0^1 U(x)x^{n-1}dx = 0.$$

Consider $T = h(X_{(n)})$. To have $\mathbb{E}(TU) = 0$, we must have

$$\int_0^1 h(x)U(x)x^{n-1}dx = 0.$$

Thus, we may consider the following function:

$$h(x) = \begin{cases} c & 0 \le x \le 1 \\ bx & x > 1, \end{cases}$$

where c and b are some constants.

From the previous discussion,

$$\mathbb{E}[h(X_{(n)})U(X_{(n)})] = 0, \qquad \theta \ge 1.$$

Since $\mathbb{E}[h(X_{(n)})] = \theta$, we obtain that

$$\begin{aligned} \theta &= cP(X_{(n)} \le 1) + b\mathbb{E}[X_{(n)} I_{(1,\infty)}(X_{(n)})] \\ &= c\theta^{-n} + [bn/(n+1)](\theta - \theta^{-n}). \end{aligned}$$

Thus, $c = 1$ and $b = (n+1)/n$. The UMVUE of θ is then

$$h(X_{(n)}) = \begin{cases} 1 & 0 \le X_{(n)} \le 1 \\ (1 + n^{-1})X_{(n)} & X_{(n)} > 1. \end{cases}$$

This estimator is better than $(1 + n^{-1})X_{(n)}$, which is the UMVUE when $\Theta = (0, \infty)$ and does not make use of the information about $\theta \ge 1$. When $\Theta = (0, \infty)$, this estimator is not unbiased.

In fact, $h(X_{(n)})$ is complete and sufficient for $\theta \in [1, \infty)$. It suffices to show that

$$g(X_{(n)}) = \begin{cases} 1 & 0 \le X_{(n)} \le 1 \\ X_{(n)} & X_{(n)} > 1. \end{cases}$$

is complete and sufficient for $\theta \in [1, \infty)$. The sufficiency follows from the fact that the joint p.d.f. of $X_1, ..., X_n$ is

$$\frac{1}{\theta^n} I_{(0,\theta)}(X_{(n)}) = \frac{1}{\theta^n} I_{(0,\theta)}(g(X_{(n)})).$$

If $\mathbb{E}[f(g(X_{(n)})] = 0$ for all $\theta > 1$, then

$$0 = \int_0^\theta f(g(x)) x^{n-1} dx = \int_0^1 f(1) x^{n-1} dx + \int_1^\theta f(x) x^{n-1} dx$$

for all $\theta > 1$. Letting $\theta \to 1$ we obtain that $f(1) = 0$. Then

$$0 = \int_1^\theta f(x) x^{n-1} dx$$

for all $\theta > 1$, which implies $f(x) = 0$ a.e. for $x > 1$. Hence, $g(X_{(n)})$ is complete.

Lecture 4: UMVUE

Lecturer: LIN Zhenhua ST5215 AY2019/2020 Semester I

4.1 UMVUE

- For squred error loss, the risk of an unbiased estimator is equal to its variance
- We can compare unbiased estimators by their variance

Definition 4.1 (UMVUE)**.** An unbiased estimator $T(X)$ of θ is called the uniformly minimum variance unbiased estimator (UMVUE) if and only if $\mathrm{Var}(T(X)) \leq \mathrm{Var}(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of θ.

- "uniformly" refers to "for any $P \in \mathcal{P}$"
- $R_T(\theta) \leq R_U(\theta)$ under squared error loss
- a UMVUE estimator is $\mathfrak{J}$-optimal in MSE with $\mathfrak{J}$ being the class of all unbiased estimators.

Theorem 4.2 (Lehmann-Scheffé)**.** *Suppose that there exists a sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$. If there exists an unbiased estimator for θ, then there is a unique unbiased estimator of θ that is of the form $h(T)$ with a Borel function h. Furthermore, $h(T)$ is the unique UMVUE of θ.*

Proof:

- By assumption, there is an unbiased estimator $\hat{\theta}$ for θ.
- Let $h(T) = \mathbb{E}(\hat{\theta} \mid T)$. Then $\mathbb{E}h(T) = \mathbb{E}\hat{\theta} = \theta$.
  - $h(T)$ is unbiased for θ
- Suppose $g(T)$ is another unbiased estimator of θ
- $\mathbb{E}\{h(T) - g(T)\} = 0$
- The completeness of T implies that $h - g = 0$ $\mathcal{P}$-a.s.
- The squared error loss is strictly convex. So any admissible estimator must be of the form $h(T)$, by Rao-Blackwell theorem
- $h(T)$ is the only (possible) admissible unbiased estimator, and thus UMVUE.

Example 4.3. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$. In previous lectures, we have shown that the order statistic $X_{(n)}$ is sufficient and complete with Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$.

- We observe that
$$\mathbb{E}X_{(n)} = n\theta^{-n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$
- $\mathbb{E}\{(n+1)X_{(n)}/n\} = \theta$
- By Lehmann-Scheffé theorem, $\hat{\theta} = (n+1)X_{(n)}/n$ is the UMVUE for θ

Example 4.4. Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathcal{P}$.

- In many cases, the vector of order statistics, $T = (X_{(1)}, ..., X_{(n)})$, is sufficient and complete for $P \in \mathcal{P}$. (For example, $\mathcal{P}$ is the collection of all Lebesgue p.d.f.'s.)
- An estimator $\varphi(X_1, ..., X_n)$ is a function of T iff the function φ is symmetric in its n arguments.
- A symmetric unbiased estimator $h(T(X))$ of any estimable θ is the UMVUE. This is because, due to symmetry, $h(T(X))$ is a function of the order statistic T, and T is sufficient and complete for many nonparametric families.

The following are some examples:

- $\bar{X}$ is the UMVUE of $\theta = \mathbb{E}X_1$;
- S^2 is the UMVUE of $\text{Var}(X_1)$;
- $n^{-1}\sum_{i=1}^n X_i^2 - S^2$ is the UMVUE of $(\mathbb{E}X_1)^2$;
- $F_n(t)$ is the UMVUE of $P(X_1 \le t)$ for any fixed t.

The previous conclusions are not true if T is *not* sufficient and complete for $P \in \mathcal{P}$. For example, if $n > 2$ and $\mathcal{P}$ contains all symmetric distributions having Lebesgue p.d.f.'s and finite means, then below we show that there is no UMVUE for $\mu = EX_1$.

Proof.

- Suppose that T is a UMVUE of μ.
- Let $\mathcal{P}_1 = \{N(\mu, 1) : \mu \in \mathcal{R}\}$. Since the sample mean $\bar{X}$ is UMVUE when $\mathcal{P}_1$ is considered, and the Lebesgue measure is dominated by any $P \in \mathcal{P}_1$, we conclude that $T = \bar{X}$ a.e. Lebesgue measure.

- Let $\mathcal{P}_2$ be the family of uniform distributions on $(\theta_1 - \theta_2, \theta_1 + \theta_2)$, $\theta_1 \in \mathcal{R}$, $\theta_2 > 0$. Then $(X_{(1)} + X_{(n)})/2$ is the UMVUE when $\mathcal{P}_2$ is considered, where $X_{(j)}$ is the jth order statistic.
- Then $\bar{X} = (X_{(1)} + X_{(n)})/2$ a.s. P for any $P \in \mathcal{P}_2$, which is impossible if $n > 2$. Hence, there is no UMVUE of μ.

□

4.2 How to Find UMVUE?

- First method: solving equations for h
  - Find a sufficient and complete statistic T and its distribution.
  - Try some function h to see if $\mathbb{E}h(T)$ is related to θ.
  - Solve for h such that $\mathbb{E}h(T) = \theta$ for all P

Example 4.5. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$. The order statistic $X(n)$ is sufficient and complete with Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$. Consider $\eta = g(\theta)$, where g is a differentiable function on $(0, \infty)$.

- An unbiased estimator $h(X_{(n)})$ of η must satisfy

$$\theta^n g(\theta) = n \int_0^\theta h(x) x^{n-1} dx \qquad \text{for all } \theta > 0.$$

- Differentiating both sizes of the previous equation and applying the result of differentiation of an integral lead to

$$n\theta^{n-1} g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

- Hence, the UMVUE of η is

$$h(X_{(n)}) = g(X_{(n)}) + n^{-1} X_{(n)} g'(X_{(n)}).$$

- In particular, if $\eta = \theta$, then the UMVUE of θ is $(1 + n^{-1})X_{(n)}$.

Example 4.6. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$.

- $T(X) = \sum_{i=1}^n X_i$ is sufficient and complete for $\theta > 0$ and has the Poisson distribution $P(n\theta)$.
- Suppose that $\eta = g(\theta)$, where g is a smooth function such that $g(x) = \sum_{j=0}^\infty a_j x^j$, $x > 0$.

- An unbiased estimator $h(T)$ of η must satisfy (for any $\theta > 0$):

$$\begin{aligned}\sum_{t=0}^{\infty} \frac{h(t)n^t}{t!}\theta^t &= e^{n\theta}g(\theta) \\ &= \sum_{k=0}^{\infty} \frac{n^k}{k!}\theta^k \sum_{j=0}^{\infty} a_j\theta^j \\ &= \sum_{t=0}^{\infty} \left(\sum_{j,k:j+k=t} \frac{n^k a_j}{k!} \right) \theta^t.\end{aligned}$$

- A comparison of coefficients in front of θ^t leads to

$$h(t) = \frac{t!}{n^t} \sum_{j,k:j+k=t} \frac{n^k a_j}{k!},$$

  i.e., $h(T)$ is the UMVUE of η.

- In particular, if $\eta = \theta^r$ for some fixed integer $r \geq 1$, then $a_r = 1$ and $a_k = 0$ if $k \neq r$ and

$$h(t) = \begin{cases} 0 & t < r \\ \frac{t!}{n^r(t-r)!} & t \geq r \end{cases}$$

- Second approach
  - Find an unbiased estimator of θ, say $U(X)$.
  - Conditioning on a sufficient and complete statistic $T(X)$: $\mathbb{E}(U \mid T)$ is the UMVUE of θ.
  - The distribution of T is not needed. We only need to work out the conditional expectation $\mathbb{E}(U \mid T)$.
  - From the uniqueness of the UMVUE, it does not matter which $U(X)$ is used.
  - Thus, $U(X)$ should be chosen so as to make the calculation of $\mathbb{E}(U \mid T)$ as easy as possible.

Example 4.7. Let $X_1, \ldots, X_n$ be IID sampled from Exp(θ). Let F_θ be the CDF and $t > 0$. Find a UMVUE for the tail probability $p = 1 - F_\theta(t)$.

- $\overline{X} = n^{-1}\sum_{i=1}^n$ is sufficient and complete, since Exp(θ) is an exponential family of full rank
- $I_{(t,\infty)}(X_1)$ is unbiased: $\mathbb{E}I_{(t,\infty)}(X_1) = P_\theta(X_1 > t) = 1 - F_\theta(t)$.
- $\mathbb{E}(I_{(t,\infty)}(X_1) \mid \overline{X})$ is a UMVUE
- Note that the distribution of $X_1/\overline{X}$ does not depend on θ – ancillary for θ
- By Basu's theorem, $X_1/\overline{X}$ and $\overline{X}$ are independent

$$\begin{aligned}P_\theta(X_1 > t \mid \overline{X} = \bar{x}) &= P_\theta(X_1/\overline{X} > t/\overline{X} \mid \overline{X} = \bar{x}) \\ &= P(X_1/\overline{X} > t/\bar{x}).\end{aligned}$$

- We need the unconditional distribution of $X_1/\overline{X}$.

$$P(X_1/\overline{X} > t/\bar{x}) = P\left(\frac{X_1}{\sum_{i=1}^n X_i} > \frac{t}{n\bar{x}}\right)$$

- $X_1 \sim \text{gamma}(k = 1, \theta)$
- $\sum_{i=2}^n X_i \sim \text{gamma}(k = n-1, \theta)$
- $\frac{X_1}{X_1 + \sum_{i=2}^n X_i} \sim \text{beta}(1, n-1)$, the PDF is $(n-1)(1-x)^{n-2} I_{(0,1)}(x)$
  - the PDF of beta(a, b) is $x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)/B(a,b)$, where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ and $\Gamma(a)$ is the gamma function
  - $\Gamma(n) = (n-1)!$
- Thus

$$P\left\{\frac{X_1}{\sum_{i=1}^n X_i} > \frac{t}{n\bar{x}}\right\} = (n-1)\int_{t/(n\bar{x})}^1 (1-x)^{n-2}\mathrm{d}x = \left(1 - \frac{t}{n\bar{x}}\right)^{n-1}$$

- So the UMVUE is

$$T(X) = \left(1 - \frac{t}{n\overline{X}}\right)^{n-1}$$

Example 4.8. Let $X_1, ..., X_n$ be i.i.d. with Lebesgue p.d.f. $f_\theta(x) = \theta x^{-2} I_{(\theta,\infty)}(x)$, where $\theta > 0$ is unknown. Suppose that $\eta = P(X_1 > t)$ for a constant $t > 0$. Find a UMVUE of η

- The smallest order statistic $X_{(1)}$ is sufficient and complete for θ.
- Hence, the UMVUE of η is

$$\begin{aligned} P(X_1 > t|X_{(1)}) &= P(X_1 > t|X_{(1)} = x_{(1)}) \\ &= P\left(\frac{X_1}{X_{(1)}} > \frac{t}{X_{(1)}}\middle| X_{(1)} = x_{(1)}\right) \\ &= P\left(\frac{X_1}{X_{(1)}} > \frac{t}{x_{(1)}}\middle| X_{(1)} = x_{(1)}\right) \\ &= P\left(\frac{X_1}{X_{(1)}} > s\right) \end{aligned}$$

  (Basu's theorem), where $s = t/x_{(1)}$.
- If $s \leq 1$, this probability is 1.

- Consider $s > 1$ and assume $\theta = 1$ in the calculation:

$$\begin{aligned}
P\left(\frac{X_1}{X_{(1)}} > s\right) &= \sum_{i=1}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= \sum_{i=2}^{n} P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= (n-1)P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_n\right) \\
&= (n-1)P\left(X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n\right) \\
&= (n-1)\int_{x_1 > sx_n, x_2 > x_n, \ldots, x_{n-1} > x_n} \prod_{i=1}^{n} \frac{1}{x_i^2} dx_1 \cdots dx_n \\
&= (n-1)\int_1^\infty \left[\int_{sx_n}^\infty \prod_{i=2}^{n-1} \left(\int_{x_n}^\infty \frac{1}{x_i^2} dx_i\right) \frac{1}{x_1^2} dx_1\right] \frac{1}{x_n^2} dx_n \\
&= (n-1)\int_1^\infty \frac{1}{s x_n^{n+1}} dx_n = \frac{(n-1)x_{(1)}}{nt}
\end{aligned}$$

- This shows that the UMVUE of $P(X_1 > t)$ is

$$h(X_{(1)}) = \begin{cases} \frac{(n-1)X_{(1)}}{nt} & X_{(1)} < t \\ 1 & X_{(1)} \geq t \end{cases}$$

4.3 A Necessary and Sufficient Condition for UMVUE

Theorem 4.9. *Let $\mathcal{U}$ be the set of all unbiased estimators of 0 with finite variances and T be an unbiased estimator of η with $\mathbb{E}(T^2) < \infty$.*

(i) A necessary and sufficient condition for $T(X)$ to be a UMVUE of η is that $\mathbb{E}[T(X)U(X)] = 0$ for any $U \in \mathcal{U}$ and any $P \in \mathcal{P}$.

(ii) Suppose that $T = h(\tilde{T})$, where $\tilde{T}$ is a sufficient statistic for $P \in \mathcal{P}$ and h is a Borel function. Let $\mathcal{U}_{\tilde{T}}$ be the subset of $\mathcal{U}$ consisting of Borel functions of $\tilde{T}$. Then a necessary and sufficient condition for T to be a UMVUE of η is that $\mathbb{E}[T(X)U(X)] = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and any $P \in \mathcal{P}$.

- Use of this theorem:
  - find a UMVUE
  - check whether a particular estimator is a UMVUE
  - show the nonexistence of any UMVUE

Proof: (i) Suppose that T is a UMVUE of η. Then $T_c = T + cU$, where $U \in \mathcal{U}$ and c is a fixed constant, is also unbiased for η and, thus,

$$\mathrm{Var}(T_c) \geq \mathrm{Var}(T) \qquad c \in \mathcal{R},\ P \in \mathcal{P},$$

which is the same as

$$c^2\mathrm{Var}(U) + 2c\mathrm{Cov}(T,U) \geq 0 \qquad c \in \mathcal{R},\ P \in \mathcal{P}.$$

This is impossible unless $\mathrm{Cov}(T,U) = \mathbb{E}(TU) = 0$ for any $P \in \mathcal{P}$.

Suppose now $\mathbb{E}(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$. Let T_0 be another unbiased estimator of η with $\mathrm{Var}(T_0) < \infty$. Then $T - T_0 \in \mathcal{U}$ and, hence,

$$\mathbb{E}[T(T - T_0)] = 0 \qquad P \in \mathcal{P},$$

which with the fact that $\mathbb{E}T = \mathbb{E}T_0$ implies that

$$\mathrm{Var}(T) = \mathrm{Cov}(T, T_0) \qquad P \in \mathcal{P}.$$

Note that $[\mathrm{Cov}(T,T_0)]^2 \leq \mathrm{Var}(T)\mathrm{Var}(T_0)$. Hence $\mathrm{Var}(T) \leq \mathrm{Var}(T_0)$ for any $P \in \mathcal{P}$.

(ii) It suffices to show that $\mathbb{E}(TU) = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and $P \in \mathcal{P}$ implies that $\mathbb{E}(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$ Let $U \in \mathcal{U}$. Then $\mathbb{E}(U|\tilde{T}) \in \mathcal{U}_{\tilde{T}}$ and the result follows from the fact that $T = h(\tilde{T})$ and

$$\mathbb{E}(TU) = \mathbb{E}[\mathbb{E}(TU|\tilde{T})] = \mathbb{E}[\mathbb{E}(h(\tilde{T})U|\tilde{T})] = \mathbb{E}[h(\tilde{T})\mathbb{E}(U|\tilde{T})].$$

Corollary 4.10. *(i) Let T_j be a UMVUE of η_j, $j = 1, ..., k$, where k is a fixed positive integer. Then $\sum_{j=1}^k c_j T_j$ is a UMVUE of $\eta = \sum_{j=1}^k c_j \eta_j$ for any constants $c_1, ..., c_k$.*

(ii) Let T_1 and T_2 be two UMVUE's of η. Then $T_1 = T_2$ a.s. P for any $P \in \mathcal{P}$.

Example 4.11. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on the interval $(0, \theta)$. We have shown that $(1 + n^{-1})X_{(n)}$ is the UMVUE for θ when the parameter space is $\Theta = (0, \infty)$. Suppose now that $\Theta = [1, \infty)$. Then $X_{(n)}$ is not complete, although it is still sufficient for θ. Thus, Lehmann-Scheffé theorem does not apply to $X_{(n)}$.

We now use Theorem 4.9(ii) to find a UMVUE of θ. Let $U(X_{(n)})$ be an unbiased estimator of 0. Since $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$,

$$0 = \int_0^1 U(x)x^{n-1}dx + \int_1^\theta U(x)x^{n-1}dx \quad \text{for all } \theta \geq 1.$$

This implies that $U(x) = 0$ a.e. Lebesgue measure on $[1, \infty)$ and

$$\int_0^1 U(x)x^{n-1}dx = 0.$$

Consider $T = h(X_{(n)})$. To have $\mathbb{E}(TU) = 0$, we must have

$$\int_0^1 h(x)U(x)x^{n-1}dx = 0.$$

Thus, we may consider the following function:

$$h(x) = \begin{cases} c & 0 \le x \le 1 \\ bx & x > 1, \end{cases}$$

where c and b are some constants.

From the previous discussion,

$$\mathbb{E}[h(X_{(n)})U(X_{(n)})] = 0, \qquad \theta \ge 1.$$

Since $\mathbb{E}[h(X_{(n)})] = \theta$, we obtain that

$$\begin{aligned} \theta &= cP(X_{(n)} \le 1) + b\mathbb{E}[X_{(n)}I_{(1,\infty)}(X_{(n)})] \\ &= c\theta^{-n} + [bn/(n+1)](\theta - \theta^{-n}). \end{aligned}$$

Thus, $c = 1$ and $b = (n+1)/n$. The UMVUE of θ is then

$$h(X_{(n)}) = \begin{cases} 1 & 0 \le X_{(n)} \le 1 \\ (1+n^{-1})X_{(n)} & X_{(n)} > 1. \end{cases}$$

This estimator is better than $(1+n^{-1})X_{(n)}$, which is the UMVUE when $\Theta = (0,\infty)$ and does not make use of the information about $\theta \ge 1$. When $\Theta = (0,\infty)$, this estimator is not unbiased.

In fact, $h(X_{(n)})$ is complete and sufficient for $\theta \in [1,\infty)$. It suffices to show that

$$g(X_{(n)}) = \begin{cases} 1 & 0 \le X_{(n)} \le 1 \\ X_{(n)} & X_{(n)} > 1. \end{cases}$$

is complete and sufficient for $\theta \in [1,\infty)$. The sufficiency follows from the fact that the joint p.d.f. of $X_1, ..., X_n$ is

$$\frac{1}{\theta^n} I_{(0,\theta)}(X_{(n)}) = \frac{1}{\theta^n} I_{(0,\theta)}(g(X_{(n)})).$$

If $\mathbb{E}[f(g(X_{(n)})] = 0$ for all $\theta > 1$, then

$$0 = \int_0^\theta f(g(x))x^{n-1}dx = \int_0^1 f(1)x^{n-1}dx + \int_1^\theta f(x)x^{n-1}dx$$

for all $\theta > 1$. Letting $\theta \to 1$ we obtain that $f(1) = 0$. Then

$$0 = \int_1^\theta f(x)x^{n-1}dx$$

for all $\theta > 1$, which implies $f(x) = 0$ a.e. for $x > 1$. Hence, $g(X_{(n)})$ is complete.

Example 4.12. Let X be a sample (of size 1) from the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$. There is no UMVUE of $\eta = g(\theta)$ for any nonconstant function g. Note that an unbiased estimator $U(X)$ of 0 must satisfy

$$\int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} U(x)dx = 0 \qquad \text{for all } \theta \in \mathcal{R}.$$

Differentiating both sides of the previous equation and applying the result of differentiation of an integral lead to

$$U(x) = U(x+1) \quad \text{a.e. } m,$$

where m is the Lebesgue measure on $\mathcal{R}$.

If T is a UMVUE of $g(\theta)$, then $T(X)U(X)$ is unbiased for 0 and, hence,

$$T(x)U(x) = T(x+1)U(x+1) \quad \text{a.e. } m,$$

where $U(X)$ is any unbiased estimator of 0.

Since this is true for all U,

$$T(x) = T(x+1) \quad \text{a.e. } m.$$

Since T is unbiased for $g(\theta)$,

$$g(\theta) = \int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} T(x)dx \qquad \text{for all } \theta \in \mathcal{R}.$$

Differentiating both sides of the previous equation and applying the result of differentiation of an integral, we obtain that

$$g'(\theta) = T\left(\theta+\tfrac{1}{2}\right) - T\left(\theta-\tfrac{1}{2}\right) = 0 \quad \text{a.e. } m.$$

4.4 Information Inequality

- What is the lower bound of the variance of an unbiased estimator?

For certain distribution families, a quantity called *Fisher information*, which measures the amount of information about an unknown parameter contained in the data, can be defined.

The families must satisfy the following regularity conditions:

(i) The family has p.d.f.s, i.e., $\mathcal{P} = \{p(x,\theta) : \theta \in \Theta\}$ where $p(x,\theta)$ is a p.d.f.

(ii) The set $\mathcal{A} = \{x : p(x,\theta) > 0\}$ does not depend on θ.

(iii) For all $x \in \mathcal{A}$ and $\theta \in \Theta$, $\frac{\partial p(x,\theta)}{\partial \theta}$ exists and is finite.

(iv) If T is any statistic such that $E_\theta|T| < \infty$ for all $\theta \in \Theta$, then

$$\frac{\partial}{\partial \theta}\int T(x)p(x,\theta)dx = \int T(x)\frac{\partial p(x,\theta)}{\partial \theta}dx,$$

whenever the right hand side is finite.

- Examples:

- The uniform distributions $\mathcal{U}(0,\theta)$ and the exponential distributions $\mathcal{E}(a,\theta)$ with unknown a and θ do not satisfy the regularity conditions.
- Any exponential family satisfies the conditions. In particular, the Normal, Gamma, Beta, Binormail, Poisson distributions, etc. satisfy the regularity conditions.

Let X be a single sample from $P \in \mathcal{P} = \{p(x,\theta) : \theta \in \Theta\}$, where Θ is an open set in $\mathcal{R}$. Suppose that the regularity conditions hold. The *Fisher information number* is defined as

$$\begin{aligned} I(\theta) &= \mathbb{E}\left(\frac{\partial}{\partial\theta}\ln p(X,\theta)\right)^2 \\ &= \int\left(\frac{\partial}{\partial\theta}\ln p(x,\theta)\right)^2 p(x,\theta)dx. \end{aligned}$$

- The greater $I(\theta)$ is, the easier it is to distinguish θ from neighboring values and, therefore, the more accurately θ can be estimated.
- $I(\theta)$ is a measure of the information that X contains about θ.
- Note that $I(\theta)$ depends on the particular parameterization.
- If $\theta = \psi(\eta)$ and ψ is differentiable, then the Fisher information that X contains about η is

  $$\left(\frac{\partial\psi(\eta)}{\partial\eta}\right)^2 I(\psi(\eta)),$$

  where $I(\psi(\eta))$ is the Fisher information number about θ.

- Extension to multi-parameter case

Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P} = \{p(x,\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, where Θ is an open set in $\mathcal{R}^k$. The $k \times k$ matrix

$$I(\theta) = \mathbb{E}\left\{\frac{\partial}{\partial\theta}\log f_\theta(X)\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^\top\right\}$$

is called the *Fisher information matrix*, where

$$\frac{\partial}{\partial\theta}\log f_\theta(X) = \left(\frac{\partial}{\partial\theta_1}\log f_\theta(X), \ldots, \frac{\partial}{\partial\theta_k}\log f_\theta(X)\right)^\top$$

- Properties of Fisher information number

  (i) If X and Y are independent with Fisher information numbers $I_X(\theta)$ and $I_Y(\theta)$, respectively, then $I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$. In particular, if $\mathbf{X} = (X_1, ..., X_n)$ where X_i's are i.i.d. and $I_1(\theta)$ is the Fisher information number of a single X_i, then $I_{\mathbf{X}}(\theta) = nI_1(\theta)$.

(ii) Suppose that the p.d.f. $p(x,\theta)$ is twice differentiable in θ and that

$$\frac{\partial}{\partial\theta}\int\frac{\partial p(x,\theta)}{\partial\theta^\top}dx=\int\frac{\partial}{\partial\theta}\frac{\partial p(x,\theta)}{\partial\theta^\top}dx,\qquad\theta\in\Theta.$$

Then

$$I(\theta)=-\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^\top}\log p(X,\theta)\right].$$

Example 4.13. Suppose $(X_1,\ldots,X_n)$ is a sample from a Poisson distribution $\mathcal{P}(\lambda)$. Then

$$\frac{\partial}{\partial\theta}\ln p(x,\lambda)=\frac{\sum_{i=1}^n x_i}{\lambda}-n$$
$$\text{and}\quad I(\lambda)=\mathrm{Var}\left(\frac{\sum_{i=1}^n x_i}{\lambda}\right)=\frac{n\lambda}{\lambda^2}=\frac{n}{\lambda}.$$

Example 4.14. Let $X_1,...,X_n$ be i.i.d. $\sim N(\mu,\nu)$. Let $\theta=(\mu,\nu)$. Then

$$\ln p(\mathbf{x},\theta)=-\frac{1}{2\nu}\sum_{i=1}^n(x_i-\mu)^2-\frac{n}{2}\ln\nu.$$

It can be calculated that

$$\frac{\partial^2}{\partial\mu^2}\ln p(\mathbf{x},\theta)=-\frac{n}{\nu},$$
$$\frac{\partial^2}{\partial\nu^2}\ln p(\mathbf{x},\theta)=-\frac{\sum_{i=1}^n(x_i-\mu)^2}{\nu^3}+\frac{n}{2\nu^2},$$
$$\frac{\partial^2}{\partial\nu\partial\mu}\ln p(\mathbf{x},\theta)=-\frac{\sum_{i=1}^n(x_i-\mu)}{\nu^2}.$$

Thus, the Fisher information matrix about θ contained in $X_1,...,X_n$ is

$$I(\theta)=\begin{pmatrix}\frac{n}{\nu}&0\\0&\frac{n}{2\nu^2}\end{pmatrix}.$$

Example 4.15. Let $X_1,...,X_n$ be i.i.d. with the Lebesgue p.d.f. $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$, where $f(x)>0$ and $f'(x)$ exists for all $x\in\mathcal{R}$, $\mu\in\mathcal{R}$, and $\sigma>0$ (a location-scale family). Let $\theta=(\mu,\sigma)$. Then, the Fisher information about θ contained in $X_1,...,X_n$ is (exercise)

$$I(\theta)=\frac{n}{\sigma^2}\begin{pmatrix}\int\frac{[f'(x)]^2}{f(x)}dx&\int\frac{f'(x)[xf'(x)+f(x)]}{f(x)}dx\\\int\frac{f'(x)[xf'(x)+f(x)]}{f(x)}dx&\int\frac{[xf'(x)+f(x)]^2}{f(x)}dx\end{pmatrix}.$$

Theorem 4.16 (Cramér-Rao lower bound)**.** *Suppose that $T(X)$ is an estimator with $\mathbb{E}[T(X)]=g(\theta)$ being a differentiable function of θ; P_θ has a p.d.f. f_θ w.r.t. a measure ν for all $\theta\in\Theta$; and f_θ is differentiable as a function of θ and satisfies*

$$\frac{\partial}{\partial\theta}\int h(x)f_\theta(x)d\nu=\int h(x)\frac{\partial}{\partial\theta}f_\theta(x)d\nu,\qquad\theta\in\Theta,\tag{4.10}$$

for $h(x)\equiv1$ and $h(x)=T(x)$. Then

$$\mathrm{Var}(T(X))\geq\left[\frac{\partial}{\partial\theta}g(\theta)\right]^\top[I(\theta)]^{-1}\frac{\partial}{\partial\theta}g(\theta).\tag{4.11}$$

- The inequality in (4.11) is called *information inequality*.
- If there is an unbiased estimator T of $g(\theta)$ whose variance is always the same as the lower bound, then T is the UMVUE.

Proof: We consider the case $k = 1$ only. When $k = 1$, (4.11) reduces to

$$\mathrm{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2}.$$

From the Cauchy-Schwartz inequality, we only need to show that

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2 = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right) \tag{4.12}$$

$$g'(\theta) = \mathrm{Cov}\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(X)\right). \tag{4.13}$$

From condition (4.10), we have

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int \frac{\partial}{\partial\theta} f_\theta(X)d\nu = \frac{\partial}{\partial\theta}\int f_\theta(X)d\nu = 0.$$

$$\mathbb{E}\left[T(X)\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int T(x)\frac{\partial}{\partial\theta} f_\theta(X)d\nu = \frac{\partial}{\partial\theta}\int T(x) f_\theta(X)d\nu = g'(\theta).$$

Thus (4.12) and (4.13) are verified.

- The general case follows from the inequality

$$\max_{\mathbf{c}} \frac{\left[\mathrm{Cov}(T, \mathbf{c}^\top \frac{\partial \ln f_\theta(X)}{\partial\theta})\right]^2}{\mathrm{Var}(T)\mathrm{Var}(\mathbf{c}^\top \frac{\partial \ln f_\theta(X)}{\partial\theta})} \leq 1.$$

  - The above inequality can be proved using the same idea for proving the univariate case of Cramér–Rao theorem.
  - But why this implies the multi-parameter case of the Cramér–Rao theorem?
  - The trick is to use $= [I(\theta)]^{-1}\frac{\partial}{\partial\theta}g(\theta)$.
  - Use $\mathbf{a} = \mathrm{Cov}(T, \frac{\partial}{\partial\theta}g(\theta))$ to simplify notations
  - Then the numerator is $(\mathbf{c}^\top\mathbf{a})^2$ while the denominator is $\mathrm{Var}(T)\mathbf{c}^\top I(\theta)\mathbf{c}$
  - Replace $\mathbf{c}$ with $[I(\theta)]^{-1}\mathbf{a}$
- The Cramér-Rao lower bound in (4.11) is not affected by any one-to-one reparameterization.
- If we use inequality (4.11) to find a UMVUE $T(X)$, then we obtain a formula for $\mathrm{Var}(T(X))$ at the same time.
- On the other hand, the Cramér-Rao lower bound in (4.11) is typically not sharp.

- Under some regularity conditions, the Cramér-Rao lower bound is attained iff f_θ is in an exponential family; see Propositions 3.2 and 3.3 and the discussion in Lehmann (1983, p. 123).

Example 4.17. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Let f_μ be the joint distribution of $X = (X_1, ..., X_n)$. Then

$$\frac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^{n} (X_i - \mu)/\sigma^2.$$

Thus, $I(\mu) = n/\sigma^2$.

Consider the estimation of μ. It is obvious that $\mathrm{Var}(\bar{X})$ attains the Cramér-Rao lower bound in (4.11).

Consider now the estimation of $\eta = \mu^2$. Since $\mathbb{E}\bar{X}^2 = \mu^2 + \sigma^2/n$, the UMVUE of η is $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$. A straightforward calculation shows that

$$\mathrm{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

On the other hand, the Cramér-Rao lower bound in this case is $4\mu^2\sigma^2/n$. Hence $\mathrm{Var}(h(\bar{X}))$ does not attain the Cramér-Rao lower bound. The difference is $2\sigma^4/n^2$.

Fisher information and exponential families

Proposition 4.18. *Suppose that the distribution of X is from an exponential family $\{f_\theta : \theta \in \Theta\}$, i.e., the p.d.f. of X w.r.t. a σ-finite measure is*

$$f_\theta(x) = \exp\big\{[\eta(\theta)]^\top T(x) - \xi(\theta)\big\} c(x), \tag{4.14}$$

where Θ is an open subset of $\mathcal{R}^k$.

(i) The regularity condition (4.10) is satisfied for any h with $\mathbb{E}|h(X)| < \infty$ and

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^\top} \log f_\theta(X)\right].$$

(ii) If $\underline{I}(\eta)$ is the Fisher information matrix for the natural parameter η, then the variance-covariance matrix $\mathrm{Var}(T) = \underline{I}(\eta)$.

(iii) If $\overline{I}(\psi)$ is the Fisher information matrix for the parameter $\psi = \mathbb{E}[T(X)]$, then $\mathrm{Var}(T) = [\overline{I}(\psi)]^{-1}$.

Proof:

(i) This is a direct consequence of Theorem 2.1 (of the textbook).

(ii) The p.d.f. under the natural parameter η is

$$f_\eta(x) = \exp\left\{\eta^\top T(x) - \zeta(\eta)\right\} c(x).$$

From Theorem 2.1 of (the textbook), $\mathbb{E}[T(X)] = \frac{\partial}{\partial \eta}\zeta(\eta)$. The result follows from

$$\frac{\partial}{\partial \eta} \log f_\eta(x) = T(x) - \frac{\partial}{\partial \eta}\zeta(\eta).$$

(iii) Since $\psi = \mathbb{E}[T(X)] = \frac{\partial}{\partial \eta}\zeta(\eta)$,

$$\underline{I}(\eta) = \frac{\partial \psi^\top}{\partial \eta} \overline{I}(\psi) \left(\frac{\partial \psi^\top}{\partial \eta}\right)^\top = \frac{\partial^2}{\partial \eta \partial \eta^\top}\zeta(\eta) \overline{I}(\psi) \left[\frac{\partial^2}{\partial \eta \partial \eta^\top}\zeta(\eta)\right]^\top.$$

By Theorem 2.1 (of the textbook) and the result in (ii),

$$\frac{\partial^2}{\partial \eta \partial \eta^\top}\zeta(\eta) = \mathrm{Var}(T) = \underline{I}(\eta).$$

Hence

$$\overline{I}(\psi) = [\underline{I}(\eta)]^{-1} \underline{I}(\eta) [\underline{I}(\eta)]^{-1} = [\underline{I}(\eta)]^{-1} = [\mathrm{Var}(T)]^{-1}.$$

- Condition (4.10) is a key regularity condition for the results in Cramér-Rao lower bound
- If f_θ is not in an exponential family, then (4.10) has to be checked.
- Typically, it does not hold if the set $\{x : f_\theta(x) > 0\}$ depends on θ

4.5 Asymptotic properties of UMVUE's

- Asymptotic normality

Let $T_n = T(\mathbf{X}_n)$ be an estimator based on a sample $\mathbf{X}_n$ of size n. Let $\mu_n(\theta)$ and $\sigma_n^2(\theta)$ be two sequences of constants which might depend on θ. If

$$\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \to_d N(0,1),$$

then T_n is said to be asymptotically normal with asymptotic mean and variance $\mu_n(\theta)$ and $\sigma_n^2(\theta)$ respectively.

- Asymptotic unbiasedness

Suppose that $T_n = T(\mathbf{X}_n)$ is asymptotically normal with asymptotic mean $\mu_n(\theta)$ and asymptotic variance $\sigma_n^2(\theta)$. If

$$\frac{\mu_n(\theta) - q(\theta)}{\sigma_n(\theta)} \to 0,$$

then T_n is said to be asymptotically unbiased for $q(\theta)$.

- Asymptotic relative efficiency

Let $T^{(1)} = \{T_n^{(1)}\}$ and $T^{(2)} = \{T_n^{(2)}\}$ be two sequences of estimators which are asymptotically unbiased for $q(\theta)$ and whose asymptotic variances σ_{n1}^2 and σ_{n2}^2 satisfy $n\sigma_{ni}^2 \to \sigma_i^2, i = 1, 2$. The asymptotic relative efficiency of $T^{(1)}$ to $T^{(2)}$ is defined by

$$e(\theta, T^{(1)}, T^{(2)}) = \frac{\sigma_2^2}{\sigma_1^2}.$$

- Asymptotic efficient estimator

Suppose that $T_n = T(\mathbf{X}_n)$ is asymptotically normal with asymptotic mean $\mu_n(\theta)$ and asymptotic variance $\sigma_n^2(\theta)$. If

$$n\sigma_n^2(\theta) \to \sigma^2(\theta) > 0, \ \sqrt{n}(\mu_n(\theta) - q(\theta)) \to 0,$$
$$\sigma^2(\theta) = \frac{[q'(\theta)]^2}{I_1(\theta)},$$

then T_n is said to be asymptotically efficient (or best asymptotically normal).

Example 4.19. Under the assumption of Hardy-Weinberg equilibrium,

$$p_1 = \theta^2, p_2 = 2\theta(1-\theta), p_3 = (1-\theta)^2.$$

Consider two estimators of θ:

$$T_1 = \sqrt{\frac{N_1}{n}}, \quad T_2 = 1 - \sqrt{\frac{N_3}{n}}.$$

- By the CLT,

$$\frac{\sqrt{n}(N_1/n - p_1)}{\sqrt{p_1(1-p_1)}} \to N(0,1) \quad \text{and} \quad \frac{\sqrt{n}(N_3/n - p_3)}{\sqrt{p_3(1-p_3)}} \to N(0,1).$$

- By the δ-method and slutsky's theorem,

$$\frac{\sqrt{n}(T_1 - \theta)}{\sqrt{\frac{1}{4}(1-\theta^2)}} \to N(0,1) \quad \text{and} \quad \frac{\sqrt{n}(T_2 - \theta)}{\sqrt{\frac{1}{4}[1-(1-\theta)^2]}} \to N(0,1).$$

- Both T_1 and T_2 are asymptotically normal and both are asymptotically unbiased.

- The asymptotic variances of T_1 and T_2 are, respectively, $\frac{1-\theta^2}{4n}$ and $\frac{1-(1-\theta)^2}{4n}$ and $\sigma_1^2 = \frac{1-\theta^2}{4}$, $\sigma_2^2 = \frac{1-(1-\theta)^2}{4}$.
- Asymptotic relative efficiency of T_1 to T_2 is

$$e(\theta, T_1, T_2) = \frac{1-(1-\theta)^2}{1-\theta^2}.$$

- It can be concluded that T_2 is better than T_1 for $\theta > 1/2$ and the two are equally efficient for $\theta = 1/2$. But none of the two is uniformly better than the other.

Consider another estimator of θ: $T_3 = \frac{N_1}{n} + \frac{N_2}{2n}$.

- By multiple CLT,

$$\sqrt{n}\begin{pmatrix} \frac{N_1}{n} - p_1 \\ \frac{N_2}{n} - p_2 \end{pmatrix} \to \begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mathbf{0}, \Sigma),$$

  where $\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 \\ -p_1p_2 & p_2(1-p_2) \end{pmatrix}$.
- By either the δ-method or continuous mapping theorem,

$$\begin{aligned} \sqrt{n}(T_3 - \theta) &= \sqrt{n}(\frac{N_1}{n} - p_1) + \frac{1}{2}\sqrt{n}(\frac{N_2}{n} - p_2) \\ &\to X + Y/2 \sim N(0, \sigma^2), \end{aligned}$$

  where

$$\sigma^2 = p_1(1-p_1) + \frac{1}{4}p_2(1-p_2) - p_1p_2 = \frac{\theta(1-\theta)}{2}.$$

- Hence, T_3 is asymptotically normal, asymptotically unbiased and its asymptotic variance is $\sigma_{n3}^2 = \frac{\theta(1-\theta)}{2n}$. Thus $n\sigma_{n3}^2 \to \sigma^2 = \frac{\theta(1-\theta)}{2}$.
- The asymptotic relative efficiency of T_3 to T_1 and T_2 are:

$$\begin{aligned} e(\theta, T_3, T_1) &= \frac{(1-\theta^2)/4}{\theta(1-\theta)/2} = \frac{1+\theta}{2\theta}, \\ e(\theta, T_3, T_2) &= \frac{\{1-(1-\theta)^2\}/4}{\theta(1-\theta)/2} = \frac{2-\theta}{2-2\theta}. \end{aligned}$$

- T_3 is uniformly better than both T_1 and T_2.
- Indeed, T_3 is asymptotically efficient, which is verified by showing that $I_n(\theta) = \frac{2n}{\theta(1-\theta)}$.
- The log p.d.f. of (N_1, N_2, N_3) is

$$\begin{aligned} \ln p(\mathbf{N}, \theta) &= N_1 \ln p_1 + N_2 \ln p_2 + N_3 \ln p_3 + C \\ &= (2N_1 + N_2)\ln\theta + (N_2 + 2N_3)\ln(1-\theta) + C, \end{aligned}$$

  where C is a generic constant.

- The derivatives of $\ln p(\mathbf{N}, \theta)$:

$$\begin{aligned} \frac{\partial}{\partial \theta} \ln p(\mathbf{N}, \theta) &= \frac{2N_1 + N_2}{\theta} - \frac{N_2 + 2N_3}{1-\theta} \\ \frac{\partial^2}{\partial \theta^2} \ln p(\mathbf{N}, \theta) &= -\frac{2N_1 + N_2}{\theta^2} - \frac{N_2 + 2N_3}{(1-\theta)^2}. \end{aligned}$$

- The Fisher information number

$$\begin{aligned} I_n(\theta) &= -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln p(\mathbf{N}, \theta)\right] = n\left[\frac{(2p_1 + p_2)}{\theta^2} + \frac{(p_2 + 2p_3)}{(1-\theta)^2}\right] \\ &= n\left[\frac{2\theta}{\theta^2} + \frac{2(1-\theta)}{(1-\theta)^2}\right] = \frac{2n}{\theta(1-\theta)}. \end{aligned}$$

5.1 U-Statistics

- It is known that, if the order statistic $(X_{(1)}, \ldots, X_{(n)})$ is sufficient and complete, then a symmetric unbiased estimator of an estimable ϑ is the UMVUE of ϑ.
- In many problems, parameters to be estimated are of the form

$$\vartheta = E[h(X_1, ..., X_m)]$$

  with a positive integer m and a Borel function h that is symmetric in its arguments.
- It is easy to see that a symmetric unbiased estimator of ϑ is

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}),$$

  where $\sum_c$ denotes the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, ..., i_m\}$ from $\{1, ..., n\}$.

Definition 5.1 (U-Statistics)**.** The statistic

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}),$$

is called a *U-statistic* with kernel h of order m.

- The use of U-statistics is an effective way of obtaining useful unbiased estimators.
- In nonparametric problems, U-statistics are often UMVUE's, whereas in parametric problems, U-statistics can be used as initial estimators to derive more efficient estimators.
- If $m = 1$, a U-statistic is simply a type of sample mean. Examples include the empirical c.d.f. evaluated at a particular t and the *sample moments* $n^{-1} \sum_{i=1}^n X_i^k$ for a positive integer k.

Example 5.2.

- Consider the estimation of $\vartheta = \mu^m$, where $\mu = EX_1$ and m is a positive integer. Using $h(x_1, ..., x_m) = x_1 \cdots x_m$, we obtain the following U-statistic unbiased for $\vartheta = \mu^m$:

$$U_n = \binom{n}{m}^{-1} \sum_c X_{i_1} \cdots X_{i_m}.$$

- Consider the estimation of $\vartheta = \sigma^2 = \text{Var}(X_1)$. Since

$$\sigma^2 = [\text{Var}(X_1) + \text{Var}(X_2)]/2 = E[(X_1 - X_2)^2/2],$$

we obtain the following U-statistic with kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) = S^2,$$

which is the sample variance.

- In some cases, we would like to estimate $\vartheta = E|X_1 - X_2|$, a measure of concentration. Using kernel $h(x_1, x_2) = |x_1 - x_2|$, we obtain the following U-statistic unbiased for $\vartheta = E|X_1 - X_2|$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|,$$

which is known as *Gini's mean difference*.

- Let $\vartheta = P(X_1 + X_2 \le 0)$. Using kernel $h(x_1, x_2) = I_{(-\infty, 0]}(x_1 + x_2)$, we obtain the following U-statistic unbiased for ϑ:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} I_{(-\infty, 0]}(X_i + X_j),$$

which is known as the *one-sample Wilcoxon statistic*.

- Variance of a U-statistic
  - For $k = 1, ..., m$, let

$$\begin{aligned} h_k(x_1, ..., x_k) &= E[h(X_1, ..., X_m) | X_1 = x_1, ..., X_k = x_k] \\ &= E[h(x_1, ..., x_k, X_{k+1}, ..., X_m)]. \end{aligned}$$

Note that $h_m = h$. It can be shown that

$$h_k(x_1, ..., x_k) = E[h_{k+1}(x_1, ..., x_k, X_{k+1})].$$

Define

$$\tilde{h}_k = h_k - E[h(X_1, ..., X_m)],$$

$k = 1, ..., m$, and $\tilde{h} = \tilde{h}_m$.

  - For any U-statistic

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}),$$

it can be represented by

$$U_n - E(U_n) = \binom{n}{m}^{-1} \sum_c \tilde{h}(X_{i_1}, ..., X_{i_m}). \tag{5.15}$$

Theorem 5.3 (Hoeffding). *For a U-statistic U_n with $E[h(X_1, ..., X_m)]^2 < \infty$,*

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k}\binom{n-m}{m-k}\zeta_k,$$

where

$$\zeta_k = \mathrm{Var}(h_k(X_1, ..., X_k)).$$

Proof: Consider two sets $\{i_1, ..., i_m\}$ and $\{j_1, ..., j_m\}$ of m distinct integers from $\{1, ..., n\}$ with exactly k integers in common. The number of distinct choices of two such sets is $\binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}$. By the symmetry of $\tilde{h}_m$ and independence of $X_1, ..., X_n$,

$$E[\tilde{h}(X_{i_1}, ..., X_{i_m})\tilde{h}(X_{j_1}, ..., X_{j_m})] = \zeta_k$$

for $k = 1, ..., m$. Then, by (5.15),

$$\begin{aligned}\mathrm{Var}(U_n) &= \binom{n}{m}^{-2} \sum_c \sum_c E[\tilde{h}(X_{i_1}, ..., X_{i_m})\tilde{h}(X_{j_1}, ..., X_{j_m})] \\ &= \binom{n}{m}^{-2} \sum_{k=1}^{m} \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}\zeta_k.\end{aligned}$$

This proves the result.

Proposition 5.4. *Under the condition of Hoeffding's theorem,*

(a) $\zeta_1 \leq \ldots \leq \zeta_m$.

(b) $(n+1)\mathrm{Var}(U_{n+1}) \leq n\mathrm{Var}(U_n)$ *for any* $n > m$.

(c) *For any fixed* m *and* $k = 1, ..., m$, *if* $\zeta_j = 0$ *for* $j < k$ *and* $\zeta_k > 0$, *then*

$$\mathrm{Var}(U_n) = \frac{k!\binom{m}{k}^2\zeta_k}{n^k} + O\left(\frac{1}{n^{k+1}}\right).$$

(d) $\frac{m^2}{n}\zeta_1 \leq \mathrm{Var}(U_n) \leq \frac{m}{n}\zeta_m$.

Proof: (a) Let $W = h_{k+1}(X_1, \ldots, X_k, X_{k+1})$ and $Y = (X_1, \ldots, X_k)$. Then $\zeta_{k+1} = \mathrm{Var}(W)$ and $\zeta_k = \mathrm{Var}(E(W|Y))$ since $h_k(X_1, \ldots, X_k) = E\{h_{k+1}(X_1, \ldots, X_{k+1})|X_1, \ldots, X_k\}$. Now we observe that $\zeta_k = \mathrm{Var}(E(W|Y)) = \mathrm{Var}(W) - E(\mathrm{Var}(W|Y)) \leq \mathrm{Var}(W) = \zeta_{k+1}$.

(b) The proof of this one requires Hoeffding's representation of the statistic $U_n - EU_n$ and is omitted here.

(c) By Hoeffding's theorem,

$$\mathrm{Var}(U_n) = \sum_{j=1}^{m} \frac{\binom{m}{j}\binom{n-m}{m-j}}{\binom{n}{m}}\zeta_j.$$

For any $j = 1, \ldots, m$,

$$\begin{aligned}\frac{\binom{m}{j}\binom{n-m}{m-j}}{\binom{n}{m}} &= j!\binom{m}{j}^2 \frac{(n-m)\cdots[n-m-(m-j-1)]}{n\cdots(n-m+1)} \\ &= j!\binom{m}{j}^2 \left[\frac{1}{n^j} + O\left(\frac{1}{n^{j+1}}\right)\right] \\ &= O\left(\frac{1}{n^j}\right).\end{aligned}$$

If $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then

$$\begin{aligned}\mathrm{Var}(U_n) &= \sum_{j=k}^{m} \frac{\binom{m}{j}\binom{n-m}{m-j}}{\binom{n}{m}}\zeta_j \\ &= \frac{\binom{m}{k}\binom{n-m}{m-k}}{\binom{n}{m}}\zeta_k + \sum_{j=k+1}^{m} \frac{\binom{m}{j}\binom{n-m}{m-j}}{\binom{n}{m}}\zeta_j \\ &= k!\binom{m}{k}^2 \zeta_k \frac{1}{n^k} + O\left(\frac{1}{n^{k+1}}\right) + \sum_{j=k+1}^{m} O\left(\frac{1}{n^j}\right) \\ &= k!\binom{m}{k}^2 \zeta_k \frac{1}{n^k} + O\left(\frac{1}{n^{k+1}}\right).\end{aligned}$$

(d) From (b), $n\mathrm{Var}(U_n)$ is non increasing. Thus, $n\mathrm{Var}(U_n) \le (n-1)\mathrm{Var}(U_{n-1}) \le \cdots \le m\mathrm{Var}(U_m) = m\zeta_m$. The first inequality is trivial if $\zeta_1 = 0$. Otherwise, from (c), $n\mathrm{Var}(U_n) \ge \lim_n[n\mathrm{Var}(U_n)] = m^2\zeta_1$.

- It follows from the Corollary that a U-statistic U_n as an estimator of its mean, its MSE converges to 0 (under the finite second moment assumption on h).
- In fact, for any fixed k, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then the MSE of U_n is of the order n^{-k}.
- In particular, the above results imply that U_n is consistent.

Example 5.5.

- Consider first $h(x_1, x_2) = x_1x_2$, which leads to a U-statistic unbiased for μ^2, where $\mu = EX_1$. Note that $h_1(x_1) = \mu x_1$, $\tilde{h}_1(x_1) = \mu(x_1 - \mu)$, $\zeta_1 = E[\tilde{h}_1(X_1)]^2 = \mu^2\mathrm{Var}(X_1) = \mu^2\sigma^2$, $\tilde{h}(x_1, x_2) = x_1x_2 - \mu^2$, and $\zeta_2 = \mathrm{Var}(X_1X_2) = E(X_1X_2)^2 - \mu^4 = (\mu^2 + \sigma^2)^2 - \mu^4$. By Hoeffding's theorem, for

$$U_n = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} X_iX_j,$$

$$\begin{aligned}\mathrm{Var}(U_n) &= \binom{n}{2}^{-1}\left[\binom{2}{1}\binom{n-2}{1}\zeta_1 + \binom{2}{2}\binom{n-2}{0}\zeta_2\right] \\ &= \frac{2}{n(n-1)}\left[2(n-2)\mu^2\sigma^2 + (\mu^2+\sigma^2)^2 - \mu^4\right] \\ &= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}.\end{aligned}$$

Comparing U_n with $\bar{X}^2 - \sigma^2/n$ which is the UMVUE under the normality and known σ^2 assumption, we find that

$$\mathrm{Var}(U_n) - \mathrm{Var}(\bar{X}^2 - \sigma^2/n) = \frac{2\sigma^4}{n^2(n-1)}.$$

- Next, consider $h(x_1, x_2) = I_{(-\infty,0]}(x_1+x_2)$, which leads to the one-sample Wilcoxon statistic. Note that $h_1(x_1) = P(x_1 + X_2 \le 0) = F(-x_1)$, where F is the c.d.f. of P. Then $\zeta_1 = \mathrm{Var}(F(-X_1))$. Let $\vartheta = E[h(X_1, X_2)]$. Then $\zeta_2 = \mathrm{Var}(h(X_1, X_2)) = \vartheta(1-\vartheta)$. Hence, for U_n being the one-sample Wilcoxon statistic,

$$\mathrm{Var}(U_n) = \frac{2}{n(n-1)}\left[2(n-2)\zeta_1 + \vartheta(1-\vartheta)\right].$$

Finally, consider $h(x_1, x_2) = |x_1 - x_2|$, which leads to Gini's mean difference.

- Note that $h_1(x_1) = E|x_1 - X_2| = \int |x_1 - y| dP(y)$, hence

$$\zeta_1 = \mathrm{Var}(h_1(X_1)) = \int\left[\int |x-y|dP(y)\right]^2 dP(x) - \vartheta^2,$$

  where $\vartheta = E|X_1 - X_2|$.

- Note that $h_2(x_1, x_2) = |x_1 - x_2|$ and

$$\zeta_2 = \mathrm{Var}(h_2(X_1, X_2)) = E|X_1 - X_2|^2 - (E|X_1 - X_2|)^2 = 2\sigma^2 - \vartheta^2.$$

- Hence, for U_n being the Gini's mean difference,

$$\begin{aligned}\mathrm{Var}(U_n) &= \frac{2}{n(n-1)}[2(n-2)\int\left(\int |x-y|dP(y)\right)^2 dP(x) \\ &\quad +2\sigma^2 - (2n-1)\vartheta^2].\end{aligned}$$

5.2 The projection method

Suppose $\mathcal{P}$ is nonparametric. In this case, the exact distribution of any U-statistic is hard to derive. We study asymptotic distributions of U-statistics by using the method of *projection*.

Definition 5.6 (Projection)**.** Let T_n be a given statistic based on $X_1, ..., X_n$. The projection of T_n on k_n random elements $Y_1, ..., Y_{k_n}$ is defined to be

$$\check{T}_n = E(T_n) + \sum_{i=1}^{k_n} [E(T_n|Y_i) - E(T_n)].$$

Let $\psi_n(X_i) = E(T_n|X_i)$. If T_n is symmetric (as a function of $X_1, ..., X_n$), then $\psi_n(X_1), ..., \psi_n(X_n)$ are i.i.d. with mean $E[\psi_n(X_i)] = E[E(T_n|X_i)] = E(T_n)$ and

$$E(\check{T}_n) = E(T_n)$$

If $E(T_n^2) < \infty$ and $\text{Var}(\psi_n(X_i)) > 0$, then

$$\frac{1}{\sqrt{n\text{Var}(\psi_n(X_1))}} \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)] \to_d N(0,1) \tag{5.16}$$

by the CLT. Let $\check{T}_n$ be the projection of T_n on $X_1, ..., X_n$. Then

$$T_n - \check{T}_n = T_n - E(T_n) - \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)]. \tag{5.17}$$

If we can show that $T_n - \check{T}_n$ has a negligible order of magnitude, then we can derive the asymptotic distribution of T_n by using (5.16)-(5.17) and Slutsky's theorem. The order of magnitude of $T_n - \check{T}_n$ can be obtained with the help of the following lemma.

Lemma 5.7. *Let T_n be a symmetric statistic with $\text{Var}(T_n) < \infty$ for every n and $\check{T}_n$ be the projection of T_n on $X_1, ..., X_n$. Then $E(T_n) = E(\check{T}_n)$ and*

$$E(T_n - \check{T}_n)^2 = \text{Var}(T_n) - \text{Var}(\check{T}_n).$$

Proof: Since $E(T_n) = E(\check{T}_n)$,

$$E(T_n - \check{T}_n)^2 = \text{Var}(T_n) + \text{Var}(\check{T}_n) - 2\text{Cov}(T_n, \check{T}_n)$$

$$\begin{aligned}
\text{Cov}(T_n, \check{T}_n) &= E(T_n\check{T}_n) - [E(T_n)]^2 \\
&= nE[T_nE(T_n|X_i)] - n[E(T_n)]^2 \\
&= nE\{E[T_nE(T_n|X_i)|X_i]\} - n[E(T_n)]^2 \\
&= nE\{[E(T_n|X_i)]^2\} - n[E(T_n)]^2 \\
&= n\text{Var}(E(T_n|X_i)) \\
&= \text{Var}(\check{T}_n)
\end{aligned}$$

by Definition 5.6 with $Y_i = X_i$ and $k_n = n$.

- This method of deriving the asymptotic distribution of T_n is known as the method of projection and is particularly effective for U-statistics. Now let $T_n = U_n$ for a U-statistic U_n and

set $k_n = n$. Let us compute $\psi_n(X_i) = E(U_n \mid X_i)$:

$$\begin{aligned}\psi_n(X_i) &= E(U_n \mid X_i) \\ &= \binom{n}{m}^{-1} \sum_c E\{h(X_{i_1}, \ldots, X_{i_m})|X_i\}\end{aligned}$$

For the term $E\{h(X_{i_1}, \ldots, X_{i_m})|X_i\}$,

- if $i \notin \{i_1, \ldots, i_m\}$, then $E\{h(X_{i_1}, \ldots, X_{i_m})|X_i\} = Eh(X_1, \ldots, X_m) = EU_n$ due to the symmetry of h and IID of $X_1, \ldots, X_n$. In total, there are $\binom{n-1}{m}$ such terms.
- if $i \in \{i_1, \ldots, i_m\}$, then $E\{h(X_{i_1}, \ldots, X_{i_m})|X_i\} = h_1(X_i)$, and there are $\binom{n-1}{m-1}$ such terms.

Thus,

$$\psi_n(X_i) = \frac{m}{n}\{h_1(X_i) - EU_n\} + EU_n$$

Let $\check{U}_n = EU_n + \sum_{i=1}^n \{\psi_n(X_-) - EU_n\} = EU_n + \sum_{i=1}^n \frac{m}{n}\{h_1(X_i) - EU_n\}$. Set

$$\tilde{h}_1(x) = h_1(x) - E[h(X_1, ..., X_m)].$$

Then $\check{U}_n = EU_n + \sum_{i=1}^n \frac{m}{n}\tilde{h}_1(X_i)$. Hence

$$\mathrm{Var}(\check{U}_n) = m^2\zeta_1/n$$

and, by Proposition 4 and Lemma 5.7,

$$E(U_n - \check{U}_n)^2 = O(n^{-2}).$$

- If $\zeta_1 > 0$, then

$$\frac{1}{\sqrt{n\mathrm{Var}(mh_1(X_1))}} \sum_{i=1}^n m[h_1(X_i) - E(U_n)] \to_d N(0, 1),$$

which leads to the result in Theorem 5.8(i) stated later.

- If $\zeta_1 = 0$, then $\tilde{h}_1 \equiv 0$ and we have to use another projection of U_n. Suppose that $\zeta_1 = \cdots = \zeta_{k-1} = 0$ and $\zeta_k > 0$ for an integer $k > 1$. Consider the projection $\check{U}_{kn}$ of U_n on $\binom{n}{k}$ random vectors $\{X_{i_1}, ..., X_{i_k}\}$, $1 \le i_1 < \cdots < i_k \le n$. We can establish a result similar to that in Lemma 5.7 and show that

$$E(U_n - \check{U}_n)^2 = O(n^{-(k+1)}).$$

Also, see Serfling (1980, §5.3.4).

Theorem 5.8. *Let U_n be a U-statistic with $E[h(X_1, ..., X_m)]^2 < \infty$.*

(i) If $\zeta_1 > 0$, then

$$\sqrt{n}[U_n - E(U_n)] \to_d N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ *but* $\zeta_2 > 0$*, then*

$$n[U_n - E(U_n)] \to_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j(\chi_{1j}^2 - 1), \tag{5.18}$$

where χ_{1j}^2*'s are i.i.d. random variables having the chi-square distribution* χ_1^2 *and* λ_j*'s are some constants (which may depend on* P*) satisfying* $\sum_{j=1}^{\infty} \lambda_j^2 = \zeta_2$*.*

Proof: We have actually proved (i). A proof for 3.5(ii) is given in Serfling (1980, §5.5.2).

- One may derive results for the cases where $\zeta_2 = 0$, but the case of either $\zeta_1 > 0$ or $\zeta_2 > 0$ is the most interesting case in applications.

We now apply Theorem 5.8 to the U-statistics in Example 2.

- Consider
$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} X_i X_j$$
Note that $\zeta_1 = \mu^2\sigma^2$. Thus, if $\mu \neq 0$, the result in Theorem 5.8(i) holds with $\zeta_1 = \mu^2\sigma^2$. If $\mu = 0$, then $\zeta_1 = 0$, $\zeta_2 = \sigma^4 > 0$, and Theorem 5.8(ii) applies.
- For the one-sample Wilcoxon statistic, $\zeta_1 = \text{Var}(F(-X_1)) > 0$ unless F is degenerate. Theorem 5.8(i) applies.
- Similarly, for Gini's mean difference, $\zeta_1 > 0$ unless F is degenerate. Theorem 5.8(i) applies.

Lecture 6: Linear Models and LSE

Lecturer: LIN Zhenhua ST5215 AY2019/2020 Semester I

6.1 Linear Models

A linear model is given below:

$$X_i = Z_i^\top \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, ..., n,$$

- X_i is the value of a response variable observed on the ith individual;
- Z_i is the value of a p-vector of explanatory variables (non-random covariates) observed on the ith individual;
- β is a p-vector of unknown parameters (main parameters of interest), $p < n$;
- ϵ_i is a random error (not observed) associated with the ith individual.

Let $X = (X_1, ..., X_n)^\top$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^\top$ $Z =$ the $n \times p$ matrix whose ith row is the vector $Z_i^\top$, $i = 1, ..., n$. A matrix form of the model is

$$X = Z\boldsymbol{\beta} + \epsilon. \tag{6.19}$$

- Suppose that the range of $\boldsymbol{\beta}$ in model (6.19) is $B \subset \mathcal{R}^p$.

A LSE (least squares estimator) of $\boldsymbol{\beta}$ is defined to be any $\hat{\boldsymbol{\beta}} \in B$ such that

$$\|X - Z\hat{\boldsymbol{\beta}}\|^2 = \min_{\mathbf{b} \in B} \|X - Z\mathbf{b}\|^2.$$

For any $a \in \mathcal{R}^p$, $a^\top \hat{\beta}$ is called an LSE of $a^\top \beta$.

- $\|A\| = \left(\sum_{ij} a_{ij}^2\right)^{1/2}$ is the Frobenius norm of a matrix A

Assume $B = \mathcal{R}^p$ unless otherwise stated. Differentiating $\|X - Z\mathbf{b}\|^2$ w.r.t. $\mathbf{b}$, we obtain the normal equation

$$Z^\top Z\mathbf{b} = Z^\top X.$$

Any solution of the normal equation is an LSE of β.

- $g(\mathbf{b}) = \|X - Z\mathbf{b}\|^2 = (X - Z\mathbf{b})^\top (X - Z\mathbf{b})$ is a quadratic form

- $\frac{\partial}{\partial \mathbf{b}}(\mathbf{b}^\top A\mathbf{b}) = 2A\mathbf{b}$ and $\frac{\partial}{\partial \mathbf{b}}(\mathbf{b}^\top A\mathbf{c}) = A\mathbf{c}$

- **The case of full rank** Z: If the rank of the matrix Z is p, in which case $(Z^\top Z)^{-1}$ exists and Z is said to be of full rank, then there is a unique LSE, which is

$$\hat{\beta} = (Z^\top Z)^{-1} Z^\top X.$$

- **The case of non full rank** Z: If Z is not of full rank, then there are infinitely many LSE's of β. Any LSE of β is of the form

$$\hat{\beta} = (Z^\top Z)^{-} Z^\top X,$$

  where $(Z^\top Z)^-$ is called a *generalized inverse* of $Z^\top Z$ and satisfies

$$Z^\top Z (Z^\top Z)^{-} Z^\top Z = Z^\top Z.$$

  Generalized inverse matrices are not unique unless Z is of full rank, in which case $(Z^\top Z)^- = (Z^\top Z)^{-1}$.

Some properties of general inverse

- $[Z(Z^\top Z)^- Z^\top]^2 = Z(Z^\top Z)^- Z^\top$.
- $Z(Z^\top Z)^- Z^\top Z = Z$.
- The rank of $Z(Z^\top Z)^- Z^\top$ is $\mathrm{tr}(Z(Z^\top Z)^- Z^\top) = r$.

6.2 Properties of LSE's of $\boldsymbol{\beta}$

To study properties of LSE's of β, we need some assumptions on the distribution of X or ϵ (conditional on Z if Z is random).

A1: (Gaussian noise) ϵ is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$.

A2: (homoscedastic noise) $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ with an unknown $\sigma^2 > 0$.

A3: (general noise) $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon)$ is an unknown matrix.

- We have mentioned that, if the matrix Z is not of full rank, then the model is not identifiable.
- Suppose that the rank of Z is $r \le p$. Then there is an $n \times r$ submatrix Z_* of Z such that $Z = Z_* Q$ and Z_* is of rank r, where Q is a fixed $r \times p$ matrix. The model is identifiable if we consider the reparameterization $\tilde{\boldsymbol{\beta}} = Q\boldsymbol{\beta}$.

- In many applications, we are interested in estimating $\vartheta = l^\top \boldsymbol{\beta}$ for some $l \in \mathcal{R}^p$. But, estimation of $l^\top \boldsymbol{\beta}$ is meaningless unless $l = Q^\top c$ for some $c \in \mathcal{R}^r$ so that

$$l^\top \boldsymbol{\beta} = c^\top Q \boldsymbol{\beta} = c^\top \tilde{\boldsymbol{\beta}}.$$

Theorem 6.1 (Theorem 3.6 of the textbook)**.** *Assume model (6.19) with assumption A3.*

(i) A necessary and sufficient condition for $l \in \mathcal{R}^p$ being $Q^\top c$ for some $c \in \mathcal{R}^r$ is $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\top Z)$, where Q is given above and $\mathcal{R}(A)$ is the smallest linear subspace containing all rows of A.

(ii) If $l \in \mathcal{R}(Z)$, then the LSE $l^\top \hat{\boldsymbol{\beta}}$ is unique and unbiased for $l^\top \boldsymbol{\beta}$.

(iii) If $l \notin \mathcal{R}(Z)$ and assumption A1 holds, then $l^\top \boldsymbol{\beta}$ is not estimable.

Proof (i) Note that $a \in \mathcal{R}(A)$ iff $a = A^\top b$ for some vector b. If $l = Q^\top c$, then

$$l = Q^\top c = Q^\top Z_*^\top Z_* (Z_*^\top Z_*)^{-1} c = Z^\top [Z_* (Z_*^\top Z_*)^{-1} c].$$

Hence $l \in \mathcal{R}(Z)$. If $l \in \mathcal{R}(Z)$, then $l = Z^\top \zeta$ for some ζ and

$$l = (Z_* Q)^\top \zeta = Q^\top c, \qquad c = Z_*^\top \zeta.$$

(ii) If $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\top Z)$, then $l = Z^\top Z \zeta$ for some ζ. Since $\hat{\beta} = (Z^\top Z)^- Z^\top X$, we have

$$\begin{aligned} E(l^\top \hat{\beta}) &= E[l^\top (Z^\top Z)^- Z^\top X] = \zeta^\top Z^\top Z (Z^\top Z)^- Z^\top Z \beta \\ &= \zeta^\top Z^\top Z \beta = l^\top \beta. \end{aligned}$$

If $\bar{\beta}$ is any other LSE of β, then, by $Z^\top Z \bar{\beta} = Z^\top X$,

$$l^\top \hat{\beta} - l^\top \bar{\beta} = \zeta^\top (Z^\top Z)(\hat{\beta} - \bar{\beta}) = \zeta^\top (Z^\top X - Z^\top X) = 0.$$

(iii) Proof via Contraposition: Under A1, if there is an estimator $h(X, Z)$ unbiased for $l^\top \beta$, then

$$l^\top \beta = \int_{\mathcal{R}^n} h(x, Z)(2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\tfrac{1}{2\sigma^2} \|x - Z\beta\|^2\right\} dx.$$

Differentiating w.r.t. β and applying Theorem 2.1 lead to

$$l^\top = Z^\top \int_{\mathcal{R}^n} h(x, Z)(2\pi)^{-n/2} \sigma^{-n-2} (x - Z\beta) \exp\left\{-\tfrac{1}{2\sigma^2} \|x - Z\beta\|^2\right\} dx,$$

which implies $l \in \mathcal{R}(Z)$.

Example 6.2 (Example 3.13 of the textbook)**.** Suppose that $n = \sum_{j=1}^m n_j$ with m positive integers $n_1, ..., n_m$ and that Consider the model:

$$X_{ij} = \mu_i + \epsilon_{ik}, \qquad j = 1, \ldots, n_i, i = 1, \ldots, m,$$

where ϵ_{ij} are i.i.d random errors with mean 0 and variance σ^2. This model is called a one-way ANOVA model. Let $\mathbf{X}_i = (X_{i1} \ldots, X_{in_i})^\top$ and $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_m^\top)^\top$. Let J_k be the k-vector of ones and

$$Z = \begin{pmatrix} J_{n_1} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \cdots & J_{n_m} \end{pmatrix}.$$

Let $\boldsymbol{\beta} = (\mu_1, ..., \mu_m)^\top$ and $\boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{1n_1}, \ldots, \epsilon_{m1}, \ldots, \epsilon_{mn_m})^\top$. Then the one-way ANOVA model can be expressed as

$$\mathbf{X} = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Since $Z^\top Z = \text{Diag}(n_1, \ldots, n_m)$, $(Z^\top Z)^{-1} = \text{Diag}(n_1^{-1}, \ldots, n_m^{-1})$. Hence the unique LSE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (Z^\top Z)^{-1} Z^\top \mathbf{X} = (\bar{X}_{1\cdot}, \ldots, \bar{X}_{m\cdot})^\top,$$

where $\bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$. Sometimes the model is expressed as

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}, j = 1, ..., n_i, i = 1, ..., m, \tag{6.20}$$

with constraint $\sum \alpha_i = 0$. Let $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_m)^\top$. The LSE of $\boldsymbol{\beta}$ is given by

$$\hat{\beta} = \left(\bar{X}, \bar{X}_{1\cdot} - \bar{X}, ..., \bar{X}_{m\cdot} - \bar{X}\right),$$

where $\bar{X}$ is total sample mean.

6.2.1 The properties under assumption A1

Theorem 6.3 (Theorem 3.7, 3.8 of the textbook)**.** *Assume model $X = Z\beta + \epsilon$ with assumption A1: ϵ is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$.*

(i) The LSE $l^\top \hat{\beta}$ is the UMVUE of $l^\top \beta$ for any estimable $l^\top \beta$.

(ii) The UMVUE of σ^2 is $\hat{\sigma}^2 = (n-r)^{-1} \|X - Z\hat{\beta}\|^2$, where r is the rank of Z.

(iii) For any estimable parameter $l^\top \beta$, the UMVUE's $l^\top \hat{\beta}$ and $\hat{\sigma}^2$ are independent; the distribution of $l^\top \hat{\beta}$ is $N(l^\top \beta, \sigma^2 l^\top (Z^\top Z)^- l)$; and $(n-r)\hat{\sigma}^2/\sigma^2$ has the chi-square distribution χ^2_{n-r}.

Proof of (i) Let $\hat{\beta}$ be an LSE of β. By $Z^\top Z\hat{\beta} = Z^\top X$,

$$(X - Z\hat{\beta})^\top Z(\hat{\beta} - \beta) = (X^\top Z - X^\top Z)(\hat{\beta} - \beta) = 0.$$

Hence,

$$\begin{aligned} \|X - Z\beta\|^2 &= \|X - Z\hat{\beta} + Z\hat{\beta} - Z\beta\|^2 \\ &= \|X - Z\hat{\beta}\|^2 + \|Z\hat{\beta} - Z\beta\|^2 \\ &= \|X - Z\hat{\beta}\|^2 - 2\beta^\top Z^\top X + \|Z\beta\|^2 + \|Z\hat{\beta}\|^2. \end{aligned}$$

Using this result and assumption A1, we obtain the following joint Lebesgue p.d.f. of X:

$$(2\pi\sigma^2)^{-n/2}\exp\left\{\frac{\beta^\top Z^\top x}{\sigma^2} - \frac{\|x-Z\hat{\beta}\|^2+\|Z\hat{\beta}\|^2}{2\sigma^2} - \frac{\|Z\beta\|^2}{2\sigma^2}\right\}.$$

By Proposition 2.1 and the fact that $Z\hat{\beta} = Z(Z^\top Z)^- Z^\top X$ is a function of $Z^\top X$, the statistic $(Z^\top X, \|X - Z\hat{\beta}\|^2)$ is complete and sufficient for $\theta = (\beta, \sigma^2)$. Note that $\hat{\beta}$ is a function of $Z^\top X$ and, hence, a function of the complete sufficient statistic. If $l^\top\beta$ is estimable, then $l^\top\hat{\beta}$ is unbiased for $l^\top\beta$ (Theorem 3.6) and, hence, $l^\top\hat{\beta}$ is the UMVUE of $l^\top\beta$.

Proof of (ii) Since $\|X - Z\beta\|^2 = \|X - Z\hat{\beta}\|^2 + \|Z\hat{\beta} - Z\beta\|^2$ and $E(Z\hat{\beta}) = Z\beta$,

$$\begin{aligned}
E\|X - Z\hat{\beta}\|^2 &= E(X - Z\beta)^\top(X - Z\beta) - E(\beta - \hat{\beta})^\top Z^\top Z(\beta - \hat{\beta}) \\
&= \text{tr}\left(\text{Var}(X) - \text{Var}(Z\hat{\beta})\right) \\
&= \sigma^2[n - \text{tr}\left(Z(Z^\top Z)^- Z^\top Z(Z^\top Z)^- Z^\top\right)] \\
&= \sigma^2[n - \text{tr}\left((Z^\top Z)^- Z^\top Z\right)].
\end{aligned}$$

Since for each row of $Z \in \mathcal{R}(Z)$, $Z\hat{\beta}$ does not depend on the choice of $(Z^\top Z)^-$ in $\hat{\beta} = (Z^\top Z)^- Z^\top X$ (Theorem 3.6). Hence, we can evaluate $\text{tr}\big((Z^\top Z)^- Z^\top Z\big)$ using a particular $(Z^\top Z)^-$.

From the theory of linear algebra, there exists a $p \times p$ matrix C such that $CC^\top = I_p$ and

$$C^\top(Z^\top Z)C = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ is an $r \times r$ diagonal matrix whose diagonal elements are positive. Then, a particular choice of $(Z^\top Z)^-$ is

$$(Z^\top Z)^- = C\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix}C^\top \tag{6.21}$$

and

$$(Z^\top Z)^- Z^\top Z = C\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}C^\top$$

whose trace is r.
Hence $\hat{\sigma}^2$ is the UMVUE of σ^2, since it is a function of the complete sufficient statistic and

$$E\hat{\sigma}^2 = (n - r)^{-1}E\|X - Z\hat{\beta}\|^2 = \sigma^2.$$

Before we prove (iii), we need the following supplementary results:

- *Cochran's Theorem*: Suppose that $X \sim N(\boldsymbol{\mu}, \sigma^2 I_n)$ and $X^\top X = X^\top A_1 X + \cdots + X^\top A_k X$, where I_n is the $n \times n$ identity matrix and A_j is an $n \times n$ symmetric matrix with rank n_j, $j = 1, \ldots, k$. A necessary and sufficient condition that $\frac{1}{\sigma^2}X^\top A_j X$ has the non-central chi-square distribution $\chi^2_{n_j}(\delta_j)$, $j = 1, \ldots, k$, and $X^\top A_j X$'s are independent is that $n = n_1 + \cdots + n_k$ and in which case $\delta_j = \frac{1}{\sigma^2}\boldsymbol{\mu}^\top A_j \boldsymbol{\mu}$ and $\sum_{j=1}^k \delta_j = \frac{1}{\sigma^2}\boldsymbol{\mu}^\top\boldsymbol{\mu}$.

- A non-central chi-square distribution $\chi^2_n(\delta)$ has p.d.f. given by

$$e^{-\delta/2}\sum_{i=0}^{\infty}\frac{(\delta/2)^i}{i!}f_{2i+n}(x),$$

  where $f_k(x)$ is the p.d.f. of the chi-square distribution χ^2_k.

- If $X_i \sim N(\mu_i, \sigma^2), i = 1, \ldots, n$, are independent, then $\frac{1}{\sigma^2}\sum_{i=1}^n X_i^2 \sim \chi^2_n(\frac{1}{\sigma^2}\sum_{i=1}^n \mu_i^2)$.

Proof of (iii)

- The estimator $\hat{\sigma}^2 = \|X - Z\hat{\boldsymbol{\beta}}\|^2/(n-r)$ where $X - Z\hat{\boldsymbol{\beta}} = [I_n - Z(Z^\top Z)^- Z^\top]X$ which is linear in X, and $l^\top\hat{\beta} = l^\top(Z^\top Z)^- Z^\top X$, a function of $(Z^\top Z)^- Z^\top X$ which is linear in X as well. Under assumption A1, both $[I_n - Z(Z^\top Z)^- Z^\top]X$ and $(Z^\top Z)^- Z^\top X$ are normally distributed.

- Since $[I_n - Z(Z^\top Z)^- Z^\top]Z(Z^\top Z)^- = 0$, $[I_n - Z(Z^\top Z)^- Z^\top]X$ and $(Z^\top Z)^- Z^\top X$ are independent. Hence $l^\top\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

- Since $l^\top\boldsymbol{\beta}$ is estimable, $l^\top\hat{\boldsymbol{\beta}} \sim N(l^\top\boldsymbol{\beta}, \sigma^2 l^\top(Z^\top Z)^- l)$.

- From $X^\top X = X^\top[Z(Z^\top Z)^- Z^\top]X + X^\top[I_n - Z(Z^\top Z)^- Z^\top]X$ and Cochran's theorem, $(n-r)\hat{\sigma}^2/\sigma^2$ has the chi-square distribution $\chi^2_{n-r}(\delta)$ with

$$\delta = \sigma^{-2}\boldsymbol{\beta}^\top Z^\top[I_n - Z(Z^\top Z)^- Z^\top]Z\boldsymbol{\beta} = 0.$$

6.2.2 Properties under assumption A2

- A linear estimator for the linear model

$$X = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{6.22}$$

  is a linear function of X, i.e., $\mathbf{c}^\top X$ for some fixed vector $\mathbf{c}$.

- $l^\top\hat{\boldsymbol{\beta}}$ is a linear estimator, since $l^\top\hat{\boldsymbol{\beta}} = l^\top(Z^\top Z)^- Z^\top X$ with $\mathbf{c} = Z(Z^\top Z)^- l$.

- The variance of $\mathbf{c}^\top X$ is given by $\mathbf{c}^\top\mathrm{Var}(X)\mathbf{c} = \mathbf{c}^\top\mathrm{Var}(\boldsymbol{\epsilon})\mathbf{c}$.

- In particular, if $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 I_n$ and $l \in \mathcal{R}(Z)$,

$$\mathrm{Var}(l^\top\hat{\beta}) = l^\top(Z^\top Z)^- Z^\top\mathrm{Var}(\boldsymbol{\epsilon})Z(Z^\top Z)^- l = \sigma^2 l^\top(Z^\top Z)^- l.$$

Theorem 6.4 (Theorem 3.9). *Assume model $X = Z\beta + \epsilon$ with assumption A2: $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 I_n$ with an unknown $\sigma^2 > 0$.*

(i) A necessary and sufficient condition for the existence of a linear unbiased estimator of $l^\top\beta$ (i.e., an unbiased estimator that is linear in X) is $l \in \mathcal{R}(Z)$.

(ii) (Gauss-Markov theorem). If $l \in \mathcal{R}(Z)$, then the LSE $l^\top \hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator *(BLUE) of $l^\top \boldsymbol{\beta}$ in the sense that it has the minimum variance in the class of linear unbiased estimators of $l^\top \boldsymbol{\beta}$.*

Proof: (i) The sufficiency is established in Theorem 3.6. Now let $c^\top X$ be unbiased for $l^\top \boldsymbol{\beta}$. Then

$$l^\top \boldsymbol{\beta} = E(c^\top X) = c^\top EX = c^\top Z\boldsymbol{\beta}.$$

Since this equality holds for all $\boldsymbol{\beta}$, $l = Z^\top c$, i.e., $l \in \mathcal{R}(Z)$.

(ii) Let $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\top Z)$. Then $l = (Z^\top Z)\zeta$ for some ζ and $l^\top \hat{\beta} = \zeta^\top (Z^\top Z)\hat{\beta} = \zeta^\top Z^\top X$ by $Z^\top Z\hat{\beta} = Z^\top X$. Let $c^\top X$ be any linear unbiased estimator of $l^\top \beta$. From the proof of (i), $Z^\top c = l$. Then Hence

$$\begin{aligned}
\mathrm{Var}(c^\top X) &= \mathrm{Var}(c^\top X - \zeta^\top Z^\top X + \zeta^\top Z^\top X) \\
&= \mathrm{Var}(c^\top X - \zeta^\top Z^\top X) + \mathrm{Var}(\zeta^\top Z^\top X) \\
&+ 2\mathrm{Cov}(\zeta^\top Z^\top X, c^\top X - \zeta^\top Z^\top X) \\
&= \mathrm{Var}(c^\top X - \zeta^\top Z^\top X) + \mathrm{Var}(l^\top \hat{\beta}) \\
&\geq \mathrm{Var}(l^\top \hat{\beta}).
\end{aligned}$$

(ii, another proof) Under A1, $l^\top \hat{\beta}$ is the UMVUE. In particular, it has the smallest variance among all linear unbiased estimators. However, as long as $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 I$, the variances of the linear unbiased estimators do not depend on the the particular assumption A1. Hence $l^\top \hat{\beta}$ is the BLUE under A2.

6.2.3 Properties under assumption A3

Theorem 6.5 (Theorem 3.10)**.** *Assume model $X = Z\beta + \epsilon$ with assumption A3: $E(\epsilon) = 0$ and* $\mathrm{Var}(\epsilon)$ *is an unknown matrix. The following are equivalent.*

(a) $l^\top \hat{\beta}$ is the BLUE of $l^\top \beta$ for any $l \in \mathcal{R}(Z)$.

(b) $E(l^\top \hat{\beta}\eta^\top X) = 0$ for any $l \in \mathcal{R}(Z)$ and any η such that $E(\eta^\top X) = 0$.

(c) $Z^\top \mathrm{Var}(\epsilon)U = 0$, where U is a matrix such that $Z^\top U = 0$ and $\mathcal{R}(U^\top) + \mathcal{R}(Z^\top) = \mathcal{R}^n$.

(d) $\mathrm{Var}(\epsilon) = Z\Lambda_1 Z^\top + U\Lambda_2 U^\top$ for some Λ_1 and Λ_2.

(e) The matrix $Z(Z^\top Z)^- Z^\top \mathrm{Var}(\epsilon)$ is symmetric.

Corollary 6.6 (Corollary 3.3 of the textbook)**.** *Consider model $X = Z\beta + \epsilon$ with a full rank Z, $\boldsymbol{\epsilon} \sim N_n(0, \Sigma)$, where Σ is an unknown positive definite matrix. Then $l^\top \hat{\beta}$ is a UMVUE of $l^\top \beta$ for any $l \in \mathcal{R}^p$ iff one of (b)-(e) in Theorem 3.10 holds.*

Roadmap of proof: (a) $\Leftrightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (e) $\Rightarrow$ (b). (a) $\Leftrightarrow$ (b). We first show that (a) and (b) are equivalent, which is an analogue of Theorem 3.2(i). Suppose that (b) holds. Let $l \in \mathcal{R}(Z)$. If $c^\top X$ is unbiased for $l^\top \beta$, then $E(\eta^\top X) = 0$ with $\eta = c - Z(Z^\top Z)^- l$. Hence, (b) implies (a) because

$$\begin{aligned}
\mathrm{Var}(c^\top X) &= \mathrm{Var}(c^\top X - l^\top \hat{\beta} + l^\top \hat{\beta}) \\
&= \mathrm{Var}(c^\top X - l^\top (Z^\top Z)^- Z^\top X + l^\top \hat{\beta}) \\
&= \mathrm{Var}(\eta^\top X + l^\top \hat{\beta}) \\
&= \mathrm{Var}(\eta^\top X) + \mathrm{Var}(l^\top \hat{\beta}) + 2\mathrm{Cov}(\eta^\top X, l^\top \hat{\beta}) \\
&= \mathrm{Var}(\eta^\top X) + \mathrm{Var}(l^\top \hat{\beta}) + 2E(l^\top \hat{\beta} \eta^\top X) \\
&= \mathrm{Var}(\eta^\top X) + \mathrm{Var}(l^\top \hat{\beta}) \\
&\geq \mathrm{Var}(l^\top \hat{\beta}).
\end{aligned}$$

Suppose now that there are $l \in \mathcal{R}(Z)$ and η such that $E(\eta^\top X) = 0$ but $\delta = E(l^\top \hat{\beta} \eta^\top X) \neq 0$. Let $c_t = t\eta + Z(Z^\top Z)^- l$. From the previous proof,

$$\mathrm{Var}(c_t^\top X) = t^2 \mathrm{Var}(\eta^\top X) + \mathrm{Var}(l^\top \hat{\beta}) + 2\delta t.$$

As long as $\delta \neq 0$, there exists a t such that $\mathrm{Var}(c_t^\top X) < \mathrm{Var}(l^\top \hat{\beta})$. This shows that $l^\top \hat{\beta}$ cannot be a BLUE and hence implies (b).

(b) $\Rightarrow$ (c). Suppose that (b) holds. Since $l \in \mathcal{R}(Z)$, $l = Z^\top \gamma$ for some γ. Let $\eta \in \mathcal{R}(U^\top)$. Then $E(\eta^\top X) = \eta^\top Z\beta = 0$ and, hence,

$$0 = E(l^\top \hat{\beta} \eta^\top X) = E[\gamma^\top Z(Z^\top Z)^- Z^\top X X^\top \eta] = \gamma^\top Z(Z^\top Z)^- Z^\top \mathrm{Var}(\boldsymbol{\epsilon}) \eta.$$

Since this equality holds for all $l \in \mathcal{R}(Z)$, it holds for all γ.

Thus,

$$Z(Z^\top Z)^- Z^\top \mathrm{Var}(\boldsymbol{\epsilon}) U = 0,$$

which implies

$$Z^\top Z(Z^\top Z)^- Z^\top \mathrm{Var}(\boldsymbol{\epsilon}) U = Z^\top \mathrm{Var}(\boldsymbol{\epsilon}) U = 0,$$

since $Z^\top Z(Z^\top Z)^- Z^\top = Z^\top$. Thus, (c) holds.

(c) $\Rightarrow$ (d). We need to use the following facts from the theory of linear algebra: there exists a nonsingular matrix C such that $\mathrm{Var}(\boldsymbol{\epsilon}) = CC^\top$ and $C = ZC_1 + UC_2$ for some matrices C_j (since $\mathcal{R}(U^\top) + \mathcal{R}(Z^\top) = \mathcal{R}^n$). Let $\Lambda_1 = C_1 C_1^\top$, $\Lambda_2 = C_2 C_2^\top$, and $\Lambda_3 = C_1 C_2^\top$.

Then

$$\mathrm{Var}(\boldsymbol{\epsilon}) = Z\Lambda_1 Z^\top + U\Lambda_2 U^\top + Z\Lambda_3 U^\top + U\Lambda_3^\top Z^\top \tag{6.23}$$

and $Z^\top \mathrm{Var}(\boldsymbol{\epsilon}) U = Z^\top Z \Lambda_3 U^\top U$, which is 0 if (c) holds. Hence, (c) implies

$$0 = Z(Z^\top Z)^- Z^\top Z \Lambda_3 U^\top U (U^\top U)^- U^\top = Z\Lambda_3 U^\top,$$

which with (6.23) implies (d).

(d) $\Rightarrow$ (e). If (d) holds, then $Z(Z^\top Z)^- Z^\top \mathrm{Var}(\boldsymbol{\epsilon}) = Z\Lambda_1 Z^\top$, which is symmetric.

To complete the proof, we need to show that (e) implies (b), which is left as an exercise.

6.3 Asymptotic Properties of LSE

Theorem 6.7 (Theorem 3.11 (Consistency) of the textbook)**.** *Consider model $X = Z\beta + \epsilon$ under assumption A3, i.e., $E(\boldsymbol{\epsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon})$ is an unknown matrix. Consider the LSE $l^\top \hat{\boldsymbol{\beta}}$ with $l \in \mathcal{R}(Z)$ for every n. Suppose that $\sup_n \lambda_+[\mathrm{Var}(\boldsymbol{\epsilon})] < \infty$, where $\lambda_+[A]$ is the largest eigenvalue of the matrix A, and that $\lim_{n\to\infty} \lambda_+[(Z^\top Z)^-] = 0$. Then $l^\top \hat{\boldsymbol{\beta}}$ is consistent in MSE for any $l \in \mathcal{R}(Z)$, i.e., $l^\top \hat{\boldsymbol{\beta}} \to l^\top \boldsymbol{\beta}$ in L_2.*

Proof: The result follows from the fact that $l^\top \hat{\beta}$ is unbiased and

$$\begin{aligned} \mathrm{Var}(l^\top \hat{\beta}) &= l^\top (Z^\top Z)^- Z^\top \mathrm{Var}(\epsilon) Z (Z^\top Z)^- l \\ &\le \lambda_+[\mathrm{Var}(\epsilon)] l^\top (Z^\top Z)^- l \le \lambda_+[\mathrm{Var}(\epsilon)] \lambda_+((Z^\top Z)^-) l^\top l. \end{aligned}$$

Theorem 6.8 (Theorem 3.12 (Asymptotic Normality) of the textbook)**.** *Consider model $X = Z\beta + \epsilon$ under assumption A3. Suppose that $0 < \inf_n \lambda_-[\mathrm{Var}(\epsilon)]$, where $\lambda_-[A]$ is the smallest eigenvalue of the matrix A, and that*

$$\lim_{n\to\infty} \max_{1\le i\le n} Z_i^\top (Z^\top Z)^- Z_i = 0. \tag{6.24}$$

Suppose further that $n = \sum_{j=1}^k m_j$ for some integers k, m_j, $j = 1, ..., k$, with m_j's bounded by a fixed integer m, $\epsilon = (\xi_1, ..., \xi_k)$, $\xi_j \in \mathcal{R}^{m_j}$, and ξ_j's are independent.

(i) If $\sup_i E|\epsilon_i|^{2+\delta} < \infty$, then for any $l \in \mathcal{R}(Z)$,

$$l^\top (\hat{\beta} - \beta) \Big/ \sqrt{\mathrm{Var}(l^\top \hat{\beta})} \to_d N(0, 1). \tag{6.25}$$

(ii) Result (6.25) holds for any $l \in \mathcal{R}(Z)$ if, when $m_i = m_j$, $1 \le i < j \le k$, ξ_i and ξ_j have the same distribution.

Proof: For $l \in \mathcal{R}(Z)$,

$$l^\top (Z^\top Z)^- Z^\top Z\beta - l^\top \beta = 0$$

and

$$l^\top (\hat{\beta} - \beta) = l^\top (Z^\top Z)^- Z^\top \epsilon = \sum_{j=1}^k c_{nj}^\top \xi_j,$$

where c_{nj} is the m_j-vector whose components are $l^\top (Z^\top Z)^- Z_i$, $i = k_{j-1} + 1, ..., k_j$, $k_0 = 0$, and $k_j = \sum_{t=1}^j m_t$, $j = 1, ..., k$. Note that

$$\sum_{j=1}^k \|c_{nj}\|^2 = l^\top (Z^\top Z)^- Z^\top Z (Z^\top Z)^- l = l^\top (Z^\top Z)^- l. \tag{6.26}$$

Also,

$$\begin{aligned}\max_{1\le j\le k}\|c_{nj}\|^2 &\le m\max_{1\le i\le n}[l^\top(Z^\top Z)^- Z_i]^2\\ &\le ml^\top(Z^\top Z)^- l\max_{1\le i\le n} Z_i^\top(Z^\top Z)^- Z_i,\end{aligned}$$

which, together with (6.26) and condition (6.24), implies that

$$\lim_{n\to\infty}\left(\max_{1\le j\le k}\|c_{nj}\|^2\Big/\sum_{j=1}^k\|c_{nj}\|^2\right)=0.$$

The results then follow from Corollary 1.3 of the textbook.

- Under the conditions of Theorem 3.12, $\mathrm{Var}(\epsilon)$ is a diagonal block matrix with $\mathrm{Var}(\xi_j)$ as the jth diagonal block, which includes the case of independent ϵ_i's as a special case.

Condition (6.24) is almost necessary for the consistency of the LSE.

Exercise 6.9. *Let $\hat{X}_i = Z_i^\top\hat{\beta}$ and $h_i = Z_i^\top(Z^\top Z)^- Z_i$. Suppose assumption A2 holds.*

(a) For any $\delta > 0$,

$$P(|\hat{X}_i - E\hat{X}_i|\ge\delta)\ge\min\{P(\varepsilon_i\ge\delta/h_i), P(\varepsilon_i\le-\delta/h_i)\}.$$

(b) $\hat{X}_i - E\hat{X}_i \xrightarrow{P} 0$ if and only if $h_i\to 0$.

Lemma 6.10 (Lemma 3.3 of the textbook)**.** *The following are sufficient conditions for (6.24).*

(a) $\lambda_+[(Z^\top Z)^-]\to 0$ and $Z_n^\top(Z^\top Z)^- Z_n\to 0$, as $n\to\infty$.

(b) There is an increasing sequence $\{a_n\}$ such that $a_n\to\infty$, $a_n/a_{n+1}\to 1$, and $Z^\top Z/a_n$ converges to a positive definite matrix.

Proof: (a) Since $Z^\top Z$ depends on n, we denote $(Z^\top Z)^-$ by A_n. Let i_n be the integer such that $h_{i_n} = \max_{1\le i\le n} h_i$. If $\lim_{n\to\infty} i_n = \infty$, then

$$\lim_{n\to\infty} h_{i_n} = \lim_{n\to\infty} Z_{i_n}^\top A_n Z_{i_n} \le \lim_{n\to\infty} Z_{i_n}^\top A_{i_n} Z_{i_n} = 0,$$

where the inequality follows from $i_n\le n$ and, thus, $A_{i_n} - A_n$ is nonnegative definite. If $i_n\le c$ for all n, then

$$\lim_{n\to\infty} h_{i_n} = \lim_{n\to\infty} Z_{i_n}^\top A_n Z_{i_n} \le \lim_{n\to\infty}\lambda_n\max_{1\le i\le c}\|Z_i\|^2 = 0.$$

Therefore, for any subsequence $\{j_n\}\subset\{i_n\}$ with $\lim_{n\to\infty} j_n = a\in(0,\infty]$, $\lim_{n\to\infty} h_{j_n} = 0$. This shows that $\lim_{n\to\infty} h_{i_n} = 0$.

(b) We show that (b) implies (a). Let A be the limit of $Z^\top Z/a_n$. Let $\lambda_-(A) > 0$ be the smallest eigenvalue of A. Then $\lambda_-(Z^\top Z/a_n)\ge\lambda_-(A)/2$ for all sufficiently large n, or equivalently,

$\lambda_+[(Z^\top Z)^-] \le 2/(a_n\lambda_-(A)) \to 0$. Note that, the diagonal element $\sum_{i=1}^n Z_{ik}^2/a_n$ of $Z^\top Z$ converges to a fixed constant C for sufficiently large n. Then $Z_{nk}^2/a_n \to 0$ for all $k = 1, \ldots, p$ and further $\|Z_n\|^2/a_n = \sum_{k=1}^p Z_{nk}^2/a_n \to 0$. Together with $\lambda_-[(Z^\top Z)^-] \le 2/(a_n\lambda_-(A))$, we conclude that $Z_n^\top (Z^\top Z)^- Z_n \to 0$. [Note that, for sufficiently large n, $Z^\top Z$ is invertible by the assumption of (b)].

Example 6.11 (Simple linear models)**.** In Example 3.12,

$$X_i = \beta_0 + \beta_1 t_i + \epsilon_i, \quad i = 1, ..., n.$$

If $n^{-1}\sum_{i=1}^n t_i^2 \to c$ and $n^{-1}\sum_{i=1}^n t_i \to d$ where c is positive and $c > d^2$, then condition (b) in Lemma 3.3 is satisfied with $a_n = n$ and, therefore, Theorem 3.12 applies.

Example 6.12 (One-way ANOVA)**.** In the one-way ANOVA model (Example 3.13),

$$X_i = \mu_j + \epsilon_i, \qquad i = k_{j-1} + 1, ..., k_j, \ j = 1, ..., m,$$

where $k_0 = 0$, $k_j = \sum_{l=1}^j n_l$, $j = 1, ..., m$, and $(\mu_1, ..., \mu_m) = \beta$,

$$\max_{1\le i\le n} Z_i^\top (Z^\top Z)^- Z_i = \lambda_+[(Z^\top Z)^-] = \max_{1\le j\le m} n_j^{-1}.$$

Conditions related to Z in Theorem 3.12 are satisfied iff $\min_j n_j \to \infty$.

Lecture 7: Asymptotically Unbiased Estimators

Lecturer: LIN Zhenhua ST5215 AY2019/2020 Semester I

7.1 Asymptotic MSE, variance and efficiency: revisited

Recall:

- Asymptotic normality: Let $T_n = T(\mathbf{X}_n)$ be an estimator based on a sample $\mathbf{X}_n$ of size n. Let $\mu_n(\theta)$ and $\sigma_n^2(\theta)$ be two sequences of constants which might depend on θ. If

$$\frac{T_n - \mu_n(\theta)}{\sigma_n(\theta)} \to_d N(0,1),$$

  then T_n is said to be asymptotically normal with asymptotic mean and variance $\mu_n(\theta)$ and $\sigma_n^2(\theta)$ respectively.

- Asymptotic unbiasedness:
  - When $E|T_n| < \infty$, T_n is asymptotically unbiased for $q(\theta)$ if $b_{T_n}(\theta) = ET_n - q(\theta) \to 0$.
  - When ET_n is not defined: Suppose that $T_n = T(\mathbf{X}_n)$ is asymptotically normal with asymptotic mean $\mu_n(\theta)$ and asymptotic variance $\sigma_n^2(\theta)$. If

$$\frac{\mu_n(\theta) - q(\theta)}{\sigma_n(\theta)} \to 0,$$

    then T_n is said to be asymptotically unbiased for $q(\theta)$.
  - these two concepts of asymptotic unbiasedness are different. The latter is stronger than the former.
  - More generally, if $a_n T_n \xrightarrow{D} T$ and $E|T| < \infty$, then ET/a_n^2 is called the asymptotic expectation of T_n. When T_n is an estimator of θ, then an asymptotic expectation of $T_n - \theta$ is called an asymptotic bias of T_n, and is denoted by $\tilde{b}_{T_n}(\theta)$. We say T_n is asymptotically unbiased if $\tilde{b}_{T_n}(\theta) \to 0$ for all θ.
  - For a given T_n, its asymptotic expectations are essentially the same.

- Asymptotic relative efficiency: Let $T^{(1)} = \{T_n^{(1)}\}$ and $T^{(2)} = \{T_n^{(2)}\}$ be two sequences of estimators which are asymptotically unbiased for $q(\theta)$ and whose asymptotic variances σ_{n1}^2 and σ_{n2}^2 satisfy $n\sigma_{ni}^2 \to \sigma_i^2, i = 1, 2$. The asymptotic relative efficiency of $T^{(1)}$ to $T^{(2)}$ is defined by

$$e(\theta, T^{(1)}, T^{(2)}) = \frac{\sigma_2^2}{\sigma_1^2}.$$

- Asymptotic efficient estimator: Suppose that $T_n = T(\mathbf{X}_n)$ is asymptotically normal with asymptotic mean $\mu_n(\theta)$ and asymptotic variance $\sigma_n^2(\theta)$. If

$$n\sigma_n^2(\theta) \to \sigma^2(\theta) > 0,\ \sqrt{n}(\mu_n(\theta) - q(\theta)) \to 0,$$
$$\sigma^2(\theta) = \frac{[q'(\theta)]^2}{I_1(\theta)},$$

  then T_n is said to be asymptotically efficient (or best asymptotically normal).

Like the bias, $\mathrm{MSE}_{T_n}(P) = E(T_n - \vartheta)^2$, is not well defined if the second moment of T_n does not exist. We now define a version of *asymptotic mean squared error* (amse) and a measure of assessing different point estimators of a common parameter.

Definition 7.1 (Definition 2.12)**.** Let T_n be an estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Assume that $a_n(T_n - \vartheta) \xrightarrow{D} Y$ with $0 < EY^2 < \infty$.

(i) The asymptotic mean squared error of T_n, denoted by $\mathrm{AMSE}_{T_n}(P)$ or $\mathrm{AMSE}_{T_n}(\theta)$ if P is in a parametric family indexed by θ, is defined to be the asymptotic expectation of $(T_n - \vartheta)^2$, i.e., $\mathrm{AMSE}_{T_n}(P) = EY^2/a_n^2$. The asymptotic variance of T_n is defined to be $\sigma^2_{T_n}(P) = \mathrm{Var}(Y)/a_n^2$.

(ii) Let T_n' be another estimator of ϑ. The *asymptotic relative efficiency* of T_n' w.t.r. T_n is defined to be $e_{T_n', T_n}(P) = \mathrm{AMSE}_{T_n}(P)/\mathrm{AMSE}_{T_n'}(P)$.

(iii) T_n is said to be *asymptotically more efficient* than T_n' iff $\limsup_n e_{T_n', T_n}(P) \le 1$ for any P and < 1 for some P.

- The amse and asymptotic variance are the same iff $EY = 0$.
- By Proposition 2.3, the amse or the asymptotic variance of T_n is essentially unique and, therefore, the concept of asymptotic relative efficiency in Definition 2.12(ii)-(iii) is well defined.

When both $\mathrm{MSE}_{T_n}(P)$ and $\mathrm{MSE}_{T_n'}(P)$ exist, one may compare T_n and T_n' by evaluating the relative efficiency $\mathrm{MSE}_{T_n}(P)/\mathrm{MSE}_{T_n'}(P)$. However, this comparison may be different from the one using the asymptotic relative efficiency in Definition 2.12(ii), since the mse and amse of an estimator may be different (Exercise 115 in §2.6). The following result shows that when the exact mse of T_n exists, it is no smaller than the amse of T_n. It also provides a condition under which the exact mse and the amse are the same.

Proposition 7.2 (Proposition 2.4)**.** *Let T_n be an estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Suppose that $a_n(T_n - \vartheta) \xrightarrow{D} Y$ with $0 < EY^2 < \infty$. Then*

(i) $EY^2 \le \liminf_n E[a_n^2(T_n - \vartheta)^2]$ and

(ii) $EY^2 = \lim_{n\to\infty} E[a_n^2(T_n - \vartheta)^2]$ *if and only if* $\{a_n^2(T_n - \vartheta)^2\}$ *is uniformly integrable.*

Proof:

(i) By Theorem 1.10(iii),

$$\min\{a_n^2(T_n - \vartheta)^2, t\} \xrightarrow{D} \min\{Y^2, t\} \quad \text{for any } t > 0.$$

Since $\min\{a_n^2(T_n - \vartheta)^2, t\}$ is bounded by t, by Theorem 1.8(viii),

$$\lim_{n\to\infty} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) = E(\min\{Y^2, t\})$$

Then

$$\begin{aligned} EY^2 &= \lim_{t\to\infty} E(\min\{Y^2, t\}) \\ &= \lim_{t\to\infty} \lim_{n\to\infty} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) \\ &= \liminf_{t,n} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) \\ &\le \liminf_{n} E[a_n^2(T_n - \vartheta)^2], \end{aligned}$$

where the third equality follows from the fact that $E(\min\{a_n^2(T_n - \vartheta)^2, t\})$ is nondecreasing in t for any fixed n. (ii) The result follows from Theorem 1.8(viii). □

Example 7.3 (Example 2.36)**.** Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Consider the estimation of $\vartheta = P(X_i = 0) = e^{-\theta}$.

- Let $T_{1n} = F_n(0)$, where F_n is the empirical c.d.f.
  - Then T_{1n} is unbiased and has $\text{MSE}_{T_{1n}}(\theta) = e^{-\theta}(1 - e^{-\theta})/n$.
  - Also, $\sqrt{n}(T_{1n} - \vartheta) \xrightarrow{D} N(0, e^{-\theta}(1 - e^{-\theta}))$ by the CLT.
  - Thus, in this case $\text{AMSE}_{T_{1n}}(\theta) = \text{MSE}_{T_{1n}}(\theta)$.
- Consider $T_{2n} = e^{-\bar{X}}$.
  - Note that $ET_{2n} = e^{n\theta(e^{-1/n}-1)}$ (follows from the m.g.f $\psi(t) = e^{\theta(e^t-1)}$).
  - Hence $b_{T_{2n}}(\theta) \to 0$.
  - Using the CLT and the δ-method, we can show that $\sqrt{n}(T_{2n} - \vartheta) \xrightarrow{D} N(0, e^{-2\theta}\theta)$.
  - By Definition 2.12(i), $\text{AMSE}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$.
- Thus, the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n} is

  $$e_{T_{1n},T_{2n}}(\theta) = \theta/(e^\theta - 1) < 1$$

  This shows that T_{2n} is asymptotically more efficient than T_{1n}.

The result for T_{2n} in Example 2.36 is a special case (with $U_n = \bar{X}$) of the following general result.

Theorem 7.4 (Theorem 2.6). *Let g be a function on $\mathcal{R}^k$ that is differentiable at $\theta \in \mathcal{R}^k$ and let U_n be a k-vector of statistics satisfying $a_n(U_n - \theta) \xrightarrow{D} Y$ for a random k-vector Y with $0 < E\|Y\|^2 < \infty$ and a sequence of positive numbers $\{a_n\}$ satisfying $a_n \to \infty$. Let $T_n = g(U_n)$ be an estimator of $\vartheta = g(\theta)$. Then, the amse and asymptotic variance of T_n are, respectively,*

$$\text{AMSE}_{T_n}(P) = E\{[\nabla g(\theta)]^\top Y\}^2/a_n^2$$

and

$$\sigma^2_{T_n}(P) = [\nabla g(\theta)]^\top \text{Var}(Y) \nabla g(\theta)/a_n^2.$$

7.2 Method of moment estimators

- An exactly unbiased estimator may not exist, or is hard to obtain. We often derive asymptotically unbiased estimators.
- The method of moments is the oldest method of deriving asymptotically unbiased estimators, although they may not be the best estimators.
- Consider a parametric problem where $X_1, ..., X_n$ are i.i.d. random variables from P_θ, $\theta \in \Theta \subset \mathcal{R}^k$, and $E|X_1|^k < \infty$. Let $\mu_j = EX_1^j$ be the jth moment of P and let

  $$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$$

  be the jth *sample moment*, which is an unbiased estimator of μ_j, $j = 1, ..., k$.
- Typically,

  $$\mu_j = h_j(\theta), \qquad j = 1, ..., k, \tag{7.27}$$

  for some functions h_j on $\mathcal{R}^k$. By substituting μ_j's on the left-hand side of (7.27) by the sample moments $\hat{\mu}_j$, we obtain a *moment estimator* $\hat{\theta}$, i.e., $\hat{\theta}$ satisfies

  $$\hat{\mu}_j = h_j(\hat{\theta}), \qquad j = 1, ..., k,$$

  which is a sample analogue of (7.27). This method of deriving estimators is called the *method of moments*.
- Let $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_k)$ and $h = (h_1, ..., h_k)$. Then $\hat{\mu} = h(\hat{\theta})$. If the inverse function h^{-1} exists, then the unique moment estimator of θ is $\hat{\theta} = h^{-1}(\hat{\mu})$.
- When h^{-1} does not exist (i.e., h is not one-to-one), any solution of $\hat{\mu} = h(\hat{\theta})$ is a moment estimator of θ. If possible, we always choose a solution $\hat{\theta}$ in the parameter space Θ. In some cases, however, a moment estimator does not exist (see Exercise 111).

- Assume that $\hat{\theta} = g(\hat{\mu})$ for a function g. If h^{-1} exists, then $g = h^{-1}$. If g is continuous at $\mu = (\mu_1, ..., \mu_k)$, then $\hat{\theta}$ is strongly consistent for θ, since $\hat{\mu}_j \xrightarrow{a.s.} \mu_j$ by the SLLN. If g is differentiable at μ and $E|X_1|^{2k} < \infty$, then $\hat{\theta}$ is asymptotically normal, by the CLT and Theorem 1.12, and

$$\text{AMSE}_{\hat{\theta}}(\theta) = n^{-1}[\nabla g(\mu)]^\top V_\mu \nabla g(\mu),$$

where V_μ is a $k \times k$ matrix whose (i,j)th element is $\mu_{i+j} - \mu_i \mu_j$.

Example 7.5 (Example 3.24)**.** Let $X_1, ..., X_n$ be i.i.d. from a population P_θ indexed by the parameter $\theta = (\mu, \sigma^2)$, where $\mu = EX_1 \in \mathcal{R}$ and $\sigma^2 = \text{Var}(X_1) \in (0, \infty)$. This includes cases such as the family of normal distributions, double exponential distributions, or logistic distributions (Table 1.2, page 20).

- Since $EX_1 = \mu$ and $EX_1^2 = \text{Var}(X_1) + (EX_1)^2 = \sigma^2 + \mu^2$, setting $\hat{\mu}_1 = \mu$ and $\hat{\mu}_2 = \sigma^2 + \mu^2$ we obtain the moment estimator

$$\hat{\theta} = \left(\bar{X},\ \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \right) = \left(\bar{X},\ \frac{n-1}{n} S^2 \right).$$

  Note that $\bar{X}$ is unbiased, but $\frac{n-1}{n} S^2$ is not.

- If X_i is normal, then $\hat{\theta}$ is sufficient and is nearly the same as an optimal estimator such as the UMVUE. On the other hand, if X_i is from a double exponential or logistic distribution, then $\hat{\theta}$ is not sufficient and can often be improved.

- Consider now the estimation of σ^2 when we know that $\mu = 0$.
  - Obviously we cannot use the equation $\hat{\mu}_1 = \mu$ to solve the problem.
  - Using $\mu_2 = \sigma^2$, we obtain the moment estimator

$$\hat{\sigma}^2 = \hat{\mu}_2 = n^{-1} \sum_{i=1}^n X_i^2.$$

  - This is still a good estimator when X_i is normal, but is not a function of sufficient statistic when X_i is from a double exponential distribution.
  - For the double exponential case one can argue that we should first make a transformation $Y_i = |X_i|$ and then obtain the moment estimator based on the transformed data. The moment estimator of σ^2 based on the transformed data is

$$\bar{Y}^2 = \left(\frac{1}{n} \sum_{i=1}^n |X_i| \right)^2,$$

    which is sufficient for σ^2. Note that this estimator can also be obtained based on absolute moment equations.

Example 7.6 (Example 3.25)**.** Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on (θ_1, θ_2), $-\infty < \theta_1 < \theta_2 < \infty$.

- Note that
$$EX_1 = (\theta_1 + \theta_2)/2 \quad \text{and} \quad EX_1^2 = (\theta_1^2 + \theta_2^2 + \theta_1\theta_2)/3.$$
- Setting $\hat{\mu}_1 = EX_1$ and $\hat{\mu}_2 = EX_1^2$ and substituting θ_1 in the second equation by $2\hat{\mu}_1 - \theta_2$ (the first equation), we obtain that
$$(2\hat{\mu}_1 - \theta_2)^2 + \theta_2^2 + (2\hat{\mu}_1 - \theta_2)\theta_2 = 3\hat{\mu}_2,$$
which is the same as
$$(\theta_2 - \hat{\mu}_1)^2 = 3(\hat{\mu}_2 - \hat{\mu}_1^2).$$
- Since $\theta_2 > EX_1$, we obtain that
$$\hat{\theta}_2 = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} + \sqrt{\tfrac{3(n-1)}{n}S^2}$$
$$\hat{\theta}_1 = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} - \sqrt{\tfrac{3(n-1)}{n}S^2}.$$
- These estimators are not functions of the sufficient and complete statistic $(X_{(1)}, X_{(n)})$.

Example 7.7 (Example 3.26)**.** Let $X_1, ..., X_n$ be i.i.d. from the binomial distribution $Bi(p, k)$ with unknown parameters $k \in \{1, 2, ...\}$ and $p \in (0, 1)$.

- Since
$$EX_1 = kp$$
and
$$EX_1^2 = kp(1-p) + k^2p^2,$$
we obtain the moment estimators
$$\hat{p} = (\hat{\mu}_1 + \hat{\mu}_1^2 - \hat{\mu}_2)/\hat{\mu}_1 = 1 - \tfrac{n-1}{n}S^2/\bar{X}$$
and
$$\hat{k} = \hat{\mu}_1^2/(\hat{\mu}_1 + \hat{\mu}_1^2 - \hat{\mu}_2) = \bar{X}/(1 - \tfrac{n-1}{n}S^2/\bar{X}).$$
- The estimator $\hat{p}$ is in the range of $(0, 1)$. But $\hat{k}$ may not be an integer. It can be improved by an estimator that is $\hat{k}$ rounded to the nearest positive integer.

7.3 Weighted LSE

In the linear model $X = Z\beta + \epsilon$, the unbiased LSE of $l^\top\beta$ may be improved by a slightly biased estimator when $V = \text{Var}(\epsilon)$ is not $\sigma^2 I_n$ and the LSE is not BLUE.

- Assume that Z is of full rank so that every $l^\top\beta$ is estimable.

- If V is known, then the BLUE of $l^\top\beta$ is $l^\top\breve{\beta}$, where

$$\breve{\beta} = (Z^\top V^{-1} Z)^{-1} Z^\top V^{-1} X \tag{7.28}$$

- If V is unknown and $\hat{V}$ is an estimator of V, then an application of the substitution principle leads to a *weighted least squares estimator*

$$\hat{\beta}_w = (Z^\top \hat{V}^{-1} Z)^{-1} Z^\top \hat{V}^{-1} X. \tag{7.29}$$

- The weighted LSE is not linear in X and not necessarily unbiased for β. If the distribution of ϵ is symmetric about 0 and $\hat{V}$ remains unchanged when ϵ changes to $-\epsilon$, then the distribution of $\hat{\beta}_w - \beta$ is symmetric about 0 and, if $E\hat{\beta}_w$ is well defined, $\hat{\beta}_w$ is unbiased for β.

- If the weighted LSE $l^\top\hat{\beta}_w$ is unbiased, then it may be a better estimator than the LSE $l^\top\hat{\beta}$, since $\mathrm{Var}(l^\top\hat{\beta}_w)$ may be smaller than $\mathrm{Var}(l^\top\hat{\beta})$.

- Asymptotic properties of the weighted LSE depend on the asymptotic behavior of $\hat{V}$. We say that $\hat{V}$ is consistent for V iff

$$\|\hat{V}^{-1}V - I_n\|_{\max} \xrightarrow{P} 0, \tag{7.30}$$

  where $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ for a matrix $A = (a_{ij})$.

Theorem 7.8 (Theorem 3.17)**.** *Consider model $X = Z\beta + \epsilon$ with a full rank Z. Let $\breve{\beta}$ and $\hat{\beta}_w$ be defined by (7.28) and (7.29), respectively, with a $\hat{V}$ consistent in the sense of (7.30). Under the conditions in Theorem 3.12,*

$$l^\top(\hat{\beta}_w - \beta)/a_n \xrightarrow{D} N(0,1),$$

where $l \in \mathcal{R}^p$, $l \neq 0$, and

$$a_n^2 = \mathrm{Var}(l^\top\breve{\beta}) = l^\top(Z^\top V^{-1} Z)^{-1} l.$$

Proof:

Using the same argument as in the proof of Theorem 3.12, we obtain that

$$l^\top(\breve{\beta} - \beta)/a_n \xrightarrow{D} N(0,1).$$

By Slutsky's theorem, the result follows from

$$l^\top\hat{\beta}_w - l^\top\breve{\beta} = o_p(a_n).$$

Define

$$\xi_n = l^\top(Z^\top\hat{V}^{-1}Z)^{-1}Z^\top(\hat{V}^{-1} - V^{-1})\epsilon$$

and

$$\zeta_n = l^\top[(Z^\top\hat{V}^{-1}Z)^{-1} - (Z^\top V^{-1}Z)^{-1}]Z^\top V^{-1}\epsilon.$$

Then

$$l^\top\hat{\beta}_w - l^\top\breve{\beta} = \xi_n + \zeta_n.$$

The result follows from $\xi_n = o_p(a_n)$ and $\zeta_n = o_p(a_n)$ (details are in the textbook). □

- Theorem 3.17 shows that as long as $\hat{V}$ is consistent in the sense of (7.30), the weighted LSE $\hat{\beta}_w$ is asymptotically as efficient as $\breve{\beta}$, which is the BLUE if V is known.
- By Theorems 3.12 and 3.17, the asymptotic relative efficiency of the LSE $l^\top\hat{\beta}$ w.r.t. the weighted LSE $l^\top\hat{\beta}_w$ is
$$\frac{l^\top(Z^\top V^{-1}Z)^{-1}l}{l^\top(Z^\top Z)^{-1}Z^\top VZ(Z^\top Z)^{-1}l} \tag{7.31}$$
which is always less than 1 and equals 1 if $l^\top\hat{\beta}$ is a BLUE (in which case $\hat{\beta}=\breve{\beta}$).
  - To see so, we note that $l=(Z^\top Z)\zeta$ as $l\in\mathcal{R}(Z^\top Z)$. So (7.31) becomes
$$\frac{\zeta^\top(Z^\top Z)(Z^\top V^{-1}Z)^{-1}(Z^\top Z)\zeta}{\zeta^\top Z^\top VZ\zeta}$$
  - Let $\gamma=Z\zeta$, then it becomes
$$\frac{\gamma^\top Z(Z^\top V^{-1}Z)^{-1}Z^\top\gamma}{\gamma^\top V\gamma}$$
  - Let $\eta=V^{1/2}\gamma$ (so that $\gamma=V^{1/2}\eta$) it becomes
$$\frac{\eta^\top V^{-1/2}Z(Z^\top V^{-1}Z)^{-1}Z^\top V^{-1/2}\eta}{\eta^\top\eta}=\frac{\eta^\top A(A^\top A)^{-1}A^\top\eta}{\eta^\top\eta}$$
where $A=V^{-1/2}Z$.
  - Note that the matrix $A(A^\top A)^{-1}A^\top$ is symmetric and positive-definite, due to the full rank of Z and V.
  - Thus, $\eta^\top A(A^\top A)^{-1}A^\top\eta\le$ largest eigenvalue of $A(A^\top A)^{-1}A^\top$.
  - The eigen-equation is $A(A^\top A)^{-1}A^\top v=\lambda v$ for eigenvector v and eigenvalue λ, which implies $A^\top v=\lambda A^\top v$. This is possible only if $\lambda=1$. (note that $\lambda>0$ and $v\neq 0$)
  - So, the largest eigenvalue of $A(A^\top A)^{-1}A^\top$ is 1.
  - Indeed, $Q=A(A^\top A)^{-1}A^\top$ is idempotent, i.e., $Q^2=Q$. The eigenvalues of an idempotent matrix is either 0 or 1.
- Finding a consistent $\hat{V}$ is possible when V has a certain type of structure, see Examples 3.29, 3.30 and 3.31 in the text book.

7.4 V-statistics

Let $X_1,...,X_n$ be i.i.d. from P. For every U-statistic U_n as an estimator of $\vartheta=E[h(X_1,...,X_m)]$, there is a closely related *V-statistic* defined by
$$V_n=\frac{1}{n^m}\sum_{i_1=1}^{n}\cdots\sum_{i_m=1}^{n}h(X_{i_1},...,X_{i_m}). \tag{7.32}$$
As an estimator of ϑ, V_n is biased; but the bias is small asymptotically as the following results show. For a fixed sample size n, V_n may be better than U_n in terms of their mse's.

Proposition 7.9 (Proposition 3.5)**.** *Let V_n be defined by (7.32).*

(i) Assume that $E|h(X_{i_1}, ..., X_{i_m})| < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the bias of V_n satisfies

$$b_{V_n}(P) = O(n^{-1}).$$

(ii) Assume that $E[h(X_{i_1}, ..., X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the variance of V_n satisfies

$$\mathrm{Var}(V_n) = \mathrm{Var}(U_n) + O(n^{-2}),$$

where U_n is the U-statistic corresponding to V_n.

Theorem 7.10 (Theorem 3.16)**.** *Let V_n be given by (7.32) with $E[h(X_{i_1}, ..., X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.*

(i) If $\zeta_1 = \mathrm{Var}(h_1(X_1)) > 0$, then

$$\sqrt{n}(V_n - \vartheta) \xrightarrow{D} N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 = \mathrm{Var}(h_2(X_1, X_2)) > 0$, then

$$n(V_n - \vartheta) \xrightarrow{D} \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2,$$

where χ_{1j}^2's and λ_j's are the same as those in Theorem 3.5.

- Note that the asymptotic distribution of the corresponding U-statistic in (ii) is

$$n[U_n - \vartheta] \xrightarrow{D} \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1).$$

- Theorem 3.16 shows that
  - if $\zeta_1 > 0$, then the amse's of U_n and V_n are the same.
  - If $\zeta_1 = 0$ but $\zeta_2 > 0$, then

$$\begin{aligned} \mathrm{AMSE}_{V_n}(P) &= \frac{m^2(m-1)^2\zeta_2}{2n^2} + \frac{m^2(m-1)^2}{4n^2} \left(\sum_{j=1}^{\infty} \lambda_j \right)^2 \\ &= \mathrm{AMSE}_{U_n}(P) + \frac{m^2(m-1)^2}{4n^2} \left(\sum_{j=1}^{\infty} \lambda_j \right)^2 \end{aligned}$$

Hence U_n is asymptotically more efficient than V_n, unless $\sum_{j=1}^{\infty} \lambda_j = 0$.

7.5 Maximum likelihood estimators

The *maximum likelihood method* is the most popular method for deriving estimators in statistical inference.

Definition 7.11. Let $X \in \mathcal{X}$ be a sample with a p.d.f. f_θ w.r.t. a σ-finite measure ν, where $\theta \in \Theta \subset \mathcal{R}^k$.

(i) For each $x \in \mathcal{X}$, $f_\theta(x)$ considered as a function of θ is called the *likelihood function* and denoted by $\ell(\theta)$.

(ii) Let $\bar{\Theta}$ be the closure of Θ. A $\hat{\theta} \in \bar{\Theta}$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} \ell(\theta)$ is called a *maximum likelihood estimate* (MLE) of θ. If $\hat{\theta}$ is a Borel function of X a.e. ν, then $\hat{\theta}$ is called a *maximum likelihood estimator* (MLE) of θ.

(iii) Let g be a Borel function from Θ to $\mathcal{R}^p$, $p \leq k$. If g is not one-to-one and $\hat{\theta}$ is an MLE of θ, then $\hat{\vartheta} = g(\hat{\theta})$ is defined to be an MLE of $\vartheta = g(\theta)$. [If g is one-to-one, then $\hat{\vartheta} = g(\hat{\theta})$, which is referred to as the invariant property of MLE].

- Note that $\bar{\Theta}$ instead of Θ is used in the definition of an MLE.
  This is because a maximum of $\ell(\theta)$ may not exist when Θ is an open set.
  In some textbooks, Θ is used, instead of $\bar{\Theta}$
- There may be multiple MLE's.
- An MLE may not have an explicit form.
- In terms of their mse's, MLE's are not necessarily better than UMVUE's.
- MLE's are frequently inadmissible. This is not surprising, since MLE's are not derived under any given loss function.
- The main theoretical justification for MLE's is provided in the theory of asymptotic efficiency considered later.
- If Θ contains finitely many points, an MLE exists and can always be obtained by comparing finitely many values $\ell(\theta)$, $\theta \in \Theta$.
- The log-likelihood function $\log \ell(\theta)$ is often more convenient to work with.
- If $\ell(\theta)$ is differentiable on Θ°, the interior of Θ, then possible candidates for MLE's are the values of $\theta \in \Theta^\circ$ satisfying the *likelihood equation* $\frac{\partial \log \ell(\theta)}{\partial \theta} = 0$. A root of the likelihood equation may be local or global minima or maxima, or simply stationary points. Extrema may also occur at the boundary of Θ or when $\|\theta\| \to \infty$.
- Furthermore, if $\ell(\theta)$ is not differentiable, then extrema may occur at nondifferentiable or discontinuity points of $\ell(\theta)$. Hence, it is important to analyze the entire likelihood function to find its maxima.

Example 7.12 (Example 3.3)**.** Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in \Theta = (0, 1)$. When $(X_1, ..., X_n) = (x_1, ..., x_n)$ is observed, the likelihood function is

$$\ell(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})},$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$.

- Note that $\bar{\Theta} = [0, 1]$ and $\Theta^\circ = \Theta$.
- The likelihood equation is
$$\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0.$$
- If $0 < \bar{x} < 1$, then this equation has a unique solution $\bar{x}$. The second-order derivative of $\log \ell(p)$ is
$$-\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2},$$
which is always negative. When p tends to 0 or 1 (the boundary of Θ), $\ell(p) \to 0$. Thus, $\bar{x}$ is the unique MLE of p.
- When $\bar{x} = 0$, $\ell(p) = (1-p)^n$ is a strictly decreasing function of p and, therefore, its unique maximum is 0. Similarly, the MLE is 1 when $\bar{x} = 1$.
- Combining these results, we conclude that the MLE of p is $\bar{x}$. When $\bar{x} = 0$ or 1, a maximum of $\ell(p)$ does not exist on $\Theta = (0, 1)$, although $\sup_{p\in(0,1)} \ell(p) = 1$; the MLE takes a value outside of Θ and, hence, is not a reasonable estimator.
- However, if $p \in (0, 1)$, the probability that $\bar{x} = 0$ or 1 tends to 0 quickly as $n \to \infty$.

Example 3.3 indicates that, for small n, a maximum of $\ell(\theta)$ may not exist on Θ and an MLE may be an unreasonable estimator; however, this is unlikely to occur when n is large. A rigorous result of this sort will be given later, where we study asymptotic properties of MLE's.

Example 7.13 (Example 3.4)**.** Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$, $n \geq 2$. Consider first the case where $\Theta = \mathcal{R} \times (0, \infty)$.

- The log-likelihood function
$$\log \ell(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi).$$
- The likelihood equation is
$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \qquad \text{and} \qquad \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{\sigma^2} = 0.$$

- Solving the equations, we obtain $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$ where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$, $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$.
- To show that $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$ is an MLE, first note that Θ is an open set and $\ell(\theta)$ is differentiable everywhere; as θ tends to the boundary of Θ or $\|\theta\| \to \infty$, $\ell(\theta)$ tends to 0; and

$$\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^n (x_i - \mu) & \frac{1}{\sigma^6}\sum_{i=1}^n (x_i - \mu)^2 - \frac{n}{2\sigma^4} \end{pmatrix}$$

  This matrix is negative definite when $\mu = \bar{x}$ and $\sigma^2 = \hat{\sigma}^2$. Hence $\hat{\theta}$ is the unique MLE.
- Sometimes we can avoid the calculation of the second-order derivatives. For instance, in this example we know that $\ell(\theta)$ is bounded and $\ell(\theta) \to 0$ as $\|\theta\| \to \infty$ or θ tends to the boundary of Θ; hence the unique solution to the likelihood equation must be the MLE.

Consider next the case where $\Theta = (0, \infty) \times (0, \infty)$, i.e., μ is known to be positive.

- The likelihood function is differentiable on $\Theta^\circ = \Theta$ and $\bar{\Theta} = [0, \infty) \times [0, \infty)$.
- If $\bar{x} > 0$, then the same argument for the previous case can be used to show that $(\bar{x}, \hat{\sigma}^2)$ is the MLE.
- If $\bar{x} \leq 0$, then the first equation in the likelihood equation does not have a solution in Θ. However, the function $\log \ell(\theta) = \log \ell(\mu, \sigma^2)$ is strictly decreasing in μ for any fixed σ^2.
  - Hence, a maximum of $\log \ell(\mu, \sigma^2)$ is $\mu = 0$, which does not depend on σ^2.
  - Then, the MLE is $(0, \tilde{\sigma}^2)$, where $\tilde{\sigma}^2$ is the value maximizing $\log \ell(0, \sigma^2)$ over $\sigma^2 \geq 0$.
  - Maximizing $\log \ell(0, \sigma^2)$ leads to $\tilde{\sigma}^2 = n^{-1}\sum_{i=1}^n x_i^2$.
  - Thus, the MLE is

$$\hat{\theta} = \begin{cases} (\bar{x}, \hat{\sigma}^2) & \bar{x} > 0 \\ (0, \tilde{\sigma}^2) & \bar{x} \leq 0. \end{cases}$$

  - Again, the MLE in this case is not in Θ if $\bar{x} \leq 0$: a maximum of $\ell(\theta)$ does not exist on Θ when $\bar{x} \leq 0$.

Example 7.14 (Example 3.5)**.** Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on an interval $\mathcal{I}_\theta$ with an unknown θ.

- First, consider the case where $\mathcal{I}_\theta = (0, \theta)$ and $\theta > 0$, $\Theta^\circ = (0, \infty)$. The likelihood function is

$$\ell(\theta) = \theta^{-n} I_{(x_{(n)}, \infty)}(\theta),$$

  where $x_{(n)} = \max(x_1, ..., x_n)$. On $(0, x_{(n)})$, $\ell \equiv 0$ and on $(x_{(n)}, \infty)$, $\ell'(\theta) = -n\theta^{n-1} < 0$ for all θ. $\ell(\theta)$ is not differentiable at $x_{(n)}$ and the method of using the likelihood equation is not applicable. Since $\ell(\theta)$ is strictly decreasing on $(x_{(n)}, \infty)$ and is 0 on $(0, x_{(n)})$, a unique maximum of $\ell(\theta)$ is $x_{(n)}$, which is a discontinuity point of $\ell(\theta)$. This shows that the MLE of θ is $X_{(n)}$.

- Next, consider the case where $\mathcal{I}_\theta = (\theta - \frac{1}{2}, \theta + \frac{1}{2})$ with $\theta \in \mathcal{R}$. The likelihood function is

$$\ell(\theta) = I_{(x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2})}(\theta),$$

  where $x_{(1)} = \min(x_1, ..., x_n)$. Again, the method of using the likelihood equation is not applicable. However, it follows from Definition 4.3 that any statistic $T(X)$ satisfying $x_{(n)} - \frac{1}{2} \le T(x) \le x_{(1)} + \frac{1}{2}$ is an MLE of θ.
  This example indicates that MLE's may not be unique and can be unreasonable.

Example 7.15 (Example 3.6)**.** Let X be an observation from the hypergeometric distribution $HG(r, n, \theta - n)$ (Table 1.1, page 18) with known r, n, and an unknown $\theta = n+1, n+2, ...$ In this case, the likelihood function is defined on integers and the method of using the likelihood equation is certainly not applicable. Note that

$$\frac{\ell(\theta)}{\ell(\theta - 1)} = \frac{(\theta - r)(\theta - n)}{\theta(\theta - n - r + x)},$$

which is larger than 1 if and only if $\theta < rn/x$ and is smaller than 1 if and only if $\theta > rn/x$. Thus, $\ell(\theta)$ has a maximum $\theta =$ the integer part of rn/x, which is the MLE of θ.

In applications, MLE's typically do not have analytic forms and some numerical methods have to be used to compute MLE's.

- The Newton-Raphson iteration method repeatedly computes

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \left[\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top} \bigg|_{\theta=\hat{\theta}^{(t)}} \right]^{-1} \frac{\partial \log \ell(\theta)}{\partial\theta} \bigg|_{\theta=\hat{\theta}^{(t)}},$$

  $t = 0, 1, ...,$ where $\hat{\theta}^{(0)}$ is an initial value and $\partial^2 \log \ell(\theta)/\partial\theta\partial\theta^\top$ is assumed of full rank for every $\theta \in \Theta$.
- If $\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top}\Big|_{\theta=\hat{\theta}^{(t)}}$ is replaced by by $\left\{ E\left(\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\top} \right) \right\} \Big|_{\theta=\hat{\theta}^{(t)}}$, where the expectation is taken under P_θ, then the method is known as the Fisher-scoring method.
- If the iteration converges, then $\hat{\theta}^{(\infty)}$ or $\hat{\theta}^{(t)}$ with a sufficiently large t is a numerical approximation to a solution of the likelihood equation.

Suppose that X has a distribution from a natural exponential family so that the likelihood function is

$$\ell(\eta) = \exp\{\eta^\top T(x) - \zeta(\eta)\} h(x),$$

where $\eta \in \Xi$ is a vector of unknown parameters.

- The likelihood equation is then

$$\frac{\partial \log \ell(\eta)}{\partial\eta} = T(x) - \frac{\partial\zeta(\eta)}{\partial\eta} = 0,$$

  which has a unique solution $T(x) = \partial\zeta(\eta)/\partial\eta$, assuming that $T(x)$ is in the range of $\partial\zeta(\eta)/\partial\eta$.

- Note that
$$\frac{\partial^2 \log \ell(\eta)}{\partial \eta \partial \eta^\top} = -\frac{\partial^2 \zeta(\eta)}{\partial \eta \partial \eta^\top} = -\mathrm{Var}(T)$$
Since $\mathrm{Var}(T)$ is positive definite, $-\log \ell(\eta)$ is convex in η and $T(x)$ is the unique MLE of the parameter $\mu(\eta) = \partial \zeta(\eta)/\partial \eta$.

- Also, the function $\mu(\eta)$ is one-to-one so that μ^{-1} exists. By the definition, the MLE of η is $\hat{\eta} = \mu^{-1}(T(x))$.

- If the distribution of X is in a general exponential family and the likelihood function is
$$\ell(\theta) = \exp\{[\eta(\theta)]^\top T(x) - \xi(\theta)\} h(x),$$
then the MLE of θ is $\hat{\theta} = \eta^{-1}(\hat{\eta})$, if η^{-1} exists and $\hat{\eta}$ is in the range of $\eta(\theta)$.

- Of course, $\hat{\theta}$ is also the solution of the likelihood equation
$$\frac{\partial \log \ell(\theta)}{\partial \theta} = \frac{\partial \eta(\theta)}{\partial \theta} T(x) - \frac{\partial \xi(\theta)}{\partial \theta} = 0.$$

7.6 Asymptotic properties of MLE's

- Let $\{\hat{\theta}_n\}$ be a sequence of estimators of θ based on a sequence of samples $\{X = (X_1, ..., X_n) : n = 1, 2, ...\}$.

- Suppose that as $n \to \infty$, $\hat{\theta}_n$ is asymptotically normal (AN) in the sense that
$$[V_n(\theta)]^{-1/2}(\hat{\theta}_n - \theta) \xrightarrow{D} N_k(0, I_k),$$
where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix depending on θ.

- If θ is one-dimensional ($k = 1$), then $V_n(\theta)$ is the asymptotic variance as well as the amse of $\hat{\theta}_n$ (§2.5.2).

- When $k > 1$, $V_n(\theta)$ is called the *asymptotic covariance matrix* of $\hat{\theta}_n$ and can be used as a measure of asymptotic performance of estimators.

- If $\hat{\theta}_{jn}$ is AN with asymptotic covariance matrix $V_{jn}(\theta)$, $j = 1, 2$, and $V_{1n}(\theta) \leq V_{2n}(\theta)$ (in the sense that $V_{2n}(\theta) - V_{1n}(\theta)$ is nonnegative definite) for all $\theta \in \Theta$, then $\hat{\theta}_{1n}$ is said to be asymptotically more efficient than $\hat{\theta}_{2n}$.

- Since the asymptotic covariance matrices are unique only in the limiting sense, we have to make our comparison based on their limits.

- When X_i's are i.i.d., $V_n(\theta)$ is usually of the form $n^{-\delta} V(\theta)$ for some $\delta > 0$ (= 1 in the majority of cases) and a positive definite matrix $V(\theta)$ that does not depend on n.

Information inequality:

- If $\hat{\theta}_n$ is AN, it is asymptotically unbiased.
- If $V_n(\theta) = \mathrm{Var}(\hat{\theta}_n)$, then, under some regularity conditions, it follows from Theorem 3.3 that we have the following information inequality

$$V_n(\theta) \geq [I_n(\theta)]^{-1},$$

  where, for every n, $I_n(\theta)$ is the Fisher information matrix for X of size n. The information inequality may lead to an optimal estimator.
- When $V_n(\theta)$ is an asymptotic covariance matrix, the information inequality may not hold (even in the limiting sense), even if the regularity conditions in Theorem 3.3 are satisfied.

Example 7.16 (Example 4.38)**.** Let $X_1, ..., X_n$ be i.i.d. from $N(\theta, 1)$, $\theta \in \mathcal{R}$. Then $I_n(\theta) = n$. For a fixed constant t, define

$$\hat{\theta}_n = \begin{cases} \bar{X} & |\bar{X}| \geq n^{-1/4} \\ t\bar{X} & |\bar{X}| < n^{-1/4}, \end{cases}$$

By Proposition 3.2, all conditions in Theorem 3.3 are satisfied. It can be shown (by using Slutsky's theorem) that $\hat{\theta}_n$ is AN with $V_n(\theta) = V(\theta)/n$, where $V(\theta) = 1$ if $\theta \neq 0$ and $V(\theta) = t^2$ if $\theta = 0$. If $t^2 < 1$, the information inequality does not hold when $\theta = 0$.

Theorem 7.17 (Theorem 4.16)**.** *Let $X_1, ..., X_n$ be i.i.d. from a p.d.f. f_θ w.r.t. a σ-finite measure ν on $(\mathcal{R}, \mathcal{B})$, where $\theta \in \Theta$ and Θ is an open set in $\mathcal{R}^k$. Suppose that for every x in the range of X_1, $f_\theta(x)$ is twice continuously differentiable in θ and satisfies*

$$\frac{\partial}{\partial \theta} \int \psi_\theta(x) d\nu = \int \frac{\partial}{\partial \theta} \psi_\theta(x) d\nu$$

for $\psi_\theta(x) = f_\theta(x)$ and $= \partial f_\theta(x)/\partial\theta$; the Fisher information matrix

$$I_1(\theta) = E\left\{ \frac{\partial}{\partial \theta} \log f_\theta(X_1) \left[\frac{\partial}{\partial \theta} \log f_\theta(X_1) \right]^\top \right\}$$

is positive definite; and for any given $\theta \in \Theta$, there exists a positive number c_θ and a positive function h_θ such that $E[h_\theta(X_1)] < \infty$ and

$$\sup_{\gamma : \|\gamma - \theta\| < c_\theta} \left\| \frac{\partial^2 \log f_\gamma(x)}{\partial \gamma \partial \gamma^\top} \right\| \leq h_\theta(x)$$

for all x in the range of X_1, where $\|A\| = \sqrt{tr(A^\top A)}$ for any matrix A.

If $\hat{\theta}_n$ is an estimator of θ (based on $X_1, ..., X_n$) and is AN with $V_n(\theta) = V(\theta)/n$, then there is a $\Theta_0 \subset \Theta$ with Lebesgue measure 0 such that the information inequality holds if $\theta \notin \Theta_0$.

- Points at which the information inequality does not hold are called points of superefficiency.

- Motivated by the fact that the set of superefficiency points is of Lebesgue measure 0 under regularity conditions, we have the following definition.

Definition 7.18 (Asymptotic efficiency)**.** Assume that the Fisher information matrix $I_n(\theta)$ is well defined and positive definite for every n. A sequence of estimators $\{\hat{\theta}_n\}$ that is AN is said to be *asymptotically efficient* or *asymptotically optimal* if and only if $V_n(\theta) = [I_n(\theta)]^{-1}$.

- Suppose that we are interested in estimating $\vartheta = g(\theta)$, where g is a differentiable function from Θ to $\mathcal{R}^p$, $1 \le p \le k$.
- If $\hat{\theta}_n$ is AN, then, by Theorem 1.12(i), $\hat{\vartheta}_n = g(\hat{\theta}_n)$ is asymptotically distributed as $N_p(\vartheta, [\nabla g(\theta)]^\top V_n(\theta) \nabla g(\theta))$.
- Thus, the information inequality becomes

$$[\nabla g(\theta)]^\top V_n(\theta) \nabla g(\theta) \ge [\tilde{I}_n(\vartheta)]^{-1},$$

  where $\tilde{I}_n(\vartheta)$ is the Fisher information matrix about ϑ contained in X.
- If $p = k$ and g is one-to-one, then

$$[\tilde{I}_n(\vartheta)]^{-1} = [\nabla g(\theta)]^\top [I_n(\theta)]^{-1} \nabla g(\theta)$$

  and, therefore, $\hat{\vartheta}_n$ is asymptotically efficient if and only if $\hat{\theta}_n$ is asymptotically efficient.
- For this reason, in the case of $p < k$, $\hat{\vartheta}_n$ is considered to be asymptotically efficient if and only if $\hat{\theta}_n$ is asymptotically efficient, and we can focus on the estimation of θ only.

Under some regularity conditions, a root of the likelihood equation (RLE), which is a candidate for an MLE, is asymptotically efficient.

Theorem 7.19 (Theorem 4.17)**.** *Assume the conditions of Theorem 4.16.*
(i) There is a sequence of estimators $\{\hat{\theta}_n\}$ such that

$$P\big(s_n(\hat{\theta}_n) = 0\big) \to 1 \qquad \text{and} \qquad \hat{\theta}_n \xrightarrow{P} \theta,$$

where $s_n(\gamma) = \partial \log \ell(\gamma)/\partial\gamma$.
(ii) Any consistent sequence $\tilde{\theta}_n$ of RLE's is asymptotically efficient.

Proof:

(i) Let $B_n(c) = \{\gamma : \|[I_n(\theta)]^{1/2}(\gamma - \theta)\| \le c\}$ for $c > 0$. Since Θ is open, for each $c > 0$, $B_n(c) \subset \Theta$ for sufficiently large n. Since $B_n(c)$ shrinks to $\{\theta\}$ as $n \to \infty$, the asymptotic existence of $\hat{\theta}_n$ is implied by the fact that for any $\epsilon > 0$, there exists $n_0 > 1$ such that

$$P\big(\log \ell(\gamma) - \log \ell(\theta) < 0 \quad \text{for all } \gamma \in \partial B_n(c)\big) \ge 1 - \epsilon, \qquad n \ge n_0, \tag{7.33}$$

where $c = 4\sqrt{2k/\epsilon}$ and $\partial B_n(c)$ is the boundary of $B_n(c)$. The measurability of $\hat{\theta}_n$ can be safely assumed; see Serfling (1980, p147).

For $\gamma \in \partial B_n(c)$, the Taylor expansion gives

$$\begin{aligned} \log \ell(\gamma) - \log \ell(\theta) &= c\lambda^\top [I_n(\theta)]^{-1/2} s_n(\theta) \\ &\quad + (c^2/2)\lambda^\top [I_n(\theta)]^{-1/2} \nabla s_n(\gamma^*) [I_n(\theta)]^{-1/2} \lambda, \end{aligned} \tag{7.34}$$

where $\lambda = [I_n(\theta)]^{1/2}(\gamma - \theta)/c$ satisfying $\|\lambda\| = 1$, $\nabla s_n(\gamma) = \partial s_n(\gamma)/\partial\gamma$, and γ^* lies between γ and θ.

Note that

$$\begin{aligned} E\frac{\|\nabla s_n(\gamma^*) - \nabla s_n(\theta)\|}{n} &\le E \max_{\gamma \in B_n(c)} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} \\ &\le E \max_{\gamma \in B_n(c)} \left\| \frac{\partial^2 \log f_\gamma(X_1)}{\partial\gamma\partial\gamma^\top} - \frac{\partial^2 \log f_\theta(X_1)}{\partial\theta\partial\theta^\top} \right\| \\ &\to 0, \end{aligned}$$

which follows from

(a) $\partial^2 \log f_\gamma(x)/\partial\gamma\partial\gamma^\top$ is continuous in a neighborhood of θ for any fixed x;

(b) $B_n(c)$ shrinks to $\{\theta\}$; and

(c) for sufficiently large n,

$$\max_{\gamma \in B_n(c)} \left\| \frac{\partial^2 \log f_\gamma(X_1)}{\partial\gamma\partial\gamma^\top} - \frac{\partial^2 \log f_\theta(X_1)}{\partial\theta\partial\theta^\top} \right\| \le 2h_\theta(X_1)$$

under the regularity condition.

By the SLLN (Theorem 1.13) and Proposition 3.1, $n^{-1}\nabla s_n(\theta) \overset{a.s.}{\to} -I_1(\theta)$ (i.e., $\|n^{-1}\nabla s_n(\theta) + I_1(\theta)\| \overset{a.s.}{\to} 0$). These results, together with (7.34), imply that

$$\log \ell(\gamma) - \log \ell(\theta) = c\lambda^\top [I_n(\theta)]^{-1/2} s_n(\theta) - [1 + o_p(1)]c^2/2. \tag{7.35}$$

Note that $\max_\lambda \{\lambda^\top [I_n(\theta)]^{-1/2} s_n(\theta)\} = \|[I_n(\theta)]^{-1/2} s_n(\theta)\|$. Hence, (7.33) follows from (7.35) and

$$\begin{aligned} P\big(\|[I_n(\theta)]^{-1/2} s_n(\theta)\| < c/4\big) &\ge 1 - (4/c)^2 E\|[I_n(\theta)]^{-1/2} s_n(\theta)\|^2 \\ &= 1 - k(4/c)^2 \;=\; 1 - \epsilon/2 \end{aligned}$$

This completes the proof of (i).

(ii) Let $A_\epsilon = \{\gamma : \|\gamma - \theta\| \le \epsilon\}$ for $\epsilon > 0$. Since Θ is open, $A_\epsilon \subset \Theta$ for sufficiently small ϵ. Let $\{\tilde{\theta}_n\}$ be a sequence of consistent RLE's, i.e., $P(s_n(\tilde{\theta}_n) = 0$ and $\tilde{\theta}_n \in A_\epsilon) \to 1$ for any $\epsilon > 0$. Hence, we can focus on the set on which $s_n(\tilde{\theta}_n) = 0$ and $\tilde{\theta}_n \in A_\epsilon$. Using the mean-value theorem for vector-valued functions, we obtain

$$-s_n(\theta) = \left[\int_0^1 \nabla s_n\big(\theta + t(\tilde{\theta}_n - \theta)\big)dt\right](\tilde{\theta}_n - \theta).$$

Note that

$$\frac{1}{n}\left\| \int_0^1 \nabla s_n\big(\theta + t(\tilde{\theta}_n - \theta)\big)dt - \nabla s_n(\theta)\right\| \le \max_{\gamma \in A_\epsilon} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n}.$$

Using the argument in proving (7.35) and the fact that $P(\tilde{\theta}_n \in A_\epsilon) \to 1$ for arbitrary $\epsilon > 0$, we obtain that

$$\frac{1}{n}\left\| \int_0^1 \nabla s_n\big(\theta + t(\tilde{\theta}_n - \theta)\big)dt - \nabla s_n(\theta)\right\| \xrightarrow{P} 0.$$

Since $n^{-1}\nabla s_n(\theta) \xrightarrow{a.s.} -I_1(\theta)$ and $I_n(\theta) = nI_1(\theta)$,

$$-s_n(\theta) = -I_n(\theta)(\tilde{\theta}_n - \theta) + o_p\big(\|I_n(\theta)(\tilde{\theta}_n - \theta)\|\big).$$

This and Slutsky's theorem (Theorem 1.11) imply that $\sqrt{n}(\tilde{\theta}_n - \theta)$ has the same asymptotic distribution as

$$\sqrt{n}[I_n(\theta)]^{-1}s_n(\theta) = n^{-1/2}[I_1(\theta)]^{-1}s_n(\theta) \xrightarrow{D} N_k\big(0, [I_1(\theta)]^{-1}\big)$$

by the CLT (Corollary 1.2), since $\mathrm{Var}(s_n(\theta)) = I_n(\theta)$. □

- Part (i) is asymptotic existence and consistency.
- If the RLE is unique, then it is consistent and asymptotically efficient, whether or not it is MLE.
- If there are more than one sequences of RLE, the theorem does not tell which one is consistent and asymptotically efficient.
- An MLE sequence is often consistent, but this needs to be verified.