Lecture Notes on Statistical Inference

Tianyu Zhang[1]

Table of Contents:

[1] BIMSA, bidenbaka@gmail.com

1.1 Set Theory

One of the main objectives of statistics is to draw conclusions about a population of objects by conducting an experiment. The first step in this endeavor is to identify the possible outcomes or, in statistical terminology, the sample space.

Definition: Sample Space

The set, S, of all possible outcomes of a particular experiment is called the sample space for the experiment.

Once the sample space has been defined, we are in a position to consider collections of possible outcomes of an experiment.

Definition: Event

An event is any collection of possible outcomes of an experiment, i.e. any subset of S.

Let A be an event, i.e. a subset of S. Note that since S is a subset of itself, therefore it is possible for $A = S$. We say that the event A occurs if the outcome of the experiment is in the set A. When speaking of probabilities, we generally speak of the probability of an event, rather than a set. But we may use the terms interchangably.

We first need to define formally the following two relationships, which allow us to order and equate sets:

$$A \subseteq B \Leftrightarrow x \in A \Rightarrow x \in B \tag{1.1}$$

$$A = B \Leftrightarrow A \subseteq B \text{ and } B \subseteq A. \tag{1.2}$$

Given any two events (or sets) A and B, we have the following elementary set operations:

$$A \cup B := \{x \,|\, x \in A \text{ or } x \in B\}. \tag{1.3}$$

$$A \cap B := \{x \,|\, x \in A \text{ and } x \in B\}. \tag{1.4}$$

$$A^c := \{x \,|\, x \notin A, x \in S\}. \tag{1.5}$$

Theorem 1.1:

For any three events A, B, and C, defined on a sample space S, one has:

(i) $A \cup B = B \cup A, A \cap B = B \cap A.$ (Commutative)

(ii) $A \cup (B \cup C) = (A \cup B) \cup C, A \cap (B \cap C) = (A \cap B) \cap C.$ (Associative)

(iii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C), A \cup (B \cap C) = (A \cup B) \cap (A \cup C).$ (Distributive)

(iv) $(A \cup B)^c = A^c \cap B^c, (A \cap B)^c = A^c \cup B^c.$ (DeMorgan's Law)

The operations of union and intersection can be extended to infinite collections of sets as well. If $A_1, A_2, \cdots$ is a countable collection of sets, all defined on a sample space S, then one has

$$\bigcup_{i=1}^{\infty} A_i = \{x \in S \,|\, x \in A_i \text{ for some } i\}, \tag{1.6}$$

$$\bigcap_{i=1}^{\infty} A_i = \{x \in S \,|\, x \in A_i \, \forall i\}. \tag{1.7}$$

We can even generalize this into the union (resp. the intersection) of arbitrarily many sets, i.e., the index set may be uncountable.

$$\bigcup_{\alpha\in\Gamma} A_\alpha := \{x \in S \,|\, x \in A_\alpha \text{ for some } \alpha \in \Gamma\}. \tag{1.8}$$

$$\bigcap_{\alpha\in\Gamma} A_\alpha := \{x \in S \,|\, x \in A_\alpha \text{ for some } \alpha \in \Gamma\}. \tag{1.9}$$

Finally, we discuss the idea of a partition of the given sample space.

Definiton: Disjoint, Pairwise Disjoint

Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \varnothing$. The events $A_1, A_2, \cdots$ are said to be pairwise disjoint (mutually exclusive) if $A_i \cap A_j = \varnothing \ \forall i \neq j$.

Note that the disjoint sets are sets with no points in common. If one draws a Venn diagram for two disjoint sets, the sets do not overlap. The collection $A_i := [i, i+1)$ for $i = 0,1,2,\cdots$, consists of pairwise disjoint sets. Note further that $\cup_{i=0}^{\infty} A_i = [0,\infty)$.

Definition: Partition

If $A_1, A_2, \cdots$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, \cdots$ forms a partition of S.

1.2 Basics of Probability Theory

For each event A in the sample space S we want to associate with A a number between zero and one that will be called the probability of A, denoted by $\mathbb{P}(A)$. It would seem natural to define the domain of P as all subsets of S; that is, for each $A \subseteq S$ we define $\mathbb{P}(A)$ as the probability that A occurs. Note that in measure theory, there are two different approach to define a measure, whether on a σ-ring or on a σ-algebra, they concepts and the definitions on both are identically the same, the choice of using σ-ring or σ-algebra is based on the author's preferrence. We shall use the approach by the σ-algebra.

Definition: σ-algebra/Borel Field

A collection of subsets of S is called a σ-algebra (or Borel field), denoted by $\mathscr{B}$, if it satisfies the following three properties:

(i) $\varnothing \in \mathscr{B}$.

(ii) If $A \in \mathscr{B}$ then $A^c \in \mathscr{B}$.

(iii) If $A_1, A_2, \cdots \in \mathscr{B}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{B}$.

The empty set $\varnothing$ is a subset of any set. Thus, $\varnothing \subset S$. Property (i) states that this subset is always in a σ-algebra. Since $S = \varnothing^c$, properties (i) and (ii) imply that S is also in $\mathscr{B}$. Moreover, by DeMorgan's Laws it follows that $\mathscr{B}$ is closed under countable intersections. That is, if $A_1, A_2, \cdots \in \mathscr{B}$, then $A_1^c, A_2^c, \cdots \in \mathscr{B}$ by property (ii), and therefore $\bigcup_{i=1}^{\infty} A_i^c \in \mathscr{B}$. However, using DeMorgan's Law, we have

$$\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c = \bigcap_{i=1}^{\infty} A_i.$$

Thus, again by property (ii), $\bigcap_{i=1}^{\infty} A_i \in \mathscr{B}$.

Associated with sample space S we can have many different σ-algebras. For example, the collection of the two sets $\{\varnothing, S\}$ is a σ-algebra, usually called the trivial σ-algebra. The only σ-algebra we will be concerned with is the smallest σ-algebra that contains all of the open sets in a given sample space S.

Definition: Probability Function

Given a sample space S and an associated σ-algebra $\mathscr{B}$, a probability function is a function $\mathbb{P}$ with domain $\mathscr{B}$ that satisfies

(i) $\mathbb{P}(A) \geq 0 \forall A \in \mathscr{B}$. (Positive Semidefinite)

(ii) $\mathbb{P}(S) = 1$.

(iii) If $A_1, A_2, \cdots \in \mathscr{B}$ are pairwise disjoint, then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. (Countably Additive)

The three properties are usually referred to as the Axioms of Probability (or refer to Kolmogorov Axioms). Any function $\mathbb{P}$ that satisfies the Axioms of Probability is called a probability function. The axiomatic definition makes no attempt to tell what particular function $\mathbb{P}$ to choose; it merely requires $\mathbb{P}$ to satisfy the axioms. For any sample space many different probability functions can be defined. Which one (s) reflects what is likely to be observed in a particular experiment is still to be discussed.

We need general methods of defining probability functions that we know will always satisfy Kolmogorov's Axioms. We do not want to have to check the Axioms for each new probability function. The following gives a common method of defining a legitimate probability function.

Theorem 1.2:

Let $S = \{s_1, \cdots, s_n\}$ be a finite set. Let $\mathscr{B}$ be any σ-algebra of subsets of S. Let $p_1, \cdots, p_n$ be nonnegative number that sum to 1. For any $A \in \mathscr{B}$, define $\mathbb{P}(A)$ by $\mathbb{P}(A) = \sum_{\{i|s_i \in A\}} p_i$. Then $\mathbb{P}$ is a probability function on $\mathscr{B}$. This remains true if $S = \{s_1, s_2, \cdots\}$ is a countable set.

Proof:

We will give the proof for the finite case while the countable case holds by induction. For any $A \in \mathscr{B}$, $\mathbb{P}(A) = \sum_{\{i|s_i \in A\}} p_i \geq 0$, since every $p_i \geq 0$. Thus, positive homogeneity follows. Moreover, we have

$$\mathbb{P}(S) = \sum_{\{i|s_i \in S\}} p_i = \sum_{i=1}^{n} p_i = 1.$$

Thus the second axiom follows. Let now $A_1, \cdots, A_k$ denote pairwise disjoint events. Then

$$\mathbb{P}\Big(\bigcup_{i=1}^{k} A_i\Big) = \sum_{\{j|s_j \in \bigcup_{i=1}^{k} A_i\}} p_j = \sum_{i=1}^{k} \sum_{\{j|s_j \in A_i\}} p_j = \sum_{i=1}^{k} \mathbb{P}(A_i).$$

The above equality is valid: The first and the third equalities are true by the definition of $\mathbb{P}(A)$. The disjointedness of the A_i's ensures that the second equality fails to be false since the same p_j's appear exactly once on each side of the equality. Therefore the countable additivity follows.

□

1.3 The Calculus of Probabilities

From the Axioms of Probability we can build up many properties of the probability function, properties that are quite helpful in the calculation of more complicated probabilities. Some of these manipulations will be discussed in this subsection.

We start with some (fairly self-evident) properties of the probability function when applied to a single event.

Theorem 1.3:

If $\mathbb{P}$ is a probability function and A is any set in $\mathscr{B}$. Then

(i) $\mathbb{P}(\varnothing) = 0$.

(ii) $\mathbb{P}(A) \leq 1$.

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

Proof:

We shall prove (iii) first since it is the easiest part.

(iii):

The sets A and A^c form a partition of the sample space, i.e. $S = A \cup A^c$.

Therefore,

$$\mathbb{P}(A \cup A^c) = P(S) = 1 \tag{1.10}$$

by the second axiom. Also, A and A^c are disjoint, so by the third axxiom,

$$\mathbb{P}(A \cup A^c) = P(A) + P(A^c). \tag{1.11}$$

Combining (1.10) and (1.11) yields (iii).

(ii):

Since $\mathbb{P}(A^c) \geq 0$, (ii) follows directly from (iii).

(i):

To prove (i), we use a similar argument on $S = S \cup \varnothing$. Since S and $\varnothing$ are disjoint, one has

$$1 = \mathbb{P}(S) = \mathbb{P}(S \cup \varnothing) = \mathbb{P}(S) + \mathbb{P}(\varnothing),$$

therefore $\mathbb{P}(\varnothing) = 0$.

□

Theorem 1.3 contains properties that are so basic that they also have the flavor of axioms, although we have formally proved them using only the original three Kolmogorov Axioms. The next theorem, which is similar in spirit of **Theorem 1.3**, contains statements that are not so self-evident.

Theorem 1.4:

If $\mathbb{P}$ is a probability function and A and B are any sets of $\mathscr{B}$. Then

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

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

(iii) If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Theorem 1.5:

If $\mathbb{P}$ is a probability function, then

(i) $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap C_i)$ for any partition $C_1, C_2, \cdots$.

(ii) $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for any sets $A_1, A_2, \cdots$. (Boole's Inequality)

1.4 Conditionaly Probability

Definition: Conditionaly Probability

If A and B are two events in S and $\mathbb{P}(B) > 0$, then the conditional probability of A given B, written as $\mathbb{P}(A \mid B)$, is $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

Note that what happens in the conditional probability calculation is that B becomes the sample space then $\mathbb{P}(B \mid B) = 1$. The intuition is that our original sample space S has been updated to B. All further occurrences are then calibrated with respect to their relation to B. In particular, note that what happens to conditional probabilities of disjoint sets. Suppose that A and B are disjoint, then $\mathbb{P}(A \cap B) = 0$. It then follows that $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A) = 0$.

Rewriting the formula of conditional probability yields the form

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B). \tag{1.12}$$

Since $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$ by the symmetry of the operation "∩", it follows that we can further express (1.12) into the form

$$\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\mathbb{P}(A). \tag{1.13}$$

Therefore we have a useful formula

$$\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)}, \tag{1.14}$$

which gives a formulfor turning around the conditional probabilities. (1.14) is often called Bayes' Rule. We now introduce a more general form.

Theorem 1.6: Bayes' Rule

Let $A_1, A_2, \cdots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1,2,\cdots$, one has that

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}.$$

Therefore, a simple calculation yields a transformation between conditioned sets, i.e. one has $\mathbb{P}(B \mid A_i) = \dfrac{\sum_{j=1}^{\infty} \mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}{\mathbb{P}(A_i \mid B)\mathbb{P}(A_i)}$.

In some cases it may happen that the occurrence of a particular event, B, has no effect on the probability of another event. That is, $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. If this is the case, then by Bayes' Rule, one has that

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

It follows that, since $\mathbb{P}(B\,|\,A)\mathbb{P}(A) = \mathbb{P}(A \cap B)$, that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. This is precisely when B has no effect on A, and we call this the statistically independent.

Definition: Statistically Independent

Two events A and B are said to be statistically independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Note that the independence could have been equivalently defined by either the form $\mathbb{P}(A\,|\,B) = \mathbb{P}(A)$ or the form $\mathbb{P}(B\,|\,A) = \mathbb{P}(B)$, with further stating that $\mathbb{P}(B) > 0$ or $\mathbb{P}(A) > 0$, respectively. The advantage of the above definition is that it treats the events symmetrically and will be easier to generalize to more than two events.

Theorem 1.7:

If A and B are independent events, then the following pairs are also independent:

(i) A and B^c.

(ii) A^c and B.

(iii) A^c and B^c.

The proof of this theorem is left as an exercise since it is only an elementary application of the above results. Now let us generalize the statistically independent to the case involving more than two events.

Definition: Mutually Independent

A collection of events $A_1, \cdots, A_n$ are mutually independent if for any subcollection $A_{i_1}, \cdots, A_{i_k}$, one has $\mathbb{P}\big(\bigcap_{j=1}^{k} A_{i_j}\big) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j})$.

1.5 Random Variables

Recall that random variables are measurable functions from the sample space to real numbers. There are two most important quantities associated with a random variable X, namely the expectation (also called the mean), and the variance. We shall denote them throughout this note by $\mathbb{E}X$ and $\mathrm{Var}(X) := \mathbb{E}(X - \mathbb{E}X)^2$.

In fact, the expectation $\mathbb{E}X$ of a random variable X on a probability space $(\Omega, \Sigma, \mathbb{P})$ is the Lebesgue integral of the function $X : \Omega \to \mathbb{R}$. This makes all theorems on Lebesgue integration applicable in probability theory, for expectations of random variables.

One more thing to note is that the change of sample space from S to Ω is for a reason. In defining a random variable, we have also defined a new sample space. Suppose we have a sample space $S = \{s_1, \cdots, s_n\}$ with a probability function $\mathbb{P}$ and we define a random variable X with range $\Omega = \{x_1, \cdots, x_m\}$. We can define a probability function $\mathbb{P}_X$ on Ω by

$$\mathbb{P}_X(X = x_i) = \mathbb{P}(\{s_j \in S\,|\,X(s_j) = x_i\}). \tag{1.15}$$

The same thing happens when Ω is countable. When Ω is not countable, $\mathbb{P}_X$ is defined by, for any set $A \subseteq \Omega$,

$$\mathbb{P}_X(X \in A) = \mathbb{P}(\{s \in S \,|\, X(s) \in A\}). \tag{1.16}$$

With every random variable X, we associate a function called the cumulative distribution of X.

Definition: Cumulative Distribution Function (CDF)

The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by $F_X(x) = \mathbb{P}_X(X \leq x)\ \forall x \in S$.

Theorem 1.8:

The function $F(x)$ is a cdf if and only if the followings hold:

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

(ii) $F(x)$ is a nondecreasing function of x.

(iii) $F(x)$ is right-continuous, i.e. $\forall x_0 \in \Omega$, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

The right continuity is a direct result from its definition $F_X(x) = \mathbb{P}_X(X \leq x)$. When one has $F_X(x) = \mathbb{P}_X(X < x)$ in contrast, F_X is left-continuous. One can turn **Theorem 1.8** to an alternative definition for cdf.

Whether a cdf is continuous or has jumps corresponds to the associated random variable being continuous or not. In fact, the association is such that it is convenient to define continuous random variables in this way.

Definition: Continuous, Discrete

A random variable X is continuous if $F_X(x)$ is a continuous function of x.

A random variable X is discrete if $F_X(x)$ is a step function of x.

We close this subsection with a theorem formally stating that F_X completely determines the probability distribution of a random variable X. This is true if $\mathbb{P}(X \in A)$ is defined only for events A in $\mathscr{B}$ the smallest sigma algebra containing all the intervals of reals in all forms $((a,b),(a,b],[a,b),[a,b],a,b \in \mathbb{R})$. If probabilities are defined for a larger class of events, it is possible for two random variables to have the same distribution function but not the same probability for every event. We shall not deal with this problem in this note, hence we need to restrict ourselves into good settings.

Definition: Identically Distributed

The random variable X and Y are identically distributed if for every set A in $\mathscr{B}$ one has $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$.

Note that two random variables that are identically distributed are not necessarily equal.

Theorem 1.9:

The random variables X and Y are identically distributed $\Leftrightarrow F_X(x) = F_Y(x) \forall x$.

Proof:

"$\Rightarrow$":

Since X and Y are identically distributed, for any set $A \in \mathscr{B}$, by definition, $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$. In particular, $\forall x$, the set $(-\infty, x]$ is in $\mathscr{B}$, hence

$$F_X(x) = \mathbb{P}(X \in (-\infty, x]) = \mathbb{P}(Y \in (-\infty, x]) = F_Y(x).$$

"$\Leftarrow$":

Showing this direction requires heavy use of sigma algebras; we will not go into these details. It suffices to say that it is necessary to prove only that the

two probability functions agree on all intervals (Chung 1974, Section 2.2).

□

Associated with a random variable X and its cdf F_X is another function, called either the probability density function (pdf) or probability mass function (pmf). The terms pdf and pmf refer, respectively, to the continuous and discrete cases. Both pdfs and pmfs are concerned with "point probabilities" of random variables.

Definition: Probability Mass Function (PMF)

The probability mass function (pmf) of a discrete random variable X is given by $f_X(x) = \mathbb{P}(X = x)\ \forall x$.

A widely accepted convention, which we shall adopt, is to use the uppercase letter for the cdf and the corresponding lowercase letter for the pmf or pdf.

We must be a little more careful in our definition of a pdf in the continuous case. If we naively try to calculate $\mathbb{P}(X = x)$ for a continuous random variable, we get the following: Since $\{X = x\} \subseteq \{x - \varepsilon < X \leq x\}\ \forall \varepsilon > 0$, one has

$$\mathbb{P}(X = x) \leq \mathbb{P}(x - \varepsilon < X \leq x) = F_X(x) - F_X(x - \varepsilon). \tag{1.17}$$

Therefore

$$0 \leq \mathbb{P}(X = x) \leq \lim_{\varepsilon \downarrow 0}(F_X(x) - F_X(x - \varepsilon)) = 0$$

by the continuity of F_X. However, if we understand the purpose of the pdf, its definition shall not be ambiguous.

In the discrete case, we can sum over values of the pmf to get the cdf. The analogous procedure in the continuous case is to substitute integrals for sums, and we get

$$\mathbb{P}(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{1.18}$$

Using the **Fundamental Theorem of Calculus**, if $f_X(x)$ is continuous, we have the further relationship

$$\frac{d}{dx}F_X(x) = f_X(x). \tag{1.19}$$

Note that the analogy with the discrete case is almost exact. We "add up" the "point probabilities" $f_X(x)$ to obtain interval probabilities. Let us summarize this into the formal definition.

Definition: Probability Density Function (PDF)

The probability density function (pdf), $f_X(x)$, of a continuous random variable X is the function that satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)dt\ \forall x$.

Remark:

The expression "X has a distribution given by $F_X(x)$" is abbreviated symbolically by "$X \sim F_X(x)$," where we read the symbol "$\sim$" as "is distributed as." We can similarly write $X \sim f_X(x)$ or, if X and Y have the same distribution, $X \sim Y$.

In the continuous case we can somewhat cavalier about the specification of interval probabilities. Since $\mathbb{P}(X = x) = 0$ if X is a continuous random variable, it follows

$$\mathbb{P}(a < x < b) = \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X < b) = \mathbb{P}(a \leq X \leq b).$$

It should be clear that the pdf (or pmf) contains the same information as the cdf. This being the case, we can use either one to solve problems and should try to choose the simpler one.

Theorem 1.10:

A function $f_X(x)$ is a pdf (or pmf) of a random variable X if and only if

(i) $f_X(x) \geq 0 \forall x$.

(ii) $\sum_x f_X(x) = 1$ (pmf), or, $\int_{-\infty}^{+\infty} f_X(x)dx = 1$ (pdf).

Proof:

If $f_X(x)$ is a pdf (or pmf), then the two properties are immediate from the definitions. In particular, for a pdf, one has $1 = \lim_{x\to\infty} F_X(x) = \int_{-\infty}^{+\infty} f_X(t)dt$.

The converse implication is equally easy to prove. Once one has $f_X(x)$, one can define $F_X(x)$ and result follows from **Theorem 1.8**.

□

From a purely mathematical viewpoint, any nonnegative function with a finite positive integral (or sum) can be turned into a pdf or pmf. For example, if $h(x)$ is any nonnegative function that is positive on a set A, 0 otherwise, and

$$\int_{\{x\in A\}} h(x)dx = K < \infty$$

For some $K > 0$, then the function $f_X(x) = h(x)/K$ is a pdf of a random variable X taking values in A.

Actually, $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ does not always hold since $F_X(x)$ may be continuous but it may not be differentiable. In fact, there exist continuous random variables for which the integral relationship does not exist for any $f_X(x)$. These cases are rather pathological and we shall not discuss them in this note. Thus, in this text, we shall always assume that $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ holds for any continuous random variable. In more advanced literature a random variable is called absolutely continuous if the integral relationship holds.

2.1 Distributions of Functions of a Random Variable

If X is a random variable with cdf $F_X(x)$, then any function of X, namely $g(X)$, is also a random variable. Often $g(X)$ is of interest itself and we write $Y = g(X)$ to denote the new random variable $g(X)$. Since Y is a function of X, we can describe the probabilistic behavior of Y in terms of X. That is, for any set A,

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A).$$

hence the distribution of Y depends on the functions F_X and g. Depending on the choice of g, it is sometimes possible to obtain a tractable expression for this probability.

Formally, if we write $y = g(x)$, the function $g(x)$ defines a mapping from the original sample space of X to a new sample space, if we denote them by Ω_X and Ω_Y, respectively. Then

$$g : \Omega_X \to \Omega_Y.$$

We associate with g an inverse mapping, denoted by g^{-1}, which is a mapping from subsets of Ω_Y to subsets of Ω_X, and is defined by

$$g^{-1}(A) = \{x \in \Omega_X \,|\, g(x) \in A\}.$$

Hence for any set $A \subseteq \Omega_Y$, one has

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) = \mathbb{P}(\{x \in \Omega_X \,|\, g(x) \in A\}) = \mathbb{P}(X \in g^{-1}(A)).$$

This defines the probability distribution of Y. It is straightforward to show that this probability distribution satisfies the Kolmogorov Axioms.

If we assume X is a discrete random variable, then Ω_X is countable. The sample space for $Y = g(X)$ is $\Omega_Y := \{y \,|\, y = g(x), x \in \Omega_X\}$, which is also a countable set. Thus, Y is also a discrete random variable. The pmf of Y is

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in g^{-1}(y)} \mathbb{P}(X = x) = \sum_{x \in g^{-1}(y)} f_X(x) \text{ for } y \in \Omega_Y.$$

and $f_Y(y) = 0$ if $y \notin \Omega_Y$. In this case, finding the pmf of Y is identifying $g^{-1}(y)$, for each $y \in \Omega_Y$, and summing the appropriate probabilities.

Example 2.1: Binomial Transformation

Recall that for nonnegative integers n and r such that $n \geq r$, one has the formula of n choose r being $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.

A discrete random variable X has a binomial distribution if its pmf is of the form

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0,1,\cdots,n,$$

where n is a positive integer and $0 \leq p \leq 1$. Consider the random variable $Y = g(X)$, where $g(x) = n - x$. Now we have $\Omega_X = \{0,1,\cdots,n\}$ and it follows that $\Omega_Y = \{y \,|\, y = g(x), x \in \Omega_X\}$. Since $y = n - x$, we have $x = n - y$, thus

$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x) = f_X(n-y) = \binom{n}{n-y} p^{n-y}(1-p)^{n-(n-y)}$$
$$= \binom{n}{y}(1-p)^y p^{n-y}. \text{ (Since } \binom{n}{y} = \binom{n}{n-y}\text{).}$$

From $X \sim \text{Binomial}(n, p)$ we arrive at $Y \sim \text{Binomial}(n, 1-p)$. ||

If X and Y are now continuous random variables, then in some cases it is possible to find simple formulas for the cdf and pdf of Y in terms of the cdf and pdf of X and the corresponding function g. The cdf of $Y = g(X)$ is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y))$$
$$= \mathbb{P}(\{x \in \Omega_X \,|\, g(x) \leq y\}) = \int_{\{x \in \Omega_X | g(x) \leq y\}} f_X(x) dx.$$

When the transformation are made, it is important to keep track of the sample spaces of the random variables; otherwise, much confusion can arise. When the transformation is from X to Y by $g(X)$, it is most convenient to use

$$\Omega_X = \{x \,|\, f_X(x) > 0\} \text{ and } \Omega_Y = \{y \,|\, y = g(x) \text{ for some } x \in \Omega_X\}.$$

The pdf of the random variable X is positive only on the set Ω_X and 0 elsewhere. Such a set is called the **support set** of a distribution, or, more informally, the support of a distribution. This terminology can also apply to a pmf, or, in general, to any nonnegative function.

Theorem 2.1:

If X and Y are random variables such that X has cdf $F_X(x)$ and $Y = g(X)$ with $\Omega_X = \{x \,|\, f_X(x) > 0\}$ and $\Omega_Y = \{y \,|\, y = g(x) \text{ for some } x \in \Omega_X\}$.

(i) If g increasing on Ω_X, then $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \Omega_Y$.

(ii) If g is decreasing on Ω_X and X is a continuous random variable. Then $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \Omega_Y$.

If the pdf of Y is continuous, it can be obtained by differentiating the cdf. That is

Theorem 2.2:

Let X have pdf $f_X(x)$ and $Y = g(X)$, where g is monotone. Assume that $\Omega_X = \{x \,|\, f_X(x) > 0\}$ and $\Omega_Y = \{y \,|\, y = g(x) \text{ for some } x \in \Omega_X\}$. Suppose that $f_X(x)$ is continuous on Ω_X and $g^{-1}(y)$ has a continuous derivative on Ω_Y. Then the pdf of Y is $f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, y \in \Omega_Y \\ 0, \text{ otherwise} \end{cases}$.

Proof:

According to **Theorem 2.1** and chain rule, one has

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \begin{cases} f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y), g \text{ is increasing} \\ -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y), g \text{ is decreasing} \end{cases}.$$

Result follows.

□

In many applications, the function g may be neither increasing nor decreasing; hence the above results will not apply in general. However, it is often the case that g will be monotone over certain invervals and that allows us to get an expression for $Y = g(X)$.

Theorem 2.3:

Let X have pdf $f_X(x)$, let $Y = g(X)$, and define the sample space $\Omega_X = \{x \,|\, f_X(x) > 0\}$ and $\Omega_Y = \{y \,|\, y = g(x) \text{ for some } x \in \Omega_X\}$. Suppose there exists a partition $A_0, A_1, \cdots, A_k$ of Ω_X such that $\mathbb{P}(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i. In addition, suppose there exist functions $g_1(x), \cdots, g_k(x)$ defined on $A_1, \cdots, A_k$, respectively, such that

(i) $g(x) = g_i(x)$ for $x \in A_i$.

(ii) $g_i(x)$ is monotone on A_i.

(iii) The set $\Omega_Y = \{y \,|\, y = g_i(x) \text{ for some } x \in \Omega_X\}$ is the same for all i.

(iv) g_i^{-1} has a continuous derivative on Ω_Y for all i.

Then one has

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right|, y \in \Omega_Y \\ 0, \text{ otherwise} \end{cases}.$$

The important part is that Ω_X can be divided into sets $A_1, \cdots, A_k$, such that $g(x)$ is monotone on each A_i. We can ignore the "excepional set" A_0 since $\mathbb{P}(X \in A_0) = 0$.

Theorem 2.4: Probability Integral Transformation

Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(x)$. Then Y is uniformly distributed on (0,1), i.e. $\mathbb{P}(Y \leq y) = y$, for $0 < y < 1$.

2.2 Expected Values

The expected value, or expectation, of a random variable is merely its average value, where we speak of "average" values as one that is weighted according to the probability distribution. The expected value of a distribution can be thought of as a measure of center, as we think of averages as being middle values. By weighting the values of the random variable according to the probability distribution, we hope to obtain a number that summarizes a typical or expected value of an observation of the random variable.

Definition: Expected Value (mean)

The expected value or mean of a random variable $g(X)$, denoted by $\mathbb{E}g(X)$, is

$$\mathbb{E}g(X) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) dx, \text{, if } X \text{ is continuous} \\ \sum_{x \in \Omega_X} g(x) f_X(x) = \sum_{x \in \Omega_X} g(x) \mathbb{P}(X = x), \text{ if } X \text{ is discrete} \end{cases}$$

provided that the integral or sum exists. If $\mathbb{E} \mid g(X) \mid = \infty$, we say that $\mathbb{E}g(X)$ does not exist.

The process of taking expectations is a linear operation, which means that the expectation of a linear function of X can be easily evaluated by nothing that for any constant a and b, that

$$\mathbb{E}(aX + b) = a\mathbb{E}X + b. \tag{2.1}$$

For example, if $X \sim \text{Binomial}(n, p)$, so that $\mathbb{E}X = np$, then

$$\mathbb{E}(X - np) = \mathbb{E}X - np = np - np = 0.$$

The expectation operator, in fact, has many properties that can help relax calculational effort. Most of these properties follow from the properties of the integral or sum, and are summarized in the following theorem:

Theorem 2.5:

Let X be a random variable and let $a, b,$ and c be constants. Then for any function $g_1(x)$ and $g_2(x)$ whose expectations exist, one has

(i) $\mathbb{E}(ag_1(X) + bg_2(X) + c) = a\mathbb{E}g_1(X) + b\mathbb{E}g_2(X) + c$.

(ii) If $g_1(x) \geq 0 \forall x$ then $\mathbb{E}g_1(X) \geq 0$.

(iii) If $g_1(x) \geq g_2(x) \forall x$ then $\mathbb{E}g_1(X) \geq \mathbb{E}g_2(X)$.

(iv) If $a \leq g_1(x) \leq b \forall x$ then $a \leq \mathbb{E}g(X) \leq b$.

When evaluating expectations of nonlinear function of X, we can proceed in one or two ways. From the definition of $\mathbb{E}g(X)$, we could directly calculate

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

But we could also find the pdf $f_Y(y)$ of $Y = g(X)$ and we would have

$$\mathbb{E}g(X) = \mathbb{E}Y = \int_{-\infty}^{\infty} f_Y(y) dy.$$

2.3 Moments and Moment Generating Functions

The various moments of a distribution are an important class of expectation:

Definition: Moment

For each integer n, the n-th moment of a random variable X (or $F_X(x)$), denoted as μ_n', is defined by $\mu_n' = \mathbb{E}X^n$.

Definition: Central Moment

The n-th central moment of X, denoted as μ_n, is defined by $\mu_n = \mathbb{E}(X - \mu)^n$, where $\mu = \mu_1' = \mathbb{E}X$, the expected value of X.

Aside from the mean $\mathbb{E}X$, perhaps the most important moment is the second central moment, more commonly known as the variance.

Definition: Variance

The variance of a random variable X is its second central moment, denoted as $\mathrm{Var}X$, is defined by $\mathrm{Var}X = \mathbb{E}(X - \mathbb{E}X)^2$.

Definition: Standard Deviation

The standard deviation of X is the positive square root of $\mathrm{Var}X$, i.e. it is defined by $\sqrt{\mathrm{Var}X}$.

The variance gives a measure of the degree of spread of a distribution around its mean. For example, the quantity $\mathbb{E}(X - b)^2$ is minimized when $b = \mathbb{E}X$. Now we consider the absolute size of this minimum. The interpretation atttached to the variance is that larger values mean X is more variable. At the extreme, if

$$\mathrm{Var}X = \mathbb{E}(X - \mathbb{E}X)^2 = 0,$$

then X is equal to $\mathbb{E}X$ with probability 1, and there is no variation in X. The standard deviation has the same qualitative interpretation: Small value means that X is very likely to be close to $\mathbb{E}X$, and large values mean X is very variable. The standard deviation is easier to interpret in that measurement unit on the standard deviation is the same as that for the original variable X. The measurement unit on the variance is the square of the original unit.

Theorem 2.6:

If X is a random variable with finite variance, then for any constants a and b, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}X$.

Proof:

According to the definition, one has

$$\begin{aligned}\mathrm{Var}(aX + b) &= \mathbb{E}\big((aX + b) - \mathbb{E}(aX + b)\big)^2 \\ &= \mathbb{E}(aX - a\mathbb{E}X)^2 \qquad (\mathbb{E}(aX + b) = a\mathbb{E}X + b).\end{aligned}$$

$$= a^2\mathbb{E}(X - \mathbb{E}X)^2 = a^2\text{Var}X.$$

□

It is sometimes easier to use an alternative formula for the variance, given by

$$\text{Var}X = \mathbb{E}X^2 - (\mathbb{E}X)^2,$$

which is easily established by noting that

$$\begin{aligned}\text{Var}X &= \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}\left(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2\right)\\ &= \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2\\ &(\mathbb{E}(2X\mathbb{E}X) = 2\mathbb{E}(X\mathbb{E}X) = 2\mathbb{E}(X)\mathbb{E}(X) = 2(\mathbb{E}X)^2 \text{ since } \mathbb{E}X \text{ is a constant})\\ &= \mathbb{E}X^2 - (\mathbb{E}X)^2.\end{aligned}$$

We now introduce a new function that is associated with a probability distribution, the moment generating function (mgf). As its name suggests, the mgf can be used to generate moments. In practice, it is easier in many cases to calculate moments directly than to use the mgf. However, the main use of the mgf is not to generate moments, but to help in characterizing a distribution. This property can lead to some extremely powerful results when used properly.

Definition: Moment Generating Function (mgf)

Let X be a random variable with cdf F_X. The moment generating function of X (of F_X), denoted by $M_X(t)$, is defined to be $M_X(t) = \mathbb{E}e^{tX}$, provided that the expectation exists for t in some neighbourhod of 0. That is, $\exists\delta > 0$ such that $\forall t \in (-\delta, \delta)$, $\mathbb{E}e^{tX}$ exists. If the expectation does not exist in a neighborhood of 0, we say that the mgf does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx, \; X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx}\mathbb{P}(X = x), \; X \text{ is discrete.}$$

It is very easy to see how the mgf generates moments. We summarize the result in the following result:

Theorem 2.7:

If X has mgf $M_X(t)$, then $\mathbb{E}X^n = M_X^{(n)}(0)$, where $M_X^{(n)}(0) := \dfrac{d^n}{dt^n}M_X(t)\Big|_{t=0}$.

That is to say, the n-th moment is equal to the n-th derivative of $M_X(t)$ evaluated at $t = 0$.

Proof:

Assume that we can differentiate under the integral sign, we have

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \frac{d}{dt}\int_{-\infty}^{\infty} e^{tx}f_X(x)dx = \int_{-\infty}^{\infty}\left(\frac{d}{dt}e^{tx}\right)f_X(x)dx\\ &= \int_{-\infty}^{\infty}(xe^{tx})f_X(x)dx =: \mathbb{E}Xe^{tX}.\end{aligned}$$

Thus, one has

$$\frac{d}{dt}M_X(t)\Big|_{t=0} = \mathbb{E}Xe^{tX}\Big|_{t=0} = \mathbb{E}X.$$

Proceeding in an analogous manner, the result follows.

□

As previously mentioned, the major usefulness of the mgf is not in its ability to generate moments, rather, its usefulness stems from the fact that, in many cases, the mgf can characterize a distribution. There are, however, some technical difficulties associated with using moments to characterize a distribution, which we will now investigate.

If the mgf exists, it characterizes an infinite set of moments. The natural question is whether characterizing the infinite set of moments uniquely determines a distribution function. The answer to this question, unfortunately, is no. Characterizing the set of moments is not enough to determine a distribution uniquely because there may be two distinct random variables having the same moments.

The problem of uniqueness of moments does not occur if the random variables have bounded support. If that is the case, then the infinite sequence of moments does uniquely determine the distribution. Furthermore, if the mgf exists in a neighborhood of 0, then the distribution is uniquely determined, no matter what its support. Thus, existence of all moments is not equivalent to existence of the mgf. The following theorem shows how a distribution is characterized.

Theorem 2.8:

Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist. Then

(i) If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all $u \Leftrightarrow \mathbb{E}X^r = \mathbb{E}Y^r$ for all intergers $r = 0,1,2,\cdots$.

(ii) If the mgfs exist and $M_X(t) = M_Y(t)\ \forall t \in (-\delta, \delta)$, where $\delta > 0$, then $F_X(u) = F_Y(u)$ for all u.

In the next theorem, which deals with a sequence of mgfs that converges, we do not treat the bounded support case separately. Note that the uniqueness assumption is automatically satisfied since the limiting mgf exists in a neighborhood of 0.

Theorem 2.9: Convergence of MGFs

Suppose that $\{X_i\}$ is a countable sequence of random variables each with mgf $M_{X_i}(t)$. Assume that $\lim_{i\to\infty} M_{X_i}(t) = M_X(t) \forall t \in (-\delta, \delta)$ and $M_X(t)$ is an mgf. Then there exists a unique cdf F_X whose moments are determined by $M_X(t)$ and, $\forall x$ where $F_X(x)$ is continuous, one has $\lim_{i\to\infty} F_{X_i}(x) = F_X(x)$. That is, convergence, for $t \in (-\delta, \delta)$, of mgfs to an mgf implies convergence of cdfs.

The proofs of **Theorem 2.8** and **Theorem 2.9** rely on the theory of Laplace transforms. The defining equation for $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$ defines a Laplace transform, i.e. $M_X(t)$ is the Laplace transform of $f_X(x)$. A key fact about Laplace transforms is their uniqueness. If $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$ is valid $\forall t \in (-\delta, \delta)$, then given $M_X(t)$ there is only one function $f_X(x)$ satisfies $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$. Given this fact, the two previous theorems are quite reasonable.

The possible nonuniqueness of the moment sequence is an annoyance. If we show that a sequence of moments converges, we will not be able to conclude formally that the random variables converge. To do so, we would have to verify the uniqueness of the moment sequence, a generaly horrible job. However, if the sequence of mgfs converges in a neighborhood of 0, then the random variables converges. Thus, we can consider the convergence of mgfs as a sufficient but not necessary condition for the sequence of random variables to converge.

Theorem 2.10:

For any constants a and b, the mgf of the random variable $aX + b$ is given by $M_{aX+b}(t) = e^{bt}M_X(at)$.

Proof:

By definition,

$$M_{aX+b}(t) = \mathbb{E}(e^{(aX+b)t}) = \mathbb{E}(e^{(aX)t}e^{bt}) = e^{bt}\mathbb{E}(e^{(at)X}) = e^{bt}M_X(at).$$

□

2.4 Differentiabting under an Integral Sign

In the previous subsection we encountered an instance in which we desired to interchange the order of integration and differentiation. This situation is encountered frequently in theoretical statistics. The purpose of this subsection is to characterize conditions under which this operation is legitimate. We will also discuss interchanging the order of differentiation and summation.

Many of these conditions can be established using standard theorems for calculus and detailed proofs can be found in most calculus books. Thus, detailed proofs will not be presented here.

We first want to establish the method of calculating

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx, \tag{2.2}$$

where $-\infty < a(\theta), b(\theta) < \infty \quad \forall\theta$. The rule for differentiating (2.2) is called Leibnitz's Ruls and is an application of the Fundamental Theorem of Calculus and the chain rule.

Theorem 2.11: Leibnitz's Rule

If $f(x,\theta)$, $a(\theta)$, and $b(\theta)$ are differentiable with respect to θ, then we have

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx$$

$$= f\big(b(\theta),\theta\big)\frac{d}{d\theta}b(\theta) - f\big(a(\theta),\theta\big)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}f(x,\theta)dx.$$

Remark:

Note that if $a(\theta)$ and $b(\theta)$ are constant, we have a special case of Leibnitz's Rule: $\dfrac{d}{d\theta}\displaystyle\int_a^b f(x,\theta)dx = \int_a^b \frac{\partial}{\partial\theta}f(x,\theta)dx.$ ||

Thus, in general, if we have the integral of a differentiable function over a finite range, differentiation of the integral poses no problem. If the range of integration is infinite, however, problems can arise.

Note that the interchange of derivative and integral in the above equation equates a partial derivative with an ordinary derivative. Formally, this must be the case since the LHS is a function of only θ while the integrand on the RHS is a function of both θ and x.

The question of whether interchanging the order of differentiation and integration is justified is really a question of whether the limits and integration can be interchanged, since a derivative is a special kind of limit. Recall that if $f(x,\theta)$ is differentiable, then

$$\frac{\partial}{\partial\theta}f(x,\theta)=\lim_{\delta\to 0}\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta},$$

so we have

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)dx=\int_{-\infty}^{\infty}\lim_{\delta\to 0}\Big(\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}\Big)dx,$$

while we have

$$\frac{d}{d\theta}\int_{-\infty}^{\infty}f(x,\theta)dx=\lim_{\delta\to 0}\int_{-\infty}^{\infty}\Big(\frac{f(x,\theta+\delta)-f(x,\theta)}{\delta}\Big)dx.$$

Therefore, if we can justify the interchanging of the order of limits and integration, differentiation under the integral sign will be justified. Treatment of this problem in full generality will, unfortunately, necessiate the use of measure theory. However, the statements and conclusions of some important results can be given. The following theorems are all corollaries of **Lebesgue's Dominated Convergence Theorem**:

Theorem 2.12:

Suppose the function $h(x,y)$ is continuous at y_0 for each x and there exists a function $g(x)$ such that

(i) $|h(x,y)|\le g(x)\,\forall x,y,$

(ii) $\int_{-\infty}^{\infty}g(x)dx<\infty.$

Then $\lim_{y\to y_0}\int_{-\infty}^{\infty}h(x,y)dx=\int_{-\infty}^{\infty}\lim_{y\to y_0}h(x,y)dx.$

The key condition in this theorem is the existence of a dominating function $g(x)$, with a finite integral, which ensures that the integrals cannot be too badly behaved. We can now apply this theorem to the case we are considering by identifying $h(x,y)$ with the difference $\dfrac{f(x,\theta+\delta)-f(x,\theta)}{\theta}$.

Corollary 2.12.1:

Suppose that $f(x,\theta)$ is differentiable at $\theta=\theta_0$, i.e.

$$\lim_{\delta\to 0}\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}=\frac{\partial}{\partial\theta}f(x,\theta)\Big|_{\theta=\theta_0}$$

exists $\forall x$ and there exists a function $g(x,\theta_0)$ and a constant $\delta_0>0$ such that

(i) $\Big|\dfrac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}\Big|\le g(x,\theta_0)\,\forall x$ and $|\delta|<\delta_0.$

(ii) $\int_{-\infty}^{\infty} g(x,\theta_0)dx < \infty.$

Then

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx\Big|_{\theta=\theta_0} = \int_{-\infty}^{\infty} \Big(\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta_0}\Big)dx. \qquad (2.3)$$

The conclusion of **Corollary 2.12.1** is cumbersome, but it is important to realize that although we seem to be treating θ as a variable, the statement of the theorem is for one value of θ. That is, for each value θ_0 for $f(x,\theta)$ is differentiable at θ_0 and satisfies (i) and (ii), the order of integration and differentiation can be interchanged. Often the distinction between θ and θ_0 is not stressed since (2.3) is written

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx. \qquad (2.4)$$

Typically, $f(x,\theta)$ is differentiable at all θ, not at just one value θ_0. In this case, condition (i) of **Corollary 2.12.1** can be replaced by another condition that often proves easier to verify. By an application of the **Mean Value Theorem**, it follows that, for fixed x and θ_0, and $|\delta| \le \delta_0$,

$$\frac{f(x,\theta_0+\delta) - f(x,\theta_0)}{\delta} = \frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta_0+\delta^*(x)}$$

for some number $\delta^*(x)$, where $|\delta^*(x)| \le \delta_0$. Therefore, (i) will be satisfied if we find a $g(x,\theta)$ that satisfies (ii) and

$$\left|\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta'}\right| \le g(x,\theta) \ \forall\theta' \text{ such that } |\theta'-\theta| \le \delta_0. \qquad (2.5)$$

Note that in (2.5) δ_0 is implicitly a function of θ. This is permitted since the theorem is applied to each value of θ individually. From (2.5) we have the corollary:

Corollary 2.12.2:

Suppose that $f(x,\theta)$ is differentiable in θ and there exists a function $g(x,\theta)$ such that

(i) $\left|\frac{\partial}{\partial\theta} f(x,\theta)\Big|_{\theta=\theta'}\right| \le g(x,\theta) \ \forall\theta'$ such that $|\theta'-\theta| \le \delta_0$

(ii) $\int_{-\infty}^{\infty} g(x,\theta)dx < \infty.$

Then $\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} f(x,\theta)dx.$

Justification for taking the derivative inside the summation is more straightforward than the integration case. The following theorem provides the details.

Theorem 2.13:

Suppose that the series $\sum_{x=0}^{\infty} h(\theta,x)$ converges $\forall\theta$ in an interval (a,b) of real numbers and

(i) $\frac{\partial}{\partial\theta}h(\theta,x)$ is continuous in θ for all x.

(ii) $\sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta,x)$ converges uniformly on every closed bounded subinterval of (a,b).

Then $\frac{d}{d\theta}\sum_{x=0}^{\infty}h(\theta,x)=\sum_{x=0}^{\infty}\frac{\partial}{\partial\theta}h(\theta,x)$.

The condition of uniform convergence is the key one to verify in order to establish that the differentiation can be taken inside the summation. We close this subsection with a theorem similar to **Theorem 2.13**, but treats the case of interchanging the order of summation and integration.

Theorem 2.14:

Suppose that the series $\sum_{x=0}^{\infty}h(\theta,x)$ converges uniformly on $[a,b]$ and for each x, $h(\theta,x)$ is a continuous function of θ. Then

$$\int_a^b\sum_{x=0}^{\infty}h(\theta,x)d\theta=\sum_{x=0}^{\infty}\int_a^b h(\theta,x)d\theta.$$

3.1 Discrete Distributions

Recall that a random variable X is said to have a discrete distribution if the range of X, the sample space, is countable. In most situations, the random variable has integer-valued outcomes.

Definition: Discrete Uniform Distribution

A random variable X has a discrete **Uniform**$(1,N)$ distribution if $\mathbb{P}(X=x\,|\,N)=\frac{1}{N}$ for $x=1,2,\cdots,N$, where N is a specified integer. This distribution puts equal mass on each of the outcomes $1,2,\cdots,N$.

We write $\mathbb{P}(X=x\,|\,N)=\frac{1}{N}$ since the distribution is dependent on values of the parameters.

Fact 3.1:

$$\sum_{i=1}^{k}i=\frac{k(k+1)}{2}\text{ and }\sum_{i=1}^{k}i^2=\frac{k(k+1)(2k+1)}{6}.$$

Moments for Discrete Uniform Random Variable:

Mean: $\mathbb{E}X=\sum_{i=1}^{N}x\mathbb{P}(X=x\,|\,N)=\sum_{x=1}^{N}x\frac{1}{N}=\frac{N+1}{2}.$

Variance: $\mathrm{Var}X=\mathbb{E}X^2-(\mathbb{E}X)^2=\sum_{x=1}^{N}x^2\frac{1}{N}-\left(\frac{N+1}{2}\right)^2$

$$=\frac{(N+1)(N-1)}{12}.$$

Remark:

This distribution can be generalized so that the sample space is any range of integers, $N_0, N_0+1, \cdots, N_1$, with pmf $\mathbb{P}(X = x \,|\, N_0, N_1) = \dfrac{1}{(N_1 - N_0 + 1)}$. ||

Definition: Bernoulli Distribution

A Bernoulli trial is an experiment with two, and only two, possible outcomes. A random variable X has a **Bernoulli**(p) distribution if

$$X = \begin{cases} 1, \text{ with probability } p \\ 0, \text{ with probability } 1-p \end{cases} \text{ where } 0 \le p \le 1.$$

Moments for Bernoulli Random Variable:

Mean: $\mathbb{E}X = p$

Variance: $\text{Var}X = p(1-p)$.

Definition: Binomial Distribution

The binomial distribution is based on the idea of a Bernoulli trial. A random variable is said to be a **Binomial**(n, p) random variable if

$$\mathbb{P}(Y = y \,|\, n, p) = \binom{n}{y} p^y (1-p)^{n-y} \text{ where } y = 0,1,2,\cdots,n.$$

Theorem 3.2: Binomial Theorem

$$\forall x, y \in \mathbb{R} \text{ and } n \ge 0 \text{ an integer one has } (x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}.$$

Moments for Binomial Random Variable:

Mean: $\mathbb{E}X = np$.

Variance: $\text{Var}X = np(1-p)$.

MGF: $M_X(t) = \left(pe^t + (1-p)\right)^n$.

Definition: Poisson distribution

The Poisson distribution has a single parameter λ, sometimes called the intensity parameter. A random variable X, taking values in the nonnegative integers, has a **Poisson**(λ) distribution if $\mathbb{P}(X = x \,|\, \lambda) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0,1,\cdots$.

Moments for Poisson Random Variables:

Mean:
$$\mathbb{E}X = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$
$$= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = \lambda.$$

Variance: $\text{Var}X = \lambda$.

MGF: $M_X(t) = e^{\lambda(e^t - 1)}$.

Definition: Negative Binomial Distribution

In a sequence of independent Bernoulli(p) trials, let the random variable X denote the trial at which the rth success occuse, where r is a fixed integer.

If $\mathbb{P}(X = x \,|\, r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$ for $x = r, r+1, \cdots$ then X is said to

be a **Negative Binomial**(r, p) random variable.

The negative binomial distribution is sometimes defined in terms of the random variable $Y = X - r$.

Moments of Negative Binomial Random Variable:

Mean: $\mathbb{E}Y = r\dfrac{(1-p)}{p}$.

Variance: $\mathrm{Var}Y = \dfrac{r(1-p)}{p^2}$.

Definition: Geometric Distribution

The geometric distribution is a special case of the negative binomial distribution. A random variable X is said to has geometric distribution if $\mathbb{P}(X = x \,|\, p) = p(1-p)^{x-1}$ for $x = 1,2,\cdots$.

Moments of Geometric Random Variable:

Mean: $\mathbb{E}X = \dfrac{1}{p}$.

Variance: $\mathrm{Var}X = \dfrac{1-p}{p^2}$.

Remark:

The geometric distribution has an interesting property known as the "memoryless" property. For integers $s > t$, it is the case that $\mathbb{P}(X > s \,|\, X > t) = \mathbb{P}(X > s - t)$, i.e. the geometric distribution "forgets" what has occured.

3.2 Continuous Distributions

Definition: Continuous Uniform Distribution

The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a, b]$. Its pdf is given by $f(x \,|\, a, b) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$.

Moments of Continuous Uniform Random Variable:

Mean: $\mathbb{E}X = \displaystyle\int_a^b \frac{x}{b-a}dx = \frac{b+a}{2}$.

Variance: $\mathrm{Var}X = \displaystyle\int_a^b \frac{(x - \frac{b+a}{2})^2}{b-a}dx = \frac{(b-a)^2}{12}$.

Definition: Gamma Function

If α is a positive constant, the integral $\displaystyle\int_0^\infty t^{\alpha-1}e^{-t}dt$ is finite. The gamma function $\Gamma(\alpha) = \displaystyle\int_0^\infty t^{\alpha-1}e^{-t}dt$.

The gamma function satisfies many useful relationships, in particular, one has that

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha), \text{ for } \alpha > 0. \tag{3.1}$$

This can be verified through integration by parts. Moreover, we have $\Gamma(1) = 1$, we have for any integer $n > 0$, $\Gamma(n) = (n-1)!$. Furthermore, as a useful special case, we have $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Since the integrand $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ is positive, it follows immediately that

$$f(t) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}, 0 < t < \infty, \tag{3.2}$$

is a pdf. The full gamma family, however, has two paramaters and can be derived by changing variables to get the pdf of the random variable $X = \beta T$, where β is a positive constant. Upon doing this, we get the Gamma(α, β) family,

$$f(x\,|\,\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, 0 < x < \infty, \alpha > 0, \beta > 0. \tag{3.3}$$

The parameter α is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter β is called the scale parameter, since most of its influence is on the spread of the distribution.

Definition: Gamma Distribution

The gamma distribution is **Gamma**(α, β) with the pdf defined as in (3.3).

Moments of Gamma Random Variables:

Mean: $\mathbb{E}X = \alpha\beta$.

Variance: $\mathrm{Var}X = \alpha\beta^2$.

MGF: $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha$ for $t < \frac{1}{\beta}$.

There are a number of important special cases of the gamma distribution. If we set $\alpha = p/2$, where p is an integer, and $\beta = 2$, then the gamma pdf becomes

$$f(x\,|\,p) = \frac{1}{\Gamma(p/2)2^{p/2}}x^{(p/2)-1}e^{-x/2}, 0 < x < \infty, \tag{3.4}$$

which is the chi squared pdf with p degrees of freedom.

Definition: Chi Squared Distribution

A random variable is said to have a chi squared distribution with p degrees of freedom if it has pdf as in (3.4).

Moments of Chi Squared Random Variables:

Mean: $\mathbb{E}X = \alpha\beta = \frac{p}{2} \cdot 2 = p$.

Variance: $\mathrm{Var}X = \alpha\beta^2 = \frac{p}{2} \cdot 2^2 = 2p$.

MGF: $M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha = \left(\frac{1}{1-2t}\right)^{p/2}$, for $t < \frac{1}{2}$.

Another important special case of the gamma distribution is obtained when we set $\alpha = 1$. We then have

$$f(x\,|\,\beta) = \frac{1}{\beta}e^{-x/\beta}, 0 < x < \infty, \tag{3.5}$$

which is the exponential pdf with scale parameter β.

Definition:

A random variable is said to have an exponential distribution if its pdf is in the form $f(x\,|\,\beta) = \frac{1}{\beta}e^{-x/\beta}, 0 < x < \infty$.

Take $\lambda := \frac{1}{\beta}$, we now describe its moments:

Moments of Exponential Random Variable:

Mean: $\mathbb{E}X = \frac{1}{\lambda}$.

Variance: $\mathrm{Var}X = \frac{1}{\lambda^2}$.

The normal distribution, sometimes called the Gaussian distribution, plays a central role in a large body of statistics. There are three main reasons for this. First, the normal distribution and distributions associated with it are very tractable analytically. Second, the normal distribution has the familiar bell shape, whose symmetry makes it an appealing choice for many population models. Although there are many other distributions that are also bell-shaped, most do not possess the analytic tractability of the normal. Third, there is the **Central Limit Theorem**, which shows that, under mild conditions, the normal distribution can be used to approximate a large variety of distributions in large samples.

Definition: Normal Distribution

The normal distribution has two parameters, μ and σ^2, which are its mean and variance. The pdf of the normal distribution with mean μ and variance σ^2 is given by

$$f(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}, -\infty < x < \infty. \tag{3.6}$$

We denote $X \sim N(\mu,\sigma^2)$ as X has a normal distribution with mean μ and variance σ^2, the random variable $Z := (X-\mu)/\sigma$ has a $N(0,1)$ distribution, aslo known as the standard normal.

Moments of Gaussian Random Variable:

Mean: $\mathbb{E}X = \mu$.

Variance: $\mathrm{Var}X = \sigma^2$.

MGF: $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$.

Definition: Beta Distribution

The beta family of distributions is a continuous family on (0,1) indexed by two parameters. The Beta(α,β) has pdf

$$f(x\,|\,\alpha,\beta) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1, \alpha,\beta > 0, \tag{3.7}$$

where $B(\alpha,\beta)$ denotes the beta function $B(\alpha,\beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$.

Remark:

The beta function is related to the gamma function $B(\alpha,\beta)=\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. ||

Moments of Beta Random Variables:

Mean: $\mathbb{E}X=\dfrac{\alpha}{\alpha+\beta}$.

Varianace: $\mathrm{Var}X=\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Definition: Cauchy Distribution

The Cauchy distribution is a symmetric, bell-shaped distribution on $(-\infty,\infty)$ with pdf

$$f(x\,|\,\theta)=\frac{1}{\pi}\frac{1}{1+(x-\theta)^2},\ -\infty<x,\theta<\infty. \tag{3.8}$$

Remark:

The mean of the Cauchy distribution does not exist. In fact, no moments of the Cauchy distributions exist, or, all absolute moments are ∞. In particular, the moment generating function does not exist. ||

Definition: Lognormal Distribution

If X is a random variable whose log is normally distribued, i.e. $\log X\sim N(\mu,\sigma^2)$, then we say X has a lognormal distribution. The pdf is

$$f(x\,|\,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-(\log x-\mu)^2/(2\sigma^2)},0<x<\infty,-\infty<\mu<\infty,\sigma>0.$$

Moments of Lognormal Random Variables:

Mean: $\mathbb{E}X=e^{\mu+(\sigma^2/2)}$.

Variance: $\mathrm{Var}X=e^{2(\mu+\sigma^2)}-e^{2\mu+\sigma^2}$.

Definition: Double Exponential Distribution

The double exponential distribution is formed by reflecting the exponential distribution around its mean. The pdf is

$$f(x\,|\,\mu,\sigma)=\frac{1}{2\sigma}e^{-|x-\mu|/\sigma},\ -\infty<x<\infty,-\infty<\mu<\infty,\sigma>0.$$

Moments of Double Exponential Distribution:

Mean: $\mathbb{E}X=\mu$.

Variance: $\mathrm{Var}X=2\sigma^2$.

3.3 Exponential Families

Definition: Exponential Family

A family of pdfs or pmfs is called an exponential family if it can be expressed as

$$f(x\,|\,\theta)=h(x)c(\theta)\exp\Big\{\sum_{i=1}^{k}w_i(\theta)t_i(x)\Big\}, \tag{3.9}$$

where $h(x)\geq 0$, $t_1(x),\cdots,t_k(x)$ are real-valued functions of the observation x (they cannot depend on θ), and $c(\theta)\geq 0$, $w_1(\theta),\cdots,w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (they cannot depend on x).

Remark:

The continuous families — normal, gamma, and beta, the discrete families — binomial, Poisson, and negative binomial, are all exponential families. ||

Theorem 3.3:

If X is a random variable with pdf or pmf of the form (3.9), then

$$\mathbb{E}\Big(\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)\Big) = -\frac{\partial}{\partial\theta_j}\log c(\theta). \tag{3.10}$$

and

$$\mathrm{Var}\Big(\sum_{i=1}^{k}\frac{\partial w_i(\theta)}{\partial\theta_j}t_i(X)\Big) = -\frac{\partial^2}{\partial\theta_j^2}\log c(\theta) - \mathbb{E}\Big(\sum_{i=1}^{k}\frac{\partial^2 w_i(\theta)}{\partial\theta_j^2}t_i(X)\Big). \tag{3.11}$$

Although the equations are so ugly, when applied to specific cases they can work out quite nicely. Their advantage is that we can replace integration or summation by differentiation, which is often more straightforward.

In general, the set of x values for which $f(x\,|\,\theta) > 0$ cannot depend on θ in an exponential family. The entire definition of the pdf or pmf must be incorporated into the form (3.9). This is most easily accomplished by incoporating the range of x into the expression for $f(x\,|\,\theta)$ through the use of an indicator function.

Definition: Indicator Function

The indicator function of a set A, most often denoted by $I_A(x)$, is the function

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

An exponential family is sometimes reparameterized as

$$f(x\,|\,\eta) = h(x)c^*(\eta)\exp\Big\{\sum_{i=1}^{k}\eta_i t_i(x)\Big\}. \tag{3.12}$$

Here the $h(x)$ and $t_i(x)$ functions are the same as in the original parameterization (3.9). The set $\mathscr{H} := \left\{\eta = (\eta_1, \cdots, \eta_k) \,\middle|\, \int_{-\infty}^{\infty} h(x)\exp\Big\{\sum_{i=1}^{k}\eta_i t_i(x)\Big\}dx < \infty\right\}$ is called the natural parameter space for the family. The natural parameterization and the natural parameter space have many useful mathematical properties, for example, $\mathscr{H}$ is convex.

In (3.9) it is often the case that the dimension of the vector θ is k, the number of terms in the sum of the exponent. This need not be so, and it is possible for the dimension of the vector θ to be less than k. Such an exponential family is called a curved exponential family.

Defnition: Curved Exponentia Family

A curved exponential family is a family of densities of the form (3.9) for which the dimension of the vector θ, $\dim\theta < k$. If $\dim\theta = k$ the family is a full exponential family.

Although the fact that the parameter space is a lower-dimensional space has some influence on the properties of the family, we will see that curved families still enjoy many of the properties of full families. In particular, **Theorem 3.3** applies to curved exponential families. For more introuduction to the exponential families, see

Lehmann (1986, Section 2.7) or Lehmann and Casella (1998, Section 1.5 and Note 1.10.6).

4.1 Joint and Marginal Distributions

Definition: Random Vector

An n-dimensional random vector is a function from a sample space S to $\mathbb{R}^n$.

Suppose, for example, that with each point in a sample space we associate an ordered pair of numbers, i.e. a point $(x, y) \in \mathbb{R}^2$. Then we have defined a two-dimensional (or bivariate) random vector (X, Y).

A random vector is called a discrete random vector when it has a countable number of possible values. For a discrete random vector, the function $f(x, y)$ defined by $f(x, y) = \mathbb{P}(X = x, Y = y)$ can be used to compute any probabilities defined in terms of (X, Y).

Definition: Joint Probability Mass Function

Let (X, Y) be a discrete bivariate random vector. Then the function $f(x, y)$ from $\mathbb{R}^2$ to $\mathbb{R}$ defined by $f(x, y) = \mathbb{P}(X = x, Y = y)$ is called the joint probability mass function or joint pmf of (X, Y). If it is necessary to stress the fact that f is the joint pmf of the vector (X, Y) rather than some other vector, the notattion $f_{X,Y}(x, y)$ will be used.

The joint pmf of (X, Y) completely defines the probability distribution of the random vector (X, Y), just as the pmf of a discrete univariate random variable completely. The joint pmf can be used to compare the probability of any event defined in terms of (X, Y). Let A be any subset of $\mathbb{R}^2$. Then $\mathbb{P}((X, Y) \in A) = \sum_{(x,y)\in A} f(x, y)$.

Expectations of functions of random vectors are computed just as with univariate random variable. Let $g(x, y)$ be a real-valued function defined for all possible values (x, y) of the discrete random vector (X, Y). Then $g(X, Y)$ is itself a random variable and its expected value $\mathbb{E}g(X, Y)$ is given by $\mathbb{E}g(X, Y) = \sum_{(x,y)\in\mathbb{R}^2} g(x, y)f(x, y)$. Moreover, for $g_1(x, y)$ and $g_2(x, y)$ being two functions and a, b, c being constants, then $\mathbb{E}(ag_1(x, y) + bg_2(x, y) + c) = a\mathbb{E}g_1(X, Y) + b\mathbb{E}g_2(x, y) + c$.

Definition: Marginal PMF

Let (X, Y) be a discrete bivariate random vector with joint pmf $f_{X,Y}(x, y)$. Then the marginal pmfs of X and Y, as $f_X(x) = \mathbb{P}(X = x)$ and $f_Y(y) = \mathbb{P}(Y = y)$, are given by $f_X(x) = \sum_{y\in\mathbb{R}} f_{X,Y}(x, y)$ and $f_Y(y) = \sum_{x\in\mathbb{R}} f_{X,Y}(x, y)$.

The marginal pmf of X or Y is the same as the pmf of X and Y as we have mentioned before. The marginal pmf of X or Y can be used to compute probabilities or expectations that involve only X or Y. But to compute a probability or expectation that simultaneously involves both X and Y, we must use the joint pmf of X and Y.

Remark:

The marginal distributions of X and Y, described by the marginal pmfs $f_X(x)$ and $f_Y(y)$, do not completely describe the joint distribution of X and Y. Indeed,

there are many different joint distributions that have the same marginal distributions. Thus, it is hopeless to try to determine the joint pmf, $f_{X,Y}(x,y)$, from the knowledge of only the marginal pmfs, $f_X(x)$ and $f_Y(y)$. ||

To this point we have discussed discrete bivariate random vectors. We can also consider random vectors whose components are continuous random variables. The probability distribution of a continuous random vector is usually described using a density function, as in the univariate case.

Definition: Joint Probability Density Function

A function $f(x,y)$ from $\mathbb{R}^2$ into $\mathbb{R}$ is called a joint probability density function or joint pdf of the continuous bivariate random vector (X,Y) if, for every $A \subseteq \mathbb{R}^2$, $\mathbb{P}((X,Y) \in A) = \int_A \int f(x,y)dxdy$.

A joint pdf is used just like a univariate pdf except now the integrals are double integrals over sets in the plane. The notation $\iint_A$ simply means that the limits of integration are set so that the function is integrated over all $(x,y) \in A$. Expectations of functions of continuous random vectors are defined as in the discrete case with integrals replacing sums and the pdf replacing the pmf. That is, if $g(x,y)$ is a real-valued function, then the expectated value of $g(X,Y)$ is defined to be

$$\mathbb{E}g(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy. \qquad (4.1)$$

It is important to realize that the joint pdf is defined for all $(x,y) \in \mathbb{R}^2$. The pdf may equal 0 on a large set A if $\mathbb{P}((X,Y) \in A) = 0$ but the pdf is defined for the points in A.

Definition: Marginal PDF

The marginal probability functions of X and Y are also defined as in the discrete case with integrals replacing sums. That is,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy, -\infty < x < \infty; f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx, -\infty < y < \infty.$$

The joint probability distribution of (X,Y) can be completely described with the joint cdf rather than with the joint pmf or joint pdf.

Definition: Joint CDF

The joint cdf is the function $F(x,y)$ defined by $F(x,y) = \mathbb{P}(X \le x, Y \le y)$ for all $(x,y) \in \mathbb{R}^2$.

The joint cdf is usually not very handy to use for a discrete random vector. But for a continuous bivariate random vector we have the important relationship, as in the univariate case, $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(s,t)dtds$. Recall the **bivariate Fundamental Theorem of Calculus**, this implies that $\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$ at continuity points of $f(x,y)$. The relationship is useful in situations where an expression for $F(x,y)$ can be found. The mixed partial derivative can be computed to find the joint pdf.

4.2 Conditional Distributions and Independence

Often when two random variables, (X, Y), are observed, the values of the two variables are related. Information about the value of X gives us some information about the value of Y even if it does not tell us the value of Y directly. Conditional probabilities regarding Y given knowledge of the X value can be computed using the joint distribution of (X, Y). Sometimes, however, knowledge about X gives us no information about Y. We will discuss these topics concerning conditional probabilities in this subsection.

Definition: Conditional Probability Mass Function

Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$.

For any x such that $\mathbb{P}(X = x) = f_X(x) > 0$, the conditional pmf of Y given that $X = x$ is the function of y denoted by $f(y \mid x)$ and defined by

$$f(y \mid x) = \mathbb{P}(Y = y \mid X = x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $\mathbb{P}(Y = y) = f_Y(y) > 0$, the conditional pmf of X given that $Y = y$ is the function of x denoted by $f(x \mid y)$ and defined by

$$f(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{f(x, y)}{f_Y(y)}.$$

Definition: Conditional Probability Density Function

Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$.

For any x such that $f_X(x) > 0$, the conditional pdf of Y given that $X = x$ is the function of y denoted by $f(y \mid x)$ and defined by $f(y \mid x) = \dfrac{f(x, y)}{f_X(x)}$.

For any y such that $f_Y(y) > 0$, the conditional pdf of X given that $Y = y$ is the function of x denoted by $f(x \mid y)$ and defined by $f(x \mid y) = \dfrac{f(x, y)}{f_Y(y)}$.

Definition: Conditional Expected Value

If $g(Y)$ is a function of Y, then the conditional expected value of $g(Y)$ given that $X = x$ is denoted by $\mathbb{E}(g(Y) \mid x)$ and is given by

$\mathbb{E}(g(Y) \mid x) = \sum_y g(y) f(y \mid x)$ and $\mathbb{E}(g(Y) \mid x) = \int_{-\infty}^{\infty} g(y) f(y \mid x) dy$ in the discrete and the continuous cases, respectively.

Definition: Independent Random Variables

Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called independent random variables if, $\forall x, y \in \mathbb{R}, f(x, y) = f_X(x) f_Y(y)$.

If X and Y are independent, the conditional pdf of Y given $X = x$ is

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} = \frac{f_X(x) f_Y(y)}{f_X(x)} = f_Y(y), \tag{4.2}$$

regardless of the value of x. Thus, for any $A \subseteq \mathbb{R}$ and $x \in \mathbb{R}$,

$$\mathbb{P}(Y \in A \,|\, x) = \int_A f(y \,|\, x)dy = \int_A f_Y(y)dy = \mathbb{P}(Y \in A).$$

The knowledge that $X = x$ gives us no additional information about Y.

Lemma 4.1: Criterion for Independent

Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent random variables $\Leftrightarrow$ there exist functions $g(x)$ and $h(y)$ such that $\forall x, y \in \mathbb{R}, f(x, y) = g(x)h(y)$.

Proof:

"$\Leftarrow$":

Taking $g(x) = f_X(x)$ and $h(y) = h_Y(y)$ yields this direction.

"$\Rightarrow$":

Let us prove the case for continuous random variables, while for discrete case we only need to replace the integrals with sums.

Consider now $\int_{-\infty}^{\infty} g(x)dx = c$ and $\int_{-\infty}^{\infty} h(y)dy = d$, where the constants c and d satisfy

$$\begin{aligned} cd &= \left(\int_{-\infty}^{\infty} g(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)dxdy \qquad \textbf{(Fubini's Theorem)} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1. \qquad \text{(Since } f(x, y) \text{ joint pdf)} \end{aligned}$$

Furthermore, the marginal pdfs are given by

$$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y)dy = g(x)d \text{ and } f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y)dx = h(y)c.$$

Thus, we have $f(x, y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y)$.

□

Certain probabilities and expectations are easy to calculate if X and Y are independent, as the following theorem states:

Theorem 4.2:

Let X and Y be independent random variables. Then

(i) For any $A, B \subseteq \mathbb{R}$, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.

(ii) If $g(x)$ is a function only of x and $h(y)$ is a function only of y. Then $\mathbb{E}(g(X)h(Y)) = (\mathbb{E}g(X))(\mathbb{E}h(Y))$.

Proof:

(ii):

For continuous random variables, part (ii) is proved by nothing but

$$\mathbb{E}(g(X)h(Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x, y)dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \quad \text{(Independent)}$$
$$= \int_{-\infty}^{\infty} h(y)f_Y(y)\int_{-\infty}^{\infty} g(x)f_X(x)dxdy$$
$$= \left(\int_{-\infty}^{\infty} g(x)f_X(x)dx\right)\left(\int_{-\infty}^{\infty} h(y)f_Y(y)dy\right) \text{ (Fubini)}$$
$$= (\mathbb{E}g(X))(\mathbb{E}g(Y)).$$

The result for discrete case is valid by replacing the integrals by sums.

(i):

Let $C := \{(x, y) \mid x \in A, y \in B\}$ and let $g(x)$ be the indicator function of the set A while letting $h(y)$ being the indicator function of the set B. Then

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}((X, Y) \in C) = \mathbb{E}(g(X)h(Y))$$
$$= (\mathbb{E}g(X))(\mathbb{E}h(Y)) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

□

Theorem 4.3:

Let X and Y be independent random variables with mgf $M_X(t)$ and $M_Y(t)$. Then the mgf of the random variable $Z = X + Y$ is given by $M_Z(t) = M_X(t)M_Y(t)$.

Proof:

Using the definition of mgf and the result of **Theorem 4.2**, we have

$$M_Z(t) = \mathbb{E}e^{tZ} = \mathbb{E}e^{t(X+Y)} = \mathbb{E}(e^{tX}e^{tY}) = (\mathbb{E}e^{tX})(\mathbb{E}e^{tY}) = M_X(t)M_Y(t).$$

□

4.3 Bivariate Transformations

In 2.1, methods of finding the distribution of a function of random variable were discussed. In this subsection we extend these ideas to the case of bivariate random vectors. Let us first state a results of Normal and Poisson random variables.

Theorem 4.4: Normal Transformation

Let $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\gamma, \tau^2)$ be independent random variables. Then the random variale $Z = X + Y$ has a $N(\mu + \gamma, \sigma^2 + \tau^2)$ distribution.

Theorem 4.5: Poisson Transformation

If $X \sim \text{Poisson}(\theta)$ and $Y \sim \text{Poisson}(\lambda)$ and X and Y are independent, then $X + Y \sim \text{Poisson}(\theta + \lambda)$.

In fact, if two random variables are independent, then the transformations of them with the other not included, is also a random variable, and, are also independent.

Theorem 4.6:

Let X and Y be independent random variables. Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then the random variables $U = g(X)$ and $V = h(Y)$ are independent.

Proof:

We will prove the case for consinuous, the discrete case follows analogously. Assume that U and Y are continuous random variables. For any $u \in \mathbb{R}$ and $v \in \mathbb{R}$, we define $A_u := \{x \mid g(x) \leq u\}$ and $B_v = \{y \mid h(y) \leq v\}$. Then the joint cdf of (U, V) is given by

$$\begin{aligned} F_{U,V}(u,v) &= \mathbb{P}(U \le u, V \le v) && \text{(Definition of cdf)} \\ &= \mathbb{P}(X \in A_u, Y \in B_v) && \text{(Definition of } U \text{ and } V\text{)} \\ &= \mathbb{P}(X \in A_u)\mathbb{P}(Y \in B_v). && \text{(\textbf{Theorem 4.2} (i))} \end{aligned}$$

The joint pdf of (U, V) is

$$f_{U,V}(u,v) = \frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v) = \left(\frac{d}{du}\mathbb{P}(X \in A_u)\right)\left(\frac{d}{dv}\mathbb{P}(Y \in B_v)\right).$$

According to **Lemma 4.1**, independence follows.

□

4.4 Hierarchical Models and Mixture Distributions

In the case we have seen so far, a random variable has a single distribution, possibly depending on parameters. While, in general, a random variable can have only one distribution, it is often easier to model a situation by thinking of things in a hierarchy.

Sometimes, calculation can be greatly simplified by using the following theorem. Recall that $\mathbb{E}(X \mid y)$ is a function of y and $\mathbb{E}(X \mid Y)$ is a random variable whose value depends on the value of Y.

Theorem 4.7: Conditional Expectation Identity

If X and Y are any two random variables. Then $\mathbb{E}X = \mathbb{E}(\mathbb{E}(X \mid Y))$.

Proof:

Let $f(x, y)$ denote the joint pdf of X and Y. By definition, we have

$$\mathbb{E} = \int\int x f(x,y) dx dy = \int \left[\int x f(x \mid y) dx\right] f_Y(y) dy,$$

where $f(x \mid y)$ and $f_Y(y)$ are the conditional pdf of X given $Y = y$ and the marginal pdf of Y, respectively. Notice that the integral in the bracket is the conditional expectation $\mathbb{E}(X \mid y)$, rewrite the above equation

$$\mathbb{E}X = \int \mathbb{E}(X \mid y) f_Y(y) dy = \mathbb{E}(\mathbb{E}(X \mid Y)),$$

as we desired. Replacing the integrals by sums yields the discrete case.

□

The term mixture distribution in the title of this subsection refers to a distribution arising from a hierarchical structure. Although there is no standardized definition for this term, we will use the following definition, which seems to be a popular one.

Definition: Mixture Distribution

A random variable X is said to have a mixture distribution if the distribution of X depends on a quantity that also has a distribution.

We have dealt with the expectation, now let us deal with the calculation of the variance. We can make use of a formula for conditional variances, similar to the one we did for conditional expectations.

Theorem 4.8: Conditional Variance Identity

For any two random variables X and Y, $\mathrm{Var}X = \mathbb{E}(\mathrm{Var}(X \mid Y)) + \mathrm{Var}(\mathbb{E}(X \mid Y))$.

Proof:

By definition, one has

$$\mathrm{Var}X = \mathbb{E}((X - \mathbb{E}X)^2) = \mathbb{E}\big(X - \mathbb{E}(X \mid Y) + \mathbb{E}(X \mid Y) - \mathbb{E}X^2\big)$$

$$= \mathbb{E}\big((X - \mathbb{E}(X \mid Y))^2\big) + \mathbb{E}\big((\mathbb{E}(X \mid Y) - \mathbb{E}X)^2\big)$$
$$+2\mathbb{E}\big((X - \mathbb{E}(X \mid Y)\big)\big(\mathbb{E}(X \mid Y) - \mathbb{E}Y)\big).$$

We leave the step in prove that $2\mathbb{E}\big((X - \mathbb{E}(X \mid Y)\big)\big(\mathbb{E}(X \mid Y) - \mathbb{E}Y)\big) = 0$ to the reader. The above equation, by definition,
$\mathbb{E}\big((X - \mathbb{E}(X \mid Y))^2\big) = \mathbb{E}(\mathrm{Var}(X \mid Y))$, $\mathbb{E}\big((\mathbb{E}(X \mid Y) - \mathbb{E}X)^2\big) = \mathrm{Var}(\mathbb{E}(X \mid Y))$.

□

4.5 Covariance and Correlation

In earlier subsections, we have discussed the absence or presence of a relationship between two random variables, independence or nonindependence. But if there is a relationship, the relationship may be strong or weak. In this subsection we discuss two numerical measures of the strength of a relationship between two random variables, the covariance and correlation.

If there is no misleading, we shall always use, for two random variables X and Y, $\mu_X := \mathbb{E}X$ and $\sigma_X^2 := \mathrm{Var}X$, $\mu_Y := \mathbb{E}Y$, $\sigma_Y^2 := \mathrm{Var}Y$, where $0 < \sigma_X^2, \sigma_Y^2 < \infty$.

Definition: Covariance

The covariance of X and Y is the value $\mathrm{Cov}(X, Y) := \mathbb{E}\big((X - \mu_X)(Y - \mu_Y)\big)$.

Definition: Correlation

The correlation of X and Y is the value $\rho_{XY} = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$. The value ρ_{XY} is also called the correlation coefficient.

While the covariance could be any number, the correlation is always between -1 and 1, with the values -1 and 1 indicating a perfect linear relationship between X and Y. We now prove another version of covariance.

Theorem 4.9:

For any random variables X and Y, $\mathrm{Cov}(X, Y) = \mathbb{E}XY - \mu_X \mu_Y$.

Proof:

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}(X - \mu_X)(Y - \mu_Y) \\
&= \mathbb{E}(XY - \mu_X Y - \mu_Y YX + \mu_X \mu_Y) && \text{(Expand the product)} \\
&= \mathbb{E}XY - \mu_X \mathbb{E}Y - \mu_Y \mathbb{E}X + \mu_X \mu_Y && (\mu_X, \mu_Y \text{ are constants}) \\
&= \mathbb{E}XY - \mu_X \mu_Y - \mu_Y \mu_X + \mu_X \mu_Y = \mathbb{E}XY - \mu_X \mu_Y.
\end{aligned}$$

□

In the next three theorems we describe some of the fundamental properties of covariance and correlaion.

Theorem 4.10:

If X and Y are independent random variables, then $\mathrm{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$.

Proof:

Since X and Y are independent, one has $\mathbb{E}XY = (\mathbb{E}X)(\mathbb{E}Y)$. Thus,

$$\mathrm{Cov}(X, Y) = \mathbb{E}XY - (\mathbb{E}X)(\mathbb{E}Y) = 0.$$

It follows that $\rho_{XY} = 0$ as well.

□

Remark:

The converse is not true in general, there are some nonindependent random variables X and Y with $\text{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$. ||

Covariance plays an important role in understanding the variation in sums of random variables, as the next theorem suggests.

Theorem 4.11:

If X and Y are any two random variables and a and b are any two constants, then

$$\text{Var}(aX + bY) = a^2\text{Var}X + b^2\text{Var}Y + 2ab\text{Cov}(X, Y).$$

If X and Y are further assumed to be independent, then

$$\text{Var}(aX + bY) = a^2\text{Var}X + b^2\text{Var}Y.$$

Covariance and correlation measure only a particular kind of linear relationship that will be described in the following theorem.

Theorem 4.12:

For any random variables X and Y.

(i) $-1 \le \rho_{XY} \le 1$.

(ii) $|\rho_{XY}| = 1 \Leftrightarrow \exists a \neq 0$ and b constants such that $\mathbb{P}(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, if $\rho_{XY} = -1$, then $a < 0$.

Later we will prove the Cauchy-Schwartz inequality which has a direct consequence that ρ_{XY} is bounded between -1 and 1.

The intuition of **Theorem 4.12** is that, if there is a line $y = ax + b$ with $a \neq 0$, such that the values of (X, Y) have a high probability of being near the line, then the correlation between X and Y will be near 1 or -1. But if no such line exists, the correlation will be near 0.

We close this subsection by introducing a very important bivariate distribution in which the correlation coefficient arises naturally as a parameter.

Definition: Bivariate Normal pdf

Let $-\infty < \mu_X, \mu_Y, \sigma_X, \sigma_Y < \infty$ and $-1 < \rho < 1$. The bivariate normal pdf with means μ_X and μ_Y, variances σ_X^2 and σ_Y^2, and correlation ρ is the bivariate pdf given by

$$f(x, y) = \left(2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}\right)^{-1}$$
$$\cdot \exp\Big\{ -\frac{1}{2(1-\rho^2)}\Big((\frac{x-\mu_X}{\sigma_X})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^2\Big)\Big\}$$

for $-\infty < x, y < \infty$.

This formula is disgusting but often used. We now give some properties of it:

Properties:

(i) The marginal distribution of X (resp. Y) is $N(\mu_X, \sigma_X^2)$ (resp. $N(\mu_Y, \sigma_Y^2)$).

(ii) The correlation between X and Y is $\rho_{XY} = \rho$.

(iii) For any constants a and b, the distribution $aX + bY$ is $N(a\mu_X + b\mu_Y, a^2\sigma_X^2 b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$.

4.6 Multivariate Distributions

Now we extend the discussion so far for bivariate random variables into the case of more, finite, or even countable random variables. Namely, we call X a random vector if $X = (X_1, \cdots, X_n)$ with each entry being a random variable,

Definition: Joint pmf

The joint pmf of the random vector $X = (X_1, \cdots, X_n)$ is a function defined by $f(x) := f(x_1, \cdots, x_n) = \mathbb{P}(X_1 = x_1, \cdots, X_n = x_n)\ \forall (x_1, \cdots, x_n) \in \mathbb{R}^n$.

The marginal pdf or pmf of any subset of the coordinates of $(X_1, \cdots, X_n)$ can be computed by integrating or summing the joint pdf or pmf over all possible values of the other coordinates.

The conditional pdf or pmf of a subset of the coordinates of $(X_1, \cdots, X_n)$ given the values of the remaining coordinates is obtained by dividing the joint pdf or pmf by the marginal pdf or pmf of the remaining coordinates.

We now introduce an important family of discrete multivariate distributions. This family generalizes the binomial family to the situation in whcih each trial has n distinct possible outcomes rather than two.

Definition: Multinomial Distributions

Let n and m be positive integers and let $p_1, \cdots, p_n$ be numbers satisfying $0 \le p_i \le 1$, where $i = 1, \cdots, n$ and $\sum_{i=1}^{n} p_i = 1$. Then the random vector $(X_1, \cdots, X_n)$ has a multinomial distribution with m trials and cell probabilities $p_1, \cdots, p_n$ if the joint pmf of $(X_1, \cdots, X_n)$ is given by

$$f(x_1, \cdots, x_n) = \frac{m!}{x_1! \cdot \cdots \cdot x_n!} p_1^{x_1} \cdot \cdots \cdot p_n^{x_n} = m! \prod_{i=1}^{n} \frac{p_i^{x_i}}{x_i!}$$ on the set of $(x_1, \cdots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^{n} x_i = m$.

The factor $\frac{m!}{x_1! \cdot \cdots \cdot x_n!}$ is called the multinomial coefficient. It is the number of ways that m objects can be divided into n groups with x_1 in the first group, x_2 in the second group, …, and x_n in the nth group. A generalization of the Binomial Theorem is the Multinomial Theorem.

Theorem 4.13: Multinomial Theorem

Let m and n be positive integers. Let A be the set of vectors $x = (x_1, \cdots, x_n)$ such that each x_i is a nonnegative integer and $\sum_{i=1}^{n} x_i = m$. Then, for any real numbers $p_1, \cdots, p_n$, one has $(p_1 + \cdots + p_n)^m = \sum_{x \in A} \frac{m!}{x_1! \cdot \cdots \cdot x_n!} p_1^{x_1} \cdot \cdots \cdot p_n^{x_n}$.

This theorem shows that a multinomial pmf sums to 1. The set A is the set of points with positive probability hence the sum of the pmf over all those points is, by this theorem, $(p_1 + \cdots + p_n)^m = 1^m = 1$.

Recall the statistical independence we introduced before is between two random variables, we now extend them to countable case, and we then state the generalized results we see earlier.

Definition: Mutually Independent

Let $X_1, \cdots, X_n$ be random vectors with joint pdf or pmf $f(x_1, \cdots, x_n)$. Let $f_{X_i}(x_i)$ denote the marginal pdf or pmf of X_i. Then $X_1, \cdots, X_n$ are called mutually independent random vectors if for every $(x_1, \cdots, x_n)$, one has $f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdot \cdots \cdot f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$. If the X_i's are all one-dimensional, then $X_1, \cdots, X_n$ are called mutually independent random variables.

Theorem 4.14: Expectation

Let $X_1, \cdots, X_n$ be mutually independent random variables. Let $g_1, \cdots, g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i for $i = 1, \cdots, n$. Then $\mathbb{E}(g_1(X_1) \cdot \cdots \cdot g_n(X_n)) = \big(\mathbb{E}g_1(X_1) \cdot \cdots \cdot \mathbb{E}g_n(X_n)\big)$.

Theorem 4.15: MGF

Let $X_1, \cdots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \cdots, M_{X_n}(t)$. Let $Z = X_1 + \cdots + X_n$. Then the mgf of Z is $M_Z(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t)$. In particular, if $X_1, \cdots, X_n$ all have the same distribution with mgf $M_X(t)$, then $M_Z(t) = \big(M_X(t)\big)^n$.

Corollary 4.16:

Let $X_1, \cdots, X_n$ be mutually independent random variables with mgfs $M_{X_1}(t), \cdots, M_{X_n}(t)$. Let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ be fixed constants. Let $Z = (a_1X_1 + b_1) + \cdots + (a_nX_n + b_n)$. Then the mgf of Z is given by $M_Z(t) = \big(e^{t(\sum b_i)}\big)M_{X_1}(a_1t) \cdot \cdots M_{X_n}(a_nt)$.

Proof:

From the definition, the mgf of Z is

$$\begin{aligned} M_Z(t) &= \mathbb{E}e^{tZ} \\ &= \mathbb{E}e^{t\sum(a_iX_i+b_i)} \\ &= \big(e^{t(\sum b_i)}\big)\mathbb{E}(e^{ta_1X_1} \cdot \cdots \cdot e^{ta_nX_n}) && \text{(Properties of exponential)} \\ &= \big(e^{t(\sum b_i)}\big)M_{X_1}(a_1t) \cdot \cdots \cdot M_{X_n}(a_nt) && \textbf{(Theorem 4.15)} \end{aligned}$$

result follows.

□

Undoubtedly, the most important application of **Corollary 4.16** is to the case of normal random variables. A linear combination of independent normal random variables is normally distributed.

Corollary 4.16.1:

Let $X_1, \cdots, X_n$ be mutually independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ be fixed constants. Then

$$Z = \sum_{i=1}^{n}(a_iX_i + b_i) \sim N\Big(\sum_{i=1}^{n}(a_i\mu_i + b), \sum_{i=1}^{n}a_i^2\sigma_i^2\Big).$$

Proof:

Recall that the mgf of a $N(\mu, \sigma^2)$ random variable is $M(t) = e^{\mu t+\sigma^2 t^2/2}$.
Substituting into the expression of **Corollary 4.16** yields

$$M_z(t) = \left(e^{t(\sum b_i)}\right) e^{\mu_1 a_1 t+\sigma_1^2 a_1^2 t^2/2} \cdot \ldots \cdot e^{\mu_n a_n t+\sigma_n^2 a_n^2 t^2/2}$$
$$= e^{(\sum (a_i\mu_i+b_i)t+(\sum a_i^2\sigma_i^2)t^2/2},$$

the mgf of the indicated normal distribution.

□

Theorem 4.17: Criterion for Independence

Let $X_1, \cdots, X_n$ be random vectors. Then $X_1, \cdots, X_n$ are mutually independent random vectors ⇔ there exists functions $g_i(x_i)$, for $i = 1,\cdots, n$, such that the joint pdf or pmf of $(X_1, \cdots, X_n)$ can be written as
$f(x_1, \cdots, x_n) = g_1(x_1) \cdot \cdots \cdot g_n(x_n)$.

Theorem 4.18:

Let $X_1, \cdots, X_n$ be independent random vectors. Let $g_i(x_i)$ be a function only of x_i for $i = 1,\cdots, n$. Then the random variables $U_i = g_i(X_i)$ for $i = 1,\cdots, n$, are mutually independent.

4.7 Inequalities

One of the most important task for us in either probability or statistics, is to find that if, for example, the variation of a given random variable, with high probability (say 99%), is bounded by a certain number. To this end, a lot of inequalities are needed. In this subsection, we shall introduce some important inequalities. Note that the inequalities are divided into categories numerical and functional, the former one is determined by "numbers" while the second one is determined by the "operating functions" according to the name.

Lemma 4.19:

Let a and b be positive numbers, and let p and q be any positive numbers (necessarily greater than 1) such that $\frac{1}{p} + \frac{1}{q} = 1$. Then $\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$ with equality if and only if $a^p = b^q$.

One of the most important variations for the **Lemma 4.19** is the famous Hölder's inequality.

Theorem 4.20: Hölder's Inequality

Let X and Y be any two random variables and let $\frac{1}{p} + \frac{1}{q} = 1$ for both p and q greater than 1. Then $|\mathbb{E}XY| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^p)^{1/p}(\mathbb{E}|Y|^q)^{1/q}$.

In fact, this is the idea derived from functional analysis, where we have a norm for the vectors having this property. Perhaps the most famous special case of Hölder's inequality is that for which $p = q = 2$. This is called the Cauchy-Schwarts Inequality.

Theorem 4.21: Cauchy-Schwartz Inequality

For any two random variables X and Y,

$$|\mathbb{E}XY| \leq \mathbb{E}|XY| \leq (\mathbb{E}|X|^2)^{1/2}(\mathbb{E}|Y|^2)^{1/2}.$$

Our next named inequality is similar in spirit to Hölder's inequality, and, in fact, follows from it.

Theorem 4.22: Minkowski's Inequality

Let X and Y be any two random variables. Then for $1 \leq p < \infty$, one has

$$(\mathbb{E}\,|X+Y|^p)^{1/p} \leq (\mathbb{E}\,|X|^p)^{1/p} + (\mathbb{E}\,|Y|^p)^{1/p}.$$

Now we introduce the functional inequalities, these inequalities rely on the property of convexity. For example, one of the most famous Jensen's Inequality.

Definition: Convex

A function $g(x)$ is convex if $\forall 0 < \lambda < 1$ and $\forall x, y$,

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

Theorem 4.23: Jensen's Inequality

For any random variable X, if $g(x)$ is a convex function, then $\mathbb{E}g(X) \geq g(\mathbb{E}X)$. With equality holds $\Leftrightarrow$ for every line $a + bx$ that is tangent to $g(x)$ at $x = \mathbb{E}X$, $\mathbb{P}(g(X) = a + bX) = 1$.

One immediate application of Jensen's Inequality is to show that $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$. Since $g(x) = x^2$ is convex. Moreover, if x is positive, then $1/x$ is convex; hence $\mathbb{E}(1/X) \geq 1/\mathbb{E}X$.

We close our section with an inequality that merely exploits the definition of covariance, but sometimes proves to be useful. If X is a random variable with finite mean μ and $g(x)$ is a nondecreasing function, then $\mathbb{E}(g(X)(X-\mu)) \geq 0$.

Theorem 4.24: Covariance Inequality

Let X be any random variable and $g(x)$ and $h(x)$ any functions such that $\mathbb{E}g(X)$, $\mathbb{E}h(X)$, and $\mathbb{E}(g(X)h(X))$ exist. Then

(i) If $g(x)$ is nondecreasing and $h(x)$ is nonincreasing then $\mathbb{E}(g(X)h(X)) \leq (\mathbb{E}g(X))(\mathbb{E}h(X))$.

(ii) If $g(x)$ and $h(x)$ are either both nondecreasing or both nonincreasing, then $\mathbb{E}(g(X)h(X)) \geq (\mathbb{E}g(X))(\mathbb{E}h(X))$.

The intuition behind the inequality is easy. In case (i) there is a negative correlation between g and h while in case (ii) there is a positive one. The inequalities merely reflect this fact. The usefulness of the **Covariance Inequality** is that it allows us to bound an expectation without higher-order moments.

5.1 Basic Concepts of Random Samples

Definition: Independent and Identically Distribued (i.i.d.)

The random variables $X_1, \cdots, X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, \cdots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \cdots, X_n$ are called independent and identically distributed random variables with pdf or pmf $f(x)$. This is commonly abbreviated to i.i.d. random variables.

When a sample $X_1, \cdots, X_n$ is drawn, some summary of the values is usually computed. Any well-defined summary may be expressed mathematically as a function $T(x_1, \cdots, x_n)$ whose domain includes the sample space of the random vector

$(X_1, \cdots, X_n)$. The function T may be real-valued or vector-valued; thus the summary is a random variable (resp. random vector), $Y = T(X_1, \cdots, X_n)$.

Since the random sample $X_1, \cdots, X_n$ has a simple probabilistic structure, the distribution of Y is particularly tractable. Because this distribution is usually derived from the distribution of the variables in the random sample, it is called the sampling distribution of Y. This distinguishes the probability distribution of Y from the distribution of the population, i.e. the marginal distribution of each X_i.

Definition: Statistic

Let $X_1, \cdots, X_n$ be a random sample of size n from a population and let $T(x_1, \cdots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \cdots, X_n)$. Then the random variable or random vector $Y := T(X_1, \cdots, X_n)$ is called a statistic. The probability distribution of a statistic Y is called the sampling distribution of Y.

Remark:

The definition of a statistic is very broad, with the only restriction being that a statistic cannot be a function of a parameter. ||

Most of the terminologies we have encountered so far are statistics, e.g. recall the mean μ and the variance σ^2. We now generalize these concepts to the form, that as a function of the random variable (resp. random vector), μ and σ^2 are themselves random variables.

Definition: Sample Mean

The sample mean is the arithmetic average of the value in a random sample. It is usually denoted by $\overline{X} := \dfrac{X_1 + \cdots + X_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

Definition: Sample Variance

The sample variance is the statistic defined by $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$.

Definition: Sample Standard Deviation

The sample standard deviation is the statistic defined by $S := \sqrt{S^2}$.

We now state and prove a very useful numerical result involving the sample mean and the sample variance.

Theorem 5.1: Numerical Identity

Let $x_1, \cdots, x_n$ be any numbers and $\overline{x} = (x_1 + \cdots + x_n)/n$. Then

(i) $\min_a \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2$.

(ii) $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

Proof:

(i):

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x} - a)^2 \qquad \text{(Add and subtract } \overline{x}\text{)}$$

$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + 2\sum_{i=1}^{n} (x_i - \overline{x})(\overline{x} - a) + \sum_{i=1}^{n} (\overline{x} - a)^2$$
$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + (\overline{x} - a)\Big((\sum_{i=1}^{n} x_i) - \overline{x}\Big) + \sum_{i=1}^{n} (\overline{x} - a)^2$$
$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} (\overline{x} - a)^2 \quad \text{(Definition of } \overline{x}\text{)}$$

We now minimize over a on both sides:

$$\min_a \sum_{i=1}^{n} (x_i - a)^2 = \min_a \Big(\sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} (\overline{x} - a)^2 \Big)$$
$$= \min_a \sum_{i=1}^{n} (\overline{x} - a)^2 \quad \text{(Minimized when } a = \overline{x}\text{)}$$

(ii):

With the same approach but this time we set $a = 0$ in $\sum_{i=1}^{n} (x_i - a)^2$, one has

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x})^2 \qquad \text{(Add and subtract } \overline{x}\text{)}$$
$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + 2\sum_{i=1}^{n} (x_i - \overline{x}) \cdot \overline{x} + \sum_{i=1}^{n} \overline{x}^2 \qquad \text{(Expand)}$$
$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} \overline{x}^2. \qquad \text{(Cross term is zero)}$$

Similarly, we expand $\sum_{i=1}^{n} (x_i - \overline{x})^2$ with the cross term vanishes, one has

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} \overline{x}^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2. \qquad (5.1)$$

Lastly for $(n-1)s^2 = (n-1) \cdot \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ by definition, we have,

by (5.1), established the identity $(n-1)s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$.

□

Theorem 5.1 is useful in both computationally and theoretically because it allows us to express s^2 in terms of sums that are easy to handle.

Lemma 5.2: Functional Identity

Let $X_1, \cdots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathbb{E}g(X_1)$ and $\mathbb{E}(\mathrm{Var}X_1)$ both exist. Then,

(i) $\mathbb{E}\sum_{i=1}^{n} g(X_i) = n\mathbb{E}g(X_1)$.

(ii) $\text{Var}\sum_{i=1}^{n} g(X_i) = n\text{Var}g(X_1)$.

Proof:

(i):

Since the X_i's are identically distributed, it follows that $\mathbb{E}g(X_i)$ is the same for all the index i, hence $\mathbb{E}\sum_{i=1}^{n} g(X_i) = n\mathbb{E}g(X_1)$. Note that we have, in fact $\mathbb{E}\sum_{i=1}^{n} g(X_i) = \sum_{i=1}^{n}\mathbb{E}g(X_i) = n\mathbb{E}g(X_1)$ where the middle equality is valid since the expecation is a linear operator, hence the independence for X_i's is not needed for (i) to be valid. Indeed, (i) is valid for any collection of n identically distributed random variables.

(ii):

$$\begin{aligned}\text{Var}\sum_{i=1}^{n} g(X_i) &= \mathbb{E}\Big(\sum_{i=1}^{n} g(X_i) - \mathbb{E}\sum_{i=1}^{n} g(X_i)\Big)^2 \quad \text{(Definition of Var)}\\ &= \mathbb{E}\Big(\sum_{i=1}^{n}\big(g(X_i) - \mathbb{E}g(X_i)\big)\Big)^2 \quad \text{(Property of Expectation)}\end{aligned}$$

In this last expression there are n^2 terms. First, there are n terms $\big((g(X_i) - \mathbb{E}g(X_i)\big)^2$, for $i = 1,\cdots,n$, and for each, we have

$$\begin{aligned}\mathbb{E}\big((g(X_i) - \mathbb{E}g(X_i)\big)^2 &= \text{Var}g(X_i) \qquad \text{(Definition of Var)}\\ &= \text{Var}g(X_1) \qquad \text{(Identically Distributed)}\end{aligned}$$

The reamaining $n(n-1)$ terms are all of the form

$$\big(g(X_i) - \mathbb{E}g(X_i)\big)\big(gX_j) - \mathbb{E}g(X_j)\big),\ i \neq j.$$

For each term, one has

$$\mathbb{E}\Big(\big(g(X_i) - \mathbb{E}g(X_i)\big)\big(g(X_j) - \mathbb{E}g(X_j)\big)\Big) = \text{Cov}\big(g(X_i), g(X_j)\big) = 0$$

due to the definition of covariance and independence. Result follows.

□

Remark:

In obtaining the second result, we have used independence, in fact, independence is a necessary condition for (ii) to hold. ||

Theorem 5.3:

Let $X_1, \cdots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

(i) $\mathbb{E}\overline{X} = \mu$.

(ii) $\text{Var}\overline{X} = \dfrac{\sigma^2}{n}$.

(iii) $\mathbb{E}S^2 = \sigma^2$.

The relationships (i) and (ii) bwtween a statistic and a population parameter are examples of unbiased statistics. The statistic $\overline{X}$ is an unbiased estimator of μ and S^2 is an unbiased estimator for σ^2.

Theorem 5.4:

Let $X_1, \cdots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is $M_{\overline{X}}(t) = \big(M_X(t/n)\big)^n$.

Of course this theorem is useful only if the expression for $M_{\overline{X}}(t)$ is a familiar mgf. Cases when this is true are somewhat limited, when it is not applicable, the following convolution formula is useful.

Theorem 5.5: Convolution Formula

If X and Y are independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$, then the pdf of $Z = X + Y$ is $f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w) dw$.

Recall the exponential family we mentioned in **Section 4**. When sampling is from an exponential family, some sums from a random sample have sampling distributions that are easy to derive. The statistics $T_1, \cdots, T_k$ in the next theorem are important summary statistics.

Theorem 5.6:

Suppose that $X_1, \cdots, X_n$ is a random sample from a pdf or pmf $f(x \mid \theta)$, where $f(x \mid \theta) = h(x)c(\theta)\exp\Big\{ \sum_{i=1}^{k} w_i(\theta) t_i(x) \Big\}$ is a member of an exponential family.

Define the statistics $T_1, \cdots, T_k$ by $T_i(X_1, \cdots, X_n) = \sum_{j=1}^{k} t_i(X_j)$ for $i = 1, \cdots, k$. If the set $\Big\{ (w_1(\theta), \cdots, w_k(\theta)) \Big| \theta \in \Theta \Big\}$ contains an open subset of $\mathbb{R}^k$, then the distribution of $(T_1, \cdots, T_k)$ is an exponential family of the form

$$f_T(u_1, \cdots, u_k \mid \theta) = H(u_1, \cdots, u_k)\big(c(\theta)\big)^2 \exp\Big\{ \sum_{i=1}^{k} w_i(\theta) u_i \Big\}.$$

Note that in the pdf or pmf of $(T_1, \cdots, T_k)$, the functions $c(\theta)$ and $w_i(\theta)$ are the same as in the original family even though the function $H(u_1, \cdots, u_k)$ is different from $h(x)$.

5.2 Sampling from the Normal Distributions

In this subsection we shall deal with the properties of sample quantities drawn from a normal population — still one of the most widely used statistical models. Sampling from a normal population leads to many useful properties of sample statistics and also to many well-known sampling distributions.

We have already seen how to calculate the means and the variances of $\overline{X}$ and S^2 in general. Now, under the additional assumption of normality, we can derive their full distributions, and more. The properties of $\overline{X}$ and S^2 are summarized in the following theorem.

Lemma 5.7: Facts about Chi Squared Random Variables

We use the notation χ_p^2 to denote a chi squared random variable with p degrees of freedom. Then

(i) If Z is a $N(0,1)$ random variable then $Z^2 \sim \chi_1^2$.

(ii) If $X_1, \cdots, X_n$ are independent and $X_i \sim \chi_{p_i}^2 \ \forall i$. Then $X_1 + \cdots + X_n \sim \chi_{p_1+\cdots+p_n}^2$.

This lemma is used to prove the following theorem, which we leave the proof to the readers.

Theorem 5.8:

Let $X_1, \cdots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$. Then

(a) $\overline{X}$ and S^2 are independent random variables.

(b) $\overline{X}$ has a $N(\mu, \frac{\sigma^2}{n})$ distribution.

(c) $\frac{(n-1)S^2}{\sigma^2}$ has a chi squared distribution with $n-1$ degrees of freedom.

Lemma 5.9:

Let $X_j \sim N(\mu_j, \sigma_j^2)$, for $j = 1, \cdots, n$, independent. For constants a_{ij} and b_{rj} where $j = 1, \cdots, n$, $i = 1, \cdots, k$, and $r = 1, \cdots, m$, where $k + m \leq n$, define $U_i := \sum_{j=1}^{n} a_{ij} X_j$ and $V_r := \sum_{j=1}^{n} b_{rj} X_j$. Then one has

(a) The random variables U_i and V_r are independent $\Leftrightarrow \mathrm{Cov}(U_i, V_r) = 0$. Furthermore, $\mathrm{Cov}(U_i, V_r) = \sum_{j=1}^{n} a_{ij} b_{rj} \sigma_j^2$.

(b) The random vectors $(U_1, \cdots, U_k)$ and $(V_1, \cdots, V_m)$ are independent $\Leftrightarrow$ U_i is independent of V_r for all pairs i and r.

This lemma shows that, if we start with independent normal random variables, covariance and independence are equivalent for linear functions of these random variables. Thus, we can check independence for normal variables by merely checking the covariance term, a much simpler calculation. Moreover, (b) allows us to infer overall independence of normal vectors by just checking pairwise independence, a property that does not hold for general random variables.

5.3 Convergence Concepts

We start with one of the weakest types of convergence, the convergence in probability, which is a special case of convergence in measure.

Definition: Converge in Probability

A sequence of random variables, $X_1, X_2, \cdots$, converges in probability to a random variable X if $\forall \varepsilon > 0$, one has $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$, or, equivalently, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$. We denote it as $X_n \xrightarrow{\text{prob}} X$.

The $X_1, X_2, \cdots$ are typically not independent and identically distributed random variables, as in a random sample.

Frequently, statisticians are concerned with situations in which the limiting random variable is a constant and the random variables in the sequence are sample means (of some sort). The most famous result of this type is the following.

Theorem 5.10: Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\text{Var}X_i = \sigma^2 < \infty$. Define $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| < \varepsilon) = 1$; i.e. $\overline{X}_n \xrightarrow{\text{Prob}} \mu$.

The WLLN quite elegantly states that, under general conditions, the sample mean approaches the population mean as $n \to \infty$. In fact, there are more general versions of the WLLN, where we need to assume only that the mean is finite. The one we employ here is applicable in most practical situations.

A natural extension of converge in probability relates to functions of random variables. That is, if the sequence $X_1, X_2, \cdots$ converges in probability to a random variable X or to a constant a, can we make any conclusions about the sequence of random variables $h(X_1), h(X_2), \cdots$ for some reasonably behaved function h? The next theorem shows that we can.

Theorem 5.11:

Suppose that $X_1, X_2, \cdots$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), \cdots$ converges in probability to $h(X)$.

One other interpretation for **Theorem 5.11** is that the continuous mappings preserves the convergence, this is true by its property that the preimage of an open set is still open.

A type of convergence that is stronger than convergence in probability is almost sure convergence. This type of convergence is similar to pointwise convergence of a sequence of functions, except that the convergence need not occur on a set with probability 0.

Definition: Almost Surely Convergence

A sequence $X_1, X_n, \cdots$ of random variables converges almost surely to a random variable X if $\forall \varepsilon > 0\ \mathbb{P}(\lim_{n\to\infty} |X_n - X| < \varepsilon) = 1$.

A very interesting example in showing that almost surely convergence is stronger than convergence in probability is that, if $f_n \to f$, then a necessary and sufficient condition for the identity,

$$\lim_{n\to\infty} \sup_{n\to\infty} f_n(x) = \sup_{n\to\infty} \lim_{n\to\infty} f_n(x), \tag{5.2}$$

being valid for all the choice of x is the almost surely convergence. However, if we relax the condition into convergence in probability, this is not always the case.

One good interpretation is that the almost surely convergence, the set

$$A := \left\{ x \,\middle|\, \forall \varepsilon > 0 \exists \delta > 0 \text{ such that } |f_n(x_1) - f(x_2)| < \delta\ \forall x_1, x_2 \text{ where } |x_1 - x_2| < \varepsilon \right\}$$

has a probability measure of zero. Therefore it guarantees that during the process of its convergence, the ordering of the original space does not vary too much, hence the

interchange is valid. Moreover, this is also valid when we change the sup into inf , max, and min. This order preserving property is, perhaps one of the reasons why the almost surely convergence is stronger than the convergence in probability. Of course, we are assuming that the original space is well-ordering.

Remark:

We shall denote that f_n converges to f almost surely by the notation $f_n \xrightarrow{\text{a.s.}} f$.

Note that

almost surely convergence $\Rightarrow$ Convergence in Probability
almost surely convergence $\nLeftarrow$ Convergence in Probability. ||

There are examples that some random variables converge in probability but fails to be almost surely convergent, one may consult [1] for details. Note that even though the converse direction fails to be true, when a sequence converges in probability, it is still possible to find a subsequence that is almost surely convergent. This is the idea of the strong law of large numbers:

Theorem 5.12: Strong Law of Large Numbers (SLLN)

Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\text{Var}X_i = \sigma^2 < \infty$, define $\overline{X}_n := \dfrac{1}{n}\sum_{i=1}^{n} X_i$. Then $\forall \varepsilon > 0$, one has that $\mathbb{P}(\lim_{n\to\infty} |\overline{X}_n - \mu| < \varepsilon) = 1$, i.e. $\overline{X}_n$ converges almost surely to μ.

Proof:

To prove SLLN is to prove that the divergence part has a probability measure 0. For the sequence to diverge, there must exist a $\delta > 0$ such that $\forall n \in \mathbb{N}$, there exist $k > n$ such that $|\overline{X}_k - \mu| > \delta$. We shall denote this set as

$$A_\delta := \bigcap_{n\geq 1}\bigcup_{k\geq n}\left\{|\overline{X}_k - \mu| > \delta\right\},$$

which has an upper bound (w.r.t. the probability measure) given by

$$\begin{aligned}
\mathbb{P}(A_\delta) &\leq \mathbb{P}\Big(\bigcup_{k\geq n}\left\{|\overline{X}_k - \mu| > \delta\right\} && \text{(Removing Intersections)}\\
&\leq \sum_{k\geq n}\mathbb{P}(\{|\overline{X}_k - \mu| > \delta\}) && \text{(\textbf{Theorem 1.5} (ii), Boole's)}\\
&\leq 2\cdot\sum_{k\geq n} c^k \text{ for } 0 < c < 1 && \text{(Left as exercise)}\\
&= \lim_{n\to\infty} 2\cdot\frac{c^n}{1-c} = 0. && \text{(Since } 0 < c < 1\text{)}
\end{aligned}$$

□

Not only the convergence for the probability measures derives useful information about the sample but also its distributions, this concept is also called the weak convergence. Unlike the other three, whether a sequence of random variables (elements) conver-ges in distribution or not depends only on their distributions.

Definition: Convergence in Distribution

A sequence of random variables $X_1, X_2, \cdots$ converges in distribution to a random variable X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)\ \forall x$ where $F_X(x)$ is continuous.

Note that although we talk of a sequence of random variables converging in distribution, it is really the cdfs, that converge, not the random variables themselves, thus it makes a major difference from the almost surely convergence and the convergence in probability. However, it is implied by the other types of convergence. We now state a result without proof and construct a relation diagram among these three types of convergence.

Theorem 5.13:

If the sequence $X_1, X_2, \cdots$ converges in probability to a random variable X, then the sequence also converges in distribution to X.

This theorem tells why the convergence in distribution is also called the "weak" convergence.

Remark:

$$\begin{array}{ccc} & \Leftarrow & \\ \text{Convergence in Probability} & \not\Rightarrow & \text{Almost Surely Convergence} \\ \Downarrow & & \Downarrow \\ \multicolumn{3}{c}{\text{Convergence in Distribution}} \end{array} \quad \|$$

In some special case, **Theorem 5.13** has a converse that turns out to be useful. We now state this result without proof.

Theorem 5.14:

The sequence of random variables $X_1, X_2, \cdots$ converges in probability to a constant $\mu \Leftrightarrow$ the sequence also converges in distribution to μ.

The sample mean is one of the statistics whose large-sample behavior is quite important. In particular, we want to investigate its limiting distribution. This is summarized in one of the most startling theorems in statistics, the Central Limit Theorem (CLT).

Theorem 5.15: Central Limit Theorem (CLT)

Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables whose mgfs exist in a neighbourhood of 0. Let $\mathbb{E}X_i = \mu$ and $\text{Var}X_i = \sigma^2 > 0$ be both finite. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and let $G_n(x)$ denote the cdf of $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Then, $\forall -\infty < x < \infty$, one has that $\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$, i.e. $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ has a limiting standard normal distribution.

CLT is valid in much more general way than it is stated. The only assumption on the parent distribution is that it has finite variance.

An approximation tool that can be used in conjunction with the CLT is known as the Slutsky's Theorem.

Theorem 5.16: Slutsky's Theorem

If $X_n \to X$ in distribution and $Y_n \xrightarrow{\text{Prob}} a$ where a is a constant. Then

(i) $Y_n X_n \to aX$ in distribution.

(ii) $X_n + Y_n \to X + a$ in distribution.

It has to be stressed out that the difference between the convergence of a sequence of random variables and the convergence of its corrsponding probability mappings are different in many senses. One of the most important, or, intuitive one, is that the convergence of the random variables themselves means that the distributions converges as well. More precisely, it is that the accumulation finally collides into one, hence it is a matter of the CDF.

The convergence statements, as well as the inequalities, play important roles in modern probability theory and statistics. In the next chapters, we shall see their practical use under, one of the thing we concern the most, the data reduction.

6.1 The Sufficiency Principle

Recall that in studying linear algebra, it is sometimes hard to deal with rather big vector spaces, even its vector subspaces; to that end, we find it useful to work only through a small collection of elements that contain all the information of the vector space, hence we introduced the basis, as well as subbasis.

Same problems may arise when we are dealing with a big set of data. We wish, therefore, to use a small collection that contains all the information of the original data. However, not every data reduction methods could discard no information, so we wish to have one that preserve as much as possible. We shall introduce three data reduction methods in this subsection. The sufficiency principle promotes a method that preserve the information while achieving summrization of the data. The likelihood principle describes a a function of the parameter, determined by the observed sample, that contains all the information about θ that is available from the sample.

Definition: Sufficient statistic

A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Theorem 6.1: Criterion for Sufficient Statistic

If $p(x\,|\,\theta)$ is the joint pdf or pmf of X and $q(t\,|\,\theta)$ is the pdf or pmf of $T(X)$, then $T(X)$ is a sufficient statistic for θ if $\forall x \in X$, $\dfrac{p(x\,|\,\theta)}{q(T(x)\,|\,\theta)}$ is constant as a function of θ.

Theorem 6.2: Factorization Theorem

Let $f(x\,|\,\theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for $\theta \Leftrightarrow$ there exist functions $g(t\,|\,\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ, $f(x\,|\,\theta) = g(T(x)\,|\,\theta)h(x)$.

It is easy to find a sufficient statistic for an exponential family of distributions using the factorization theorem.

Theorem 6.3:

Let $X_1, \cdots, X_n$ be i.i.d. observations from a pdf or pmf $f(x\,|\,\theta)$ that belongs to an experimental family given by $f(x\,|\,\theta) = h(x)c(\theta)\exp\Big\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\Big\}$ where $\theta = (\theta_1, \cdots, \theta_d)$, for $d \le k$. Then $T(X) = \Big(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\Big)$ is a

sufficient statistic for θ.

Remark:

$T(X) = X$ is always a sufficient statistic. Moreover, every one-to-one function of a sufficient statistic is a sufficient statistic. ||

Because of the numerous sufficient statistics in a problem, we might ask whether one sufficient statistic is any better than another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still remaining all the information about θ might be considered preferable. The definition of such a statistic is the minimal sufficient statistic.

Definition: Minimal Sufficient Statistic

A sufficient statistic $T(X)$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(X)$, $T(x)$ is a function of $T'(X)$.

That is to say, $T'(x) = T'(y) \Rightarrow T(x) = T(y)$, or, equivalently, if $\{B_{t'} \mid t' \in \mathcal{T}'\}$ are the partition sets of $T'(X)$ and $\{A_t \mid t \in \mathcal{T}\}$ are the partition sets for $T(x)$, then every $B_{t'}$ is a subset of A_t. Thus, the partition associated with a minimal sufficient statistic, is the coarsest possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

Theorem 6.4: Criterion for Minimal Sufficient Statistic

Let $f(x \mid \theta)$ be the pmf or pdf of a sample X. Suppose that there exist a function $T(x)$ such that for every two sample points x and y, the ratio $\dfrac{f(x \mid \theta)}{f(y \mid \theta)}$ is constant as a function of $\theta \Leftrightarrow T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.

However, a minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.

Definition: Ancillary Statistic

A statistic $S(X)$ whose distribution does not depend on the parameter θ is called an ancillary statistic.

Alone, an ancillary statistic contains no information about θ. An ancillary statistic is an observation on a random variable whose distribution is fixed and knwon, unrelated to θ. Paradoxically, an ancillary statistic, when used in conjunction with other statistics, sometimes does contain valuable information for inferences about θ.

Ancillary statistic is not necessary to be independent from the minimal sufficient statistic. Indeed, an ancillary statistic can sometimes give important information for inference about θ. For many important situations, however, a minimal sufficient statistic is independent of any ancillary statistic.

Definition: Complete Statistic

Let $f(t \mid \theta)$ be a family of pdfs or pmfs for a statistic $T(X)$. The family of distributions is called complete if $\mathbb{E}_\theta g(T) = 0 \forall \theta$ then $\mathbb{P}_\theta(g(T) = 0) = 1$ $\forall \theta$. Equivalently, $T(X)$ is called a complete statistic.

We now use completeness to state a condition under which a minimal sufficient statistic is independent of every ancillary statistic.

Theorem 6.5: Basu's Theorem

If $T(X)$ is a complete and minimal sufficient statistic, then $T(X)$ is independent of every ancillary statistic.

Basu's theorem is useful since it allows us to determine the independence of two statistics without ever finding their joint distribution. However, to use Basu's theorem, one needs to show that a statistic is complete, which could be a tedius work. Fortunately, most problems we are concerned with satisfy the following theorem.

Theorem 6.6: Complete Statistic in the Exponential Family

Let $X_1, \cdots, X_n$ be i.i.d. observations from an exponential family with pdf or pmf of the form $f(x \mid \theta) = h(x)c(\theta)\exp\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\}$, where $\theta = (\theta_1, \cdots, \theta_k)$.

Then the statistic $T(X) := \left(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\right)$ is complete if $\left\{\left(w_1(\theta), \cdots, w_k(\theta)\right) \middle| \theta \in \Theta\right\}$ contains an open set in $\mathbb{R}^k$.

The proof of this theorem depends on the uniqueness of a Laplace transform. It should be noted that the minimality of the sufficient statistic was not used in the proof of Basu's theorem. Indeed, the theorem is true with this word omitted, since a fundamental property of a complete statistic is that it is minimal. However, the condition that it contains an open set is necessarily needed.

Theorem 6.7:

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

So even though the word "minimal" is redundant in the statement of Basu's theorem, it was stated in this way as a reminder that the statistic $T(X)$ in the theorem is a minimal sufficient statistic.

6.2 The Likelihood Principle

In this subsection we study a specific, important statistic called the likelihood function that also can be used to summarize data. There are many ways to use the likelihood function but the main consideration in this subsection is an argument which indicates that if certain other principles are accepted, the likelihood function **must** be used as a data reduction device.

Definition: Likelihood Function

Let $f(x \mid \theta)$ denote the joint pdf or pmf of the sample $X = (X_1, \cdots, X_n)$. Then, given that $X = x$ is observed, the function of θ defined by $L(\theta \mid x) = f(x \mid \theta)$ is called the likelihood function.

In this form, it is intuitively that the likelihood function has relationships with the original distribution function $f(x \mid \theta)$. It turns out that our instinct is true. The comparison between likelihood function implies the comparison between the corresponding probability measures.

Suppose that X is a continuous real-valued random variable with continuous pdf in x. Then, $\forall \varepsilon > 0$, $\mathbb{P}_\theta(x - \varepsilon < X < x + \varepsilon)$ is approximately $2\varepsilon f(x \mid \theta) = 2\varepsilon L(\theta, x)$ by

definition. Therefore, $\frac{\mathbb{P}_{\theta_1}(x-\varepsilon<X<x+\varepsilon)}{\mathbb{P}_{\theta_2}(x-\varepsilon<X<x+\varepsilon)} \approx \frac{L(\theta_1 \mid x)}{L(\theta_2 \mid x)}$. Let us summarize this observation into the following remark.

Remark:

Likelihood functions behave very much as the pmf or pdf. The only distinction is that pdf and pmf $f(x \mid \theta)$ consider θ as fixed and x as the variable while the likelihood functions behave the other way around. ||

Fact 6.8: Likelihood Principle

If x and y are two sample points such that $L(\theta \mid x)$ is proportional to $L(\theta \mid y)$, i.e. there exists a constant C such that $L(\theta \mid x) = C(x, y)L(\theta, y) \forall \theta \in \Theta$. Then the conclusion drawn from x and y are identical.

The likelihood principle specifies how the likelihood function should be used as a data reduction device. When $C(x, y) = 1$, the likelihood principle tells us that two sample points x and y result in the same likelihood function then they convey the same information about θ. Likelihood principle may go even further, it states that even if two sample points have only proportional likelihoods, then they contain equivalent information about θ.

Definition: Evidence

Define an experiment E to be a triple $(X, \Theta, \{f(x \mid \theta)\})$, where X is a random vector with pmf $f(x \mid \theta)$ for some $\theta \in \Theta$. An experimenter, knowing what experiment E was performed and having observed a particular sample $X = x$, will make some inference or draw some conclusion about θ. This conclusion we denote as $E_V(E, X)$, which stands for the evidence about θ arising from E and x.

Fact 6.9: Formal Sufficiency Principle

Consider experiment $E = (X, \Theta, \{f(x \mid \theta\})$ and suppose $T(X)$ is a sufficient statistic for θ. If x and y are sample points such that $T(x) = T(y)$ then $E_V(E, x) = E_V(E, y)$.

The formal sufficiency principle says that the two experiments are identically the same once their summarization over the data coincide.

Fact 6.9: Conditionality Principle

Suppose that $E_1 := (X_1, \Theta_1, \{f_1(x \mid \theta)\})$ and $E_2 := (X_2, \Theta_2, \{f_2(x_2 \mid \theta)\})$ are two experiments, where only the unknown parameter θ need be common between the two experiments. Consider the mixed experiment in which the random variable J is observed where $\mathbb{P}(J = 1) = \mathbb{P}(J = 2) = \frac{1}{2}$ (independent of $\theta, x_1,$ or x_2), and then the experiment E_j is performed. Formally, the experiment performed is $E^* = (X^*, \Theta, \{f^*(x^* \mid \theta)\})$, where $X^* = (j, X_j)$ and $f^*(x^* \mid \theta) = f^*\big((j, x_j) \,\big|\, \theta\big) = \frac{1}{2} f_j(x_j \mid \theta)$. Then, $E_V(E^*, (j, x_j)) = E_V(E_j, x_j)$.

The conditional principle simply states that if one or two experiments is randomly chosen and the chosen experiment is done, yielding data x, the information about θ depends only on the experiment performed. That is, it is the same information as

would have been obtained if it were decided (non-randomly) to do that experiment from the beginning, and data x had been observed. The fact that this experiment was performed, rather than some other, has not increased, decreased, or changed any knowledge of θ.

Fact 6.10: Formal Likelihood Principle

Suppose that we have two experiments, $E_1 = (X_1, \Theta_1, \{f_1(x_1 \mid \theta)\})$ and $E_2 = (X_2, \Theta_2, \{f_2(x_2 \mid \theta)\})$, where the unknown parameter θ is the same in both experiments. Suppose that x_1^* and x_2^* are sample points from E_1 and E_2, respectively such that $L(\theta \mid x_2^*) = C(x_1^*, x_2^*)L(\theta \mid x_1^*)$ for all θ and for some constant C that may depend on x_1^* and x_2^* but not on θ. Then $E_V(E_1, x_1^*) = E_V(E_2, x_2^*)$.

The formal likelihood principle is different from the likelihood principle we saw before because the formal likelihood principle concerns two experiments while the likelihood principle concerns one.

Fact 6.11: Likelihood Principle Corollary

If $E = \{X, \Theta, \{f(x \mid \theta)\})$ is an experiment, then $E_V(E, x)$ should depend on E and x only through $L(\theta, x)$.

We now state and investigate the Birnbaum's theorem whose result turns out to be somewhat surprising.

Theorem 6.12: Birnbaum's Theorem

Formal Sufficiency Principle + Conditional Principle $\Leftrightarrow$ Formal Likelihood Principle.

Many common statistical procedure violates the formal likelihood principple, hence by Birnbaum's Theorem, we are then violating either the sufficiency principle or the conditional principle. It must be realized that before considering the sufficiency principle, or the likelihood principle, we must be comfortable with the model.

7.1 Methods of Finding Estimators

This section is divided into two parts. The first part deals with methods for finding estimators, and the second part deals with evaluating these (and other) estimators. In general these two activities are intertwined. Often the methods of evaluating estimators will suggest new ones.

Definition: Point Estimator

A point estimator is any function $W(X_1, \cdots, X_n)$ of a sample; i.e. any statistic is a point estimator.

Note that an estimator is a function of the sample, while an estimate is the realized value of an estimator. It is useful to have some techniques that will at least give us some reasonable candidates for consideration.

There are four different ways of finding estimators we shall mention in this subsection. They are: the methods of moments, maximum likelihood estimators (MLE), the Bayes Estimators, and the EM algorithm. We follow this order in introduction.

The method of moments is, perhaps, the oldest method of finding point estimators, it has the virtue of being quite simple to use and almost always yields some sort of

estimate. In many cases, unfortunately, this method yields estimators that may be improved upon. However, it is a good place to start when old methods prove intractable.

Algorithm 7.1: Methods of Moments

Let $X_1, \cdots, X_n$ be a sample from population with pdf or pmf $f(x \mid \theta_1, \cdots, \theta_k)$. Methods of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneously equations. More precisely, define

$$m_1 := \frac{1}{n}\sum_{i=1}^{n} X_i', \mu_1' = \mathbb{E}X,$$
$$m_2 := \frac{1}{n}\sum_{i=1}^{n} X_i^2, \mu_2' := \mathbb{E}X^2,$$
$$\cdots\cdots$$
$$m_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k, \mu_k := \mathbb{E}X^k.$$

The population moments μ_j' will typically be function of $\theta_1, \cdots, \theta_k$, namely $\mu_j'(\theta_1, \cdots, \theta_k)$. The method of moments estimators $(\tilde{\theta}_1, \cdots, \tilde{\theta}_k)$ of $(\theta_1, \cdots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \cdots, \theta_k)$ in terms of $(m_1, \cdots, m_k)$.

$$m_1 = \mu_1'(\theta_1, \cdots, \theta_k),$$
$$m_2 = \mu_2'(\theta_1, \cdots, \theta_k),$$
$$\cdots\cdots$$
$$m_k = \mu_k'(\theta_1, \cdots, \theta_k).$$

The method of moments can be very useful in obtaining approximations to the distribution of statistics. This technique, is sometimes called the moment matching, gives us an approximation that is based on matching moments of distributions. In theory, the moments of distribution of any statistics could be matched, however, in practical terms, it is best to have distributions that are similar.

The method of maximum likelihood, on the other hand, is by far the most popular technique for deriving estimators. Recall that if $X_1, \cdots, X_n$ are an i.i.d. sample from a pupulation with pdf or pmf $f(x \mid \theta_1, \cdots, \theta_k)$, the likelihood function is defined by

$$L(\theta \mid x) := L(\theta_1, \cdots, \theta_k \mid x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \cdots, \theta_k).$$

Definition: Maximum Likelihood Estimator (MLE)

For each sample point x, let $\hat{\theta}(x)$ be a paramater value at which $L(\theta \mid x)$ attains its maximum as a function of θ, with x fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample X is $\hat{\theta}(X)$. In short, it is the value of θ that maximizes the likelihood function.

Notice that, by this construction, the range of the MLE coincides with the range of the parameter. We also use the abbreviation MLE to stand for Maximum Likelihood Estimate when we are talking about the realized value of the estimator. Intuitively, the

MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. In general, the MLE is a good point estimator, processing some of the optimality properties.

It could be a little confusing that how can maximizing over the likelihood function gives us the best approximation, since intuitively, $L(\theta \mid x) = f(x \mid \theta)$ is how the likelihood function is defined, how can maximizing the "distribution" raise the probability? The key distinction here is that the likelihood function is a function of the parameters, treating the data as fixed, while the probability distribution is a function of the data, given specific parameter values. The likelihood function doesn't represent a probability distribution over data points; it measures the fit between the data and the parameter values. So, when we talk about maximizing the likelihood, we mean finding parameter values that make the observed data most probable under the given statistical model. It's not about making the data itself more probable but rather about finding the parameter values that make the observed data most consistent with the assumed model. In other words, maximizing the likelihood is about choosing the parameter values that align with the data we've observed. It is a way to find the "best-fitting" parameters that explain the data in a probabilistic sense based on the model we have specified.

Now the problem turns out to be an "optimization" one. In finding the maximum, one common practice is to have the first derivative being zero, however, this is a necessary condition but not a sufficient one. Moreover, the zeros of the first derivative locate only extreme points in the interior of the domain of a function. Furthermore, if the extrema occurs at the boundary then the first derivative may not be 0, thus the boundary points must be checked separately for extrema.

We also wish the translation invariance to be one of the properties of the MLE. It is unfortunately that sometimes a slightly change of the sample will produce a vastly change between MLEs, which makes its use suspects.

Let us start with the first problem: finding the global maximum. This is always hard since guranteeing the globality is very tedius. Instead of differentiation, one general technique is taking the global upper bound. Followed from some properties of the convexity, it turns out that the log MLE, $\log L(\theta \mid x)$, which is convex, is easier to work with, and since the log function is strictly increasing on $(0,\infty)$, the extrema of $L(\theta \mid x)$ and $\log L(\theta \mid x)$ must coincide.

Remark:

If $L(\theta \mid x)$ cannot be maximized analytically, it may be possible to use a computer and maximize $L(\theta \mid x)$ numericall. ||

Now for the second problem, a very useful property of MLEs is its invariance property. Informally, the invariance property of MLEs says that if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ for some function τ. If the mapping $\theta \mapsto \tau(\theta)$ is one-to-one, then we are done. In this case optimizing over θ has no difference in optimizing over $\tau(\theta)$. However, not all functions are one-to-one. Thus we need a more general theorem and in fact a more general definition of the likelihood function for $\tau(\theta)$.

Definition: Induced Likelihood Function

Define for $\tau(\theta)$ the induced likelihood function L^* given by

$$L^*(\eta \mid x) := \sup_{\{\theta \mid \tau(\theta)=\eta\}} L(\theta \mid x).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta \mid x)$ will be called the MLE of $\eta = \tau(\theta)$, and it can be seen by the definition that the maximum of L^* and L coincide. Hence it follows, no matter bijective or not the τ is, the translation invariance is always valid.

Theorem 7.2: Invariant Property of MLEs

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Proof:

Let $\hat{\eta}$ be the value that maximizes the induced likelihood function $L^*(\eta \mid x)$.

WTS I: $L^*(\hat{\eta} \mid x) = L^*(\tau(\hat{\theta}) \mid x)$.

By definition, the maximum of L and L^* coincide, therefore, it follows that

$$L^*(\hat{\eta} \mid x) = \sup_{\eta} \sup_{\{\theta \mid \tau(\theta)=\eta\}} L(\theta \mid x) = \sup_{\theta} L(\theta \mid x) = L(\hat{\theta} \mid x),$$

where the last equality is by the definition of $\hat{\theta}$. On the other hand, we have

$$\begin{aligned} L(\hat{\theta} \mid x) &= \sup_{\{\theta \mid \tau(\theta)=\tau(\hat{\theta})\}} L(\theta \mid x) \quad (\hat{\theta} \text{ is the MLE}) \\ &= L^*(\tau(\hat{\theta}) \mid x). \quad \text{(Definition of } L^*) \end{aligned}$$

Hence, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ and the invariance follows.

□

Remark:

The invariance property for MLE is still valid for the multivariate case. ||

Note that in most instances, MLE cannot be solved for explicitly and must be found by numerical methods. When facing such problems, it is often wise to spend a little extra time investigating the stability of the solution.

Now we move to the discussion of the Bayes Estimators. The Bayesian approach to statistics is fundamentally different from the classical ones. In the calssical approach to the parameter, θ, is thought to be an unknown, but fixed, quanitity. A random sample $X_1, \cdots, X_n$ is drawn from a population indexed by θ and, based on the observed values in the sample, knowledge about the value of θ is obtained. In the Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution called the prior distribution, which is based on the experimenters' belief. A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the posterior distribution.

Note that the posterior distribution is a conditional distribution, conditional upon observing the sample. The posterior distribution is now used to make statements about θ, which is still considered as a random quantity. For instance, the mean of the posterior distribution can be used as a point estimate of θ.

In general, for any sampling distribution, there is a natural family of prior distributions, called the conjugate family.

Definition: Conjugate Family

Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x \mid \theta)$ indexed by θ. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the

class Π $\forall f \in \mathcal{F}$, all priors in Π, and all $x \in \Omega_X$.

Loosely speaking, one may interprete the conjugate family as that it is closed under taking Bayesian estimators.

Example 7.1: Normal Bayes Estimators

Let $X \sim N(\theta, \sigma^2)$ and suppose that the prior distribution on θ is $N(\mu, \tau^2)$. Here we assume that σ^2, μ, and τ^2 are known. The posterior distribution of θ is also normal, with mean and variance given by

$$\mathbb{E}(\theta \mid x) = \frac{\tau^2}{\tau^2 + \sigma^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu, \ \mathrm{Var}(\theta \mid x) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}.$$

Notice that Normal families are their own conjugate families.

Again use the posterior mean, we have the Bayes estimator of θ is $\mathbb{E}(\theta \mid X)$. The Bayes estimator is, again, a linear combination of the prior and the sample means.

Remark:

Notice that as τ^2, the prior variance, is allowed to tend to infinity, the Bayes estimator tends toward the sample mean. We can interpret this as saying that, as the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information. On the other hand, if the prior information is good, so that $\sigma^2 > \tau^2$, then more weight is given to the prior mean. ||

A last method that we will look at for finding estimators is inherently different in its approach and specifically designed to find MLEs. Rather than detailing a procedure for solving for the MLE, we specify an algorithm that is guaranteed to converge to the MLE. This algorithm is called the EM (Expectation-Maximization) algorithm. It is based on the idea of replacing one difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original problem. It is particularly suited to "missing data" problems, as the very fact that there are missing data can sometimes make calculations cumbersome. However, we will see that filling in the "missing data" will often make the calculation go more smoothly.

7.2 Methods of Evaluating Estimators

The methods discussed in the previous subsection have outlined reasonable techniques for finding point estimators of parameters. A difficulty that arises, however, is that since we can usually apply more than one of these methods in a particular situation, we are often faced with the task of choosing between estimators. Of course, it is possible that different methods of finding estimators will yield the same answer, which makes evaluation a bit easier, but, in many cases, different methods will lead to different estimators.

The general topic of evaluating statistical procedures is part of the branch of statistics known as decision theory. However, no procedure should be considered until some clues about its performance have been gathered. In this subsection we introduce some basic criteria for evaluating estimators, and examine several estimators against these criteria.

We first investigate finite-sample measures of the quality of an estimator, beginning with its mean squared error.

Definition: Mean Squared Error (MSE)

The mean squared error of an estimator W of a parameter θ is the function of θ defined by $\mathbb{E}_\theta(W-\theta)^2$.

The MSE measures the average squared difference between the estimator W and the parameter θ, a somewhat reasonable measure of performance for a point estimator. For example, any increasing function of the absolute distance $|W-\theta|$ would serve to measure the goodness of an estimator (Mean Absolute Error, for example, $\mathbb{E}_\theta(|W-\theta|)$, is a reasonable alternative), but MSE has at least two advantages over the other distance measures:

(i) MSE is quite tractable analytically.

(ii) MSE has the interpretation

$$\mathbb{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W + (\mathbb{E}_\theta W-\theta)^2 := \mathrm{Var}_\theta W + (\mathrm{Bias}_\theta W)^2.$$

Therefore we derive the concepts "biased" and "unbiased" in a very natural way, they are defined as follows.

Definition: Bias

The bias of a point estimator W of a parameter θ is the difference between the expected value of W and θ. That is, $\mathrm{Bias}_\theta W := \mathbb{E}_\theta W - \theta$.

Definition: Unbiased

An estimator whose bias is identically (in θ) equal to 0 is called unbiased and satisfies $\mathbb{E}_\theta W = \theta \; \forall \theta \in \Theta$.

MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy). An estimator that has good MSE properties has small combined variance and bias.

To find an estimator with good MSE properties, we need to find estimators that control both variance and bias. Clearly, unbiased estimators do a good job in controlling bias. For an unbiased estimator, one has $\mathbb{E}_\theta(W-\theta)^2 = \mathrm{Var}_\theta W$. That is, if the estimator is unbiased, then its MSE equal to its variance.

Remark:

Although many unbiased estimators are also reasonable, controlling bias does not necessarily control the MSE. In particular, it is sometimes the case that a trade-off occurs between the variance and the bias in such a way that a small increase in bias could result in a larger decrease in variance, resulting in an improvement in MSE. ||

Disadvantage: MSE

It can be argued that the MSE, while being reasonable for location parameter, is not reasonable to scale parameters since MSE penalizes equally for overestimation and underestimation, which is fine in the location case; in the scale case however, 0 is a natural lower bound, so the estimation is not symmetric.

In many cases, the MSEs of two estimators will cross each other, showing that each estimator is better with respect to the other in only a small portion of the parameter space. However, even this partial information can sometimes provide guidelines for

choosing between given estimators. In some worse cases however, only more information is gathered but no absolute answer is obtained.

One of the reason is that the class of all estimators is too large as a class. So instead of stucking in MSE, we have another alternative that is to reduce the size of this class. A popular way of restricting the class of estimators is to consier only unbiased estimators.

If W_1 and W_2 are both unbiased estimators of a parameter θ, i.e. $\mathbb{E}_\theta W_1 = \mathbb{E}_\theta W_2 = \theta$ then their MSE are equal to their variances, so we should choose the estimator with the smaller variance. If we can find an unbiased estimator with uniformly smallest variance — a best unbiased estimator — then we are done.

Suppose that there is an estimator W^* of θ with $\mathbb{E}_\theta W^* = \tau(\theta) \neq \theta$ and we are interested in investigating the worth of W^*. Consider the class of estimators given by

$$C_\tau := \{W \mid \mathbb{E}_\theta W = \tau(\theta)\}.$$

For all the choice of $W_1, W_2 \in C_\tau$, $\text{Bias}_\theta(W_1) = \text{Bias}_\theta(W_2)$ so one has

$$\mathbb{E}_\theta(W_1 - \theta)^2 - \mathbb{E}_\theta(W_2 - \theta)^2 = \text{Var}_\theta(W_1) - \text{Var}_\theta(W_2)$$

and MSE comparisons, within the class C_τ, can be based on variance alone. Thus, although we speak in terms of unbiased estimators, we really are comparing estimators with the same expected value $\tau(\theta)$.

Definition: Best Unbiased Estimator (BUE)

An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $\mathbb{E}_\theta W^* = \tau(\theta) \forall \theta$, and for any other estimator W with $\mathbb{E}_\theta W = \tau(\theta)$.

Definition: Uniform Minimum Variance Unbiased Estimators (UMVUE)

A BUE W^* is said to be a uniform minimum variance unbiased estimator if for any other estimator W with $\mathbb{E}_\theta W = \tau(\theta)$, one always has $\text{Var}_\theta W^* \leq \text{Var}_\theta W \ \forall \theta$.

Suppose that, for estimating a parameter $\tau(\theta)$ of a distribution $f(x \mid \theta)$, we can specify the lower bound, say $B(\theta)$, on the variance of any unbiased estimator of $\tau(\theta)$. If we can find an unbiased estimator W^* such that $\text{Var}_\theta W^* = B(\theta)$, then we have found the BUE. This is the approach taken with the use of the Cramér-Rao lower bound.

Theorem 7.3: Cramér-Rao Inequality

Let $X_1, \cdots, X_n$ be a sample with pdf $f(x \mid \theta)$, and let $W(X) = W(X_1, \cdots, X_n)$ be any estimator satisfying

(i) $\dfrac{d}{d\theta}\mathbb{E}_\theta W(X) = \displaystyle\int_{\Omega_X} \frac{\partial}{\partial\theta} W(x) f(x \mid \theta) dx$

(ii) $\text{Var}_\theta W(X) < \infty$.

Then $\text{Var}_\theta W(X) \geq \dfrac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(X)\right)^2}{\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right)}$.

If we add the assumption of independent samples, the calculatin of the lower bound could be simplified. The expectation in the denominator becomes a univariate calculation, as the following corollary implies.

Corollary 7.3.1: Cramér-Rao Inequality, i.i.d. case

Let $X_1, \cdots, X_n$ be an i.i.d. sample with pdf $f(x \mid \theta)$ and let

$W(X) := W(X_1, \cdots, X_n)$ be any estimator such that

(i) $\dfrac{d}{d\theta}\mathbb{E}_\theta W(X) = \displaystyle\int_{\Omega_X} \frac{\partial}{\partial\theta} W(x) f(x\,|\,\theta)\,dx$

(ii) $\mathrm{Var}_\theta W(X) < \infty$.

Then $\mathrm{Var}_\theta W(X) \geq \dfrac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(X)\right)^2}{n\mathbb{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X\,|\,\theta)\right)^2\right)}$.

Note that the Cramér-Rao lower bound does not only work for the continuous random variables but also the discrete ones. The quantity $\mathbb{E}_\theta\left(\left(\dfrac{\partial}{\partial\theta}\log f(X\,|\,\theta)\right)^2\right)$ is called the information number, or Fisher information of the sample. This terminology reflects the fact that the information number gives a bound on the variance of the BUE of θ. As the information number increases, the bound on the variance of BUE gets smaller.

For any differentiable function $\tau(\theta)$, we now have a lower bound on the variance of any estimator W such that $\mathbb{E}_\theta W = \tau(\theta)$. The bound depends only on $\tau(\theta)$ and $f(x\,|\,\theta)$ and is a uniform lower bound for the variance. Any candidate estimator satisfying $\mathbb{E}_\theta W = \tau(\theta)$ and attaining this lower bound is a BUE of $\tau(\theta)$.

Remark:

> Even if the Cramér-Rao is applicable, there is no guarantee that the bound is sharp. That is to say, the value of the Cramér-Rao lower bound may be strictly smaller than the variance of any unbiased estimator.

In fact, the most we can say by applying Cramér-Rao is that there exists a parameter $\tau(\theta)$ with an unbiased estimator that achieves the Cramér-Rao lower bound; however, in other typical situations, for other parameters, the bound may not be attainable. Hence we need results dealing with its attainment.

Corollary 7.3.2: Attainment of Cramér-Rao Lower Bound

> Let $X_1, \cdots, X_n$ be i.i.d. $f(x\,|\,\theta)$ where $f(x\,|\,\theta)$ satisfies the conditions of Cramér-Rao Theorem. Let $L(\theta\,|\,x) := \displaystyle\prod_{i=1}^{n} f(x_i\,|\,\theta)$ denote the likelihood function. If $W(X) = W(X_1, \cdots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(X)$ attains the Cramér-Rao lower bound if and only if
>
> $$a(\theta)\big(W(x) - \tau(\theta)\big) = \frac{\partial}{\partial\theta}\log L(\theta\,|\,x)$$
>
> for some function $a(\theta)$.

The attainment of the Cramér-Rao lower bound still leaves some questions unanswered. Firstly, what if the $f(x\,|\,\theta)$ does not satisfy the assumptions of the **Cramér-Rao Theorem**? Secondly, what if the bound is still unattainable for legal estimators?

One way of answering these questions is to search for methods that are more widely applicable and yield sharper (i.e. greater) lower bounds. Much research has been done on this topic, with perhaps the most famous one is Chapman and Robbins (1951). We leave this to interested readers and we now introduce the study of BUE from another view, using the concept of sufficiency.

In the previous discussion, the concept of sufficiency was not used in our search f-or unbiased estimates. We will now see the consideration of sufficiency is a powerful tool indeed. The main result of this method relates the sufficient statistic to unbiased estimate. Recall that $\mathbb{E}X = \mathbb{E}\big(\mathbb{E}(X \mid Y)\big)$ and $\mathrm{Var}X = \mathrm{Var}\big(\mathbb{E}(X \mid Y)\big) + \mathbb{E}\big(\mathrm{Var}(X \mid Y)\big)$.

Theorem 7.4: Rao-Blackwell

Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\varphi(T) := \mathbb{E}(W \mid T)$. Then

(i) $\mathbb{E}_\theta \varphi(T) = \tau(\theta)$.

(ii) $\mathrm{Var}_\theta \varphi(T) \leq \mathrm{Var}_\theta W \ \forall \theta$.

That is , $\varphi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Therefore, conditioning any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only statistics that are functions of a sufficient statistic in our search for best unbiased estimator.

In fact, conditioning on anything will result in an improvement, but the problem is that the resulting quantity will probably depend on θ and therefore not be an estima-tor.

We now state and prove a powerful result stating that a best unbiased estimator is unique.

Theorem 7.5:

If W is a best unbiased estimator of $\tau(\theta)$ then W is unique.

Proof:

Suppose that W' is another best unbiased estimator, and consider the estimator $W^* = \frac{1}{2}(W + W')$. Note that $\mathbb{E}_\theta W^* = \tau(\theta)$ and

$$\begin{aligned}\mathrm{Var}_\theta W^* &= \mathrm{Var}_\theta\left(\frac{1}{2}W + \frac{1}{2}W'\right) = \frac{1}{4}\mathrm{Var}_\theta W + \frac{1}{4}\mathrm{Var}_\theta W' + \frac{1}{2}\mathrm{Cov}_\theta(W, W') \\ &\leq \frac{1}{4}\mathrm{Var}_\theta W + \frac{1}{4}\mathrm{Var}_\theta W' + \frac{1}{2}\big(\mathrm{Var}_\theta W \cdot \mathrm{Var}_\theta W'\big)^{\frac{1}{2}} \quad \text{(Cauchy-Schwartz)} \\ &= \mathrm{Var}_\theta W. \qquad (\mathrm{Var}_\theta W = \mathrm{Var}_\theta W' \text{ by assumption})\end{aligned}$$

But if the above inequality is strict, then the best unbiasedness of W is contradicted, so we must have equality for all θ. Since the inequality is an application of Cauchy-Schwartz we can have equality only if $W' = a(\theta)W + b(\theta)$. Now applying properties of covariance, we have

$$\begin{aligned}\mathrm{Cov}_\theta(W, W') &= \mathrm{Cov}_\theta\big(W, a(\theta)W + b(\theta)\big) \\ &= \mathrm{Var}_\theta(W, a(\theta)W) = a(\theta)\mathrm{Var}_\theta W,\end{aligned}$$

but $\mathrm{Cov}_\theta(W, W') = \mathrm{Var}_\theta W$ hence $a(\theta) = 1$. Since $\mathbb{E}_\theta W' = \tau(\theta)$ we must have $b(\theta) = 0$ therefore $W = W'$, uniqueness follows.

□

To see when an unbiased estimator is best unbiased, we might ask how could we improve upon a given unbiased estimator? The relationship of an unbiased estimator W with unbiased estimators of 0 (i.e. $\mathbb{E}_\theta U = 0 \forall \theta$) is crucial in evaluating whether W is best unbiased. This relationship, in fact, characterizes the best unbiasedness.

Theorem 7.6:

If $\mathbb{E}_\theta W = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta) \Leftrightarrow W$ is uncorrelated with all unbiased estimators of 0.

Remark: Random Noise

Note that an unbiased estimator of 0 is nothing more than random noise; i.e. there is no information in an estimator of 0. Therefore, if an estimator could be improved by adding random noise to it, the estimator probably is defective. ||

Although we now have an interesting characterization of BUEs, its usefulness is limited in application. It is often a difficult task to verify that an estimator is uncorrelated with all unbiased estimators of 0 since it is usually difficult to describe all unbiased estimators of 0.

It is worthwhile to note once again that what is important is the completeness of the family of distributions of the sufficient statistic. Completeness of the original family is of no consequence. This follows from the Rao-Blackwell Theorem, which says that we can restrict attention to functions of a sufficient statistic, so all expectations will be taken with respect to its distribution.

We sum up the relationship between completeness and best unbiasedness in the following theorem.

Theorem 7.7:

Let T be a complete sufficient statistic for a parameter θ and let $\varphi(T)$ be any estimator based only on T. Then $\varphi(T)$ is the unique BUE of its expected value.

In many situations, there will be no obvious candidate for an unbiased estimator of a function $\tau(\theta)$, much less a candidate for BUE. However, in the presence of completeness, **Theorem 7.7** tells us that if we can find any unbiased estimator, then we can find the best unbiased estimator.

Theorem 7.8: Lehmann-Scheffé

Unbiased estimators based on complete sufficient statistics are unique.

The last method we introduce in this subsection is the loss function optimality. So far, our evaluations of point estimators have been based on their MSE, which is a special case of a function called a loss function. The study of the performance, and the optimality, of estimators evaluated through loss functions is a branch of decision theory.

Definition: Action Space

After the data $X = x$ is observed, where $X \sim f(x \mid \theta)$ for $\theta \in \Theta$, a decision regarding θ is made. The set of all allowable actions are then called the action space, denoted as $\mathcal{A}$.

Remark:

Often in point estimation problems $\mathcal{A}$ is equal to Θ, the parameter space, but this will change in other problems such as hypothesis testing. ||

The loss function in a point estimation problem reflects the fact that if an action a is close to θ, then the decision a is reasonable and little loss is incurred. Therefore the loss function is a nonnegative function that generally increases as the distance between a and θ increases. If θ is real-valued, two commonly used loss functions are

$$\text{Absolute Error Loss, } L(\theta, a) = |a - \theta|,$$

and

Squared Error Loss, $L(\theta, a) = (a - \theta)^2$.

Definition: Risk Function

In a loss function or decision theoretic analysis, the quality of an estimator is quantified in its risk function; i.e. for an estimator $\delta(x)$ of θ, the risk function, a function, a function of θ, is $R(\theta, \delta) := \mathbb{E}_\theta L(\theta, \delta(X))$.

Since the true value of θ is unknown, we would like to use an estimator that has a small value of $R(\theta, \delta)$ for all values of θ. This would mean that, regardless of the true value of θ, the estimator will have a small expected loss. If the qualities of two different estimators, δ_1 and δ_2, are to be compared, then they will be compared by comparing their risk functions $R(\theta, \delta_1)$ and $R(\theta, \delta_2)$. If $R(\theta, \delta_1) < R(\theta, \delta_2)$ for all θ then δ_1 is preferred. More typically, the two risk functions will cross. Then the judgement as to which estimator is better may not be so clear-cut.

8.1 Methods of Finding Hypothesis Tests

We have studied in last section a method of inference called point estimation. Now we move to another inference method called the hypothesis testing. We follow the same structure as we did in the last section to start with finding and then evaluating.

Definition: Hypothesis

A hypothesis is a statement about a population parameter.

The definition of a hypothesis is rather general, but the improvement point is that a hypothesis makes a statement about the population. The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.

Definition: Null and Alternative Hypothesis

The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

In a hypothesis testing problem, after observing the sample the experimenter must decide either to accept H_0 as true or to reject H_0 as false and decide H_1 is true.

Definition: Hypothesis Testing Procedure/ Hypothesis Test

A hypothesis testing procedure or hypothesis test is a rule that specifies

(i) For which sample values the decision is made to accept H_0 as true.

(ii) For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

The likelihood ratio method of hypothesis testing is related to the maximum likelihood estimators and likelihood ratio tests are as widely applicable as maximum likelihood estimation. Recall that if $X_1, \cdots, X_n$ is a random sample from a population with pdf or pmf $f(x \mid \theta)$ (θ may be a vector), the likelihood function is defined as

$$L(\theta \mid x_1, \cdots, x_n) = L(\theta \mid x) = f(x \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

Let Θ denote the entire parameter space. Likelihood ratio tests are defined as follows.

Definition: Likelihood Ratio Test Statistic

The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is $\lambda(x) := \dfrac{\sup_{\Theta_0} L(\theta \mid x)}{\sup_{\Theta} L(\theta \mid x)}$.

Definition: Likelihood Ratio Test (LRT)

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{x \mid \lambda(x) \leq c\}$ where c is any constant such that $0 \leq c \leq 1$.

Recall that in the MLE, the maximization of the likelihood function is, not about making the data itself more probable but rather about finding the parameter values that make the observed data most consistent with the assumed model. The motivation for the LRT is quite the same.

It coule be best interpreted in the situation in which $f(x \mid \theta)$ is a pmf of a discrete r-andom variable. In this case, the numeraotr is maximized over the whole parameter space Θ while the denominator is maximized over the Θ_0. The less the ratio is shows that more consistent our model is.

Connectio with MLEs:

If we think of maximizing over both the entire parameter space and a subset of the parameter space, then the correspondence between the LRTs and MLEs become very clear. Suppose that $\hat{\theta}$, an MLE of θ, exists; $\hat{\theta}$ is obtained by doing an unrestricted maximization of $L(\theta \mid x)$. We can also consider the MLE of θ, call I $\hat{\theta}_0$, obtained by doing the restriced maximization, assuming that Θ_0 is the parameter space. That is, $\hat{\theta}_0 = \hat{\theta}_0(x)$ is the value of $\theta \in \Theta_0$ that maximizes $L(\theta \mid x)$. Then, the LRT statistics is given by $\lambda(x) = \dfrac{L(\hat{\theta}_0 \mid x)}{L(\hat{\theta} \mid x)}$.

For a sufficient statistic of a random sample X, namely $T(X)$, we know that all the information about θ could be found in $T(X)$, the test based on T should be as good as the test based on the complete sample X. In fact, the tests are equivalent.

Theorem 8.1:

If $T(X)$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(x)$ are the LRT statistics based on T and X, respectively. Then $\lambda^*(T(x)) = \lambda(x)\ \forall x \in \Omega_X$.

Proof:

According to the **Factorization Theorem**, the pdf or pmf of X can be written as $f(x \mid \theta) = g(T(x) \mid \theta)h(x)$, where $g(t \mid \theta)$ is the pdf or pmf of T and $h(x)$ does not depend on θ. Thus,

$$\begin{aligned}
\lambda(x) := \frac{\sup_{\Theta_0} L(\theta \mid x)}{\sup_{\Theta} L(\theta \mid x)} &= \frac{\sup_{\Theta_0} f(x \mid \theta)}{\sup_{\Theta} f(x \mid \theta)} \\
&= \frac{\sup_{\Theta_0} g(T(x) \mid \theta)h(x)}{\sup_{\Theta} g(T(x) \mid \theta)h(x)} && (T \text{ is sufficient}) \\
&= \frac{\sup_{\Theta_0} g(T(x) \mid \theta)}{\sup_{\Theta} g(T(x) \mid \theta)} && (h \text{ does not depend on } \theta)
\end{aligned}$$

$$= \frac{\sup_{\Theta_0} L^*(\theta \mid T(x))}{\sup_{\Theta} L^*(\theta \mid T(x))} \qquad (g \text{ is the pdf or pmf of } T)$$
$$=: \lambda^*(T(x)).$$

□

Now we move to the Bayesian tests. One way a Bayesian hypothesis tester may choose to use the posterior distribution is to decide to accept H_0 as true if

$$\mathbb{P}(\theta \in \Theta_0 \mid X) \geq \mathbb{P}(\theta \in \Theta_0^c \mid X)$$

and to reject H_0 otherwise. In the terminology of the previous sections, the test statistic, a function of the sample, is $\mathbb{P}(\theta \in \Theta_0^c \mid X)$ and the rejection region is given by $\left\{x \,\middle|\, \mathbb{P}(\theta \in \Theta_0^c \mid x) > \frac{1}{2}\right\}$. Alternatively, if the Bayesian hypothesis testers wish to guard against falsely rejecting H_0, he must decide to reject H_0 only if $\mathbb{P}(\theta \in \Theta_0^c \mid X)$ is greater than some certain large number, say, 0.99.

In some situations, tests for complicated null hypothesis can be developed from tests for simpler null hypothesis. We will discuss two methods to close this subsection.

Algorithm 8.2: Union-Intersection Method

The Union-Intersectio method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection. Namely, $H_0 : \theta \in \bigcap_{\gamma \in \Gamma} \Theta_\gamma$, where Γ is an arbitrary index set. Suppose that tests are available for each of the problems of testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Say the rejection region for the test of $H_{0\gamma}$ is $\{x \mid T_\gamma(x) \in R_\gamma\}$. Then the rejection region for the union-intersection test is $\bigcup_{\gamma \in \Gamma} \{x \mid T_\gamma(x) \in R_\gamma\}$.

The rationale is simple. If any one of the hypothesis $H_{0\gamma}$ is rejected then H_0 should be rejected. On the other hand, H_0 is true only if each of the hypothesis $H_{0\gamma}$ is accepted as true.

In some cases a simple expression for the rejection region of a Union-Intersection test can be found. In particular, suppose that each of the individual test has a rejection region of the form $\{x \mid T_\gamma(x) > c\}$, where c does not depend on γ. The rejection region for the union-intersection test can be expressed as

$$\bigcup_{\gamma \in \Gamma} \{x \mid T_\gamma(x) > c\} = \{x \mid \sup_{\gamma \in \Gamma} T_\gamma(x) > c\}.$$

Thus the test statistic for testing H_0 is $T(x) = \sup_{\gamma \in \Gamma} T_\gamma(x)$.

The Union-Intersection method of test construction is useful if the null hypothesis is conveniently expressed as an intersection. Another method, the Intersection-Union method, may be useful if the null hypothesis is conveniently expressed as a union.

Algorithm 8.3: Intersection-Union Method

Suppose we wish to test the null hypothesis $H_0 : \theta \in \bigcup_{\gamma \in \Gamma} \Theta_\gamma$. Suppose that for

each $\gamma \in \Gamma$, $\{x \mid T_\gamma(x) \in R_\gamma\}$ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$. Then the rejection region for the Intersection-Union test of H_0 versus H_1 is $\bigcap_{\gamma\in\Gamma} \{x \mid T_\gamma(x) \in R_\gamma\}$. H_0 is false if and only if all of the $H_{0\gamma}$ is false, so H_0 can be rejected if and only if each of the individual hypothesis $H_{0\gamma}$ can be rejected.

Again, the Intersection-Union test can be greatly simplified if the rejection regions for the individual hypothesis are all of the form $\{x \mid T_\gamma(x) \geq c\}$, where c is independent of γ. In such cases, the rejection region of H_0 is

$$\bigcap_{\gamma\in\Gamma} \{x \mid T_\gamma(x) \geq c\} = \{x \mid \inf_{\gamma\in\Gamma} T_\gamma(x) \geq c\}.$$

Here, the Intersection-Union test statistic is $\inf_{\gamma\in\Gamma} T_\gamma(x)$, and the test rejects H_0 for large values of this statistic.

8.2 Methods of Evaluating Tests

In deciding to accept or reject the null hypothesis H_0, an experimenter might be making a mistake. Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes. In this subsection we discuss how these error probabilities can be controlled. In some cases, it can even be determined which tests have the smallest possible error probabilities.

We will go through five methods in this subsection, they are: (1) Error Probabilities and Power Function, (2) Most Powerful Tests, (3) Sizes of Union-Intersection and Intersection-Union Tests, and (4) p-Values. We now start with the first one.

A hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ might make one or two types of errors. These two types of errors traditionally have been given the names Type I Error and Type II Error.

Definition: Type I Error

If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a Type I Error.

Definition: Type II Error

If $\theta \in \Theta_0^c$ but the hypothesis test incorrectly decides to accept H_0, then the test has made a Type II Error.

Suppose that R denotes the rejection region for a test. Then for $\theta \in \Theta_0$, the test will make a mistake if $x \in R$, so the probability of a Type I Error is $\mathbb{P}_\theta(X \in R)$. For $\theta \in \Theta_0^c$, the probability of a Type II Error is $\mathbb{P}_\theta(X \in R^c)$. This switching from R to R^c is a bit confusing but if we realize that $\mathbb{P}_\theta(X \in R^c) = 1 - \mathbb{P}_\theta(X \in R)$. This consideration leads to the following definition of the power function.

Definition: Power Function

The power function of a hypothesis test with rejection region R is the function of θ defined by

$$\beta(\theta) := \mathbb{P}_\theta(X \in R) = \begin{cases} \text{probability of a Type I Error}, \theta \in \Theta_0 \\ \text{one minus the probability of a Type II Error}, \theta \in \Theta_0^c \end{cases}.$$

Remark:

The ideal power function is 0 $\forall \theta \in \Theta_0$ and 1 $\forall \theta \in \Theta_0^c$. Except in trivial situations, this ideal cannot be attained. Qualitively, a good test has power function near 1 for most $\theta \in \Theta_0^c$ and near 0 for most $\theta \in \Theta_0$. ||

Typically, the power function of a test will depend on the sample size n. If n can be chosen by the experimenter, consideration of the power function might be helpful in determining what sample size is appropriate for an experiment.

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small. In searching for a good test, it is common to restrict consideration to tests that control the Type I Error probability at a specified level. Within this class of tests we then search for tests that have Type II Error probability that is as small as possible. The following two terms are useful when discussing tests that control Type I Error probabilities.

Definition: Size α Test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$.

Definition: Level α Test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$.

Some authors do not make distinction between these two definitions. We made the distinction here to stress out the fact that sometimes having a size α test is difficult, so in practical terms, one should make compromises with the alternative level α test.

Remark:

Typical α level tests use $\alpha = 0.01$, 0.05, and 0.10, **but be aware that in fixing the level α test, the experimenter is controlling only the Type I Error.** An LRT is one rejects H_0 if $\lambda(X) \leq c$, for example. ||

Other than α levels, there are other features of a test that might also be of concern. For example, we would like a test to be more likely to reject H_0 if $\theta \in \Theta_0^c$ than if $\theta \in \Theta_0$. This property is called unbiased.

Definition: Unbiased Power Function

A test with power function $\beta(\theta)$ is unbiased if $\beta(\theta') \geq \beta(\theta'') \forall \theta' \in \Theta_0^c$ and $\forall \theta'' \in \Theta_0$.

In most problems there are many unbiased tests. Likewise, there are many size α tests, LRTs, etc. In some cases we have imposed enough restrictions to narrow the consideration to one test. In other cases there remain many tests from which to choose. We discussed only the one that rejects H_0 for large values of T. In the following discussion we will discuss other criteria for selecting one out of a class of tests, criteria that are all related to the power functions of the tests.

We have seen that the α tests could control the probability of a Type I Error, i.e. level α tests have Type I Error probabilities at most α for all $\theta \in \Theta_0$. A good test in such a class would also have a small Type II Error probability, i.e. a large power function for $\theta \in \Theta_0^c$. If one test has a smaller Type II Error probability than all other tests in the class, it would certainly be a strong contender for the best test in the class, a notion that is formalized in the next definition.

Definition: Uniformly Most Powerful (UMP) Test

Let $\mathscr{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathscr{C}$, with power function $\beta(\theta)$, is a uniformly most powerful class $\mathscr{C}$ test if $\beta(\theta) \geq \beta'(\theta) \forall \theta \in \Theta_0^c$ and $\forall \beta' \in \mathscr{C}$.

In this subsection, the class $\mathscr{C}$ will be the class of all level α tests. The test described in the above definition is then called a UMP level α test. For this test to be interesting, restriction to the class $\mathscr{C}$ must involve some restriction on the Type I Error probability. A minimization of the Type II Error probability without some control of the Type I Error is not very interesting.

The requirements in this definition are so strong that UMP does not exist in many realistic problems. But in problems that have UMP tests, a UMP test might well be considered the best test in the class. Thus, we would like to be able to identify UMP tests if they exist. The following famous theorem clearly describes which tests are UMP level α tests in the situation where the null and alternative hypotheses both consist of only one probability distribution for the sample.

Theorem 8.4: Neymann-Pearson Lemma

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(x \mid \theta_i)$, $i = 0,1$, using a test with rejection region R such that

(i) $x \in R$, if $f(x \mid \theta_1) > k f(x \mid \theta_0)$,

(ii) $x \in R^c$, if $f(x \mid \theta_1) < k f(x \mid \theta_0)$,

for some $k \geq 0$ and $\alpha = \mathbb{P}_{\theta_0}(X \in R)$. Then

(a) Any test that satisfies (i) and (ii) is a UMP level α test. (Sufficiency)

(b) If there exists a test satisfies (i) and (ii) with $k > 0$, then every UMP level α test is a size α test and every UMP level α test satisfies the first condition except perhaps on a set with probability measure 0, i.e. on a set A such that $\mathbb{P}_{\theta_0}(X \in A) = \mathbb{P}_{\theta_1}(X \in A) = 0$. (Necessity)

The following corollary connects the Neyman-Pearson Lemma to sufficiency.

Corollary 8.4.1:

Under the same settings as in **Theorem 8.4**. Suppose that $T(X)$ is a sufficient statistic for θ and $g(t \mid \theta_i)$ is the pdf or pmf of T corresponding to θ_i for $i = 0,1$. Then any test based on T with rejection region S is a UMP level α test if it satisfies

(1) $t \in S$, if $g(t \mid \theta_1) > k g(t \mid \theta_0)$,

(2) $t \in S^c$, if $g(t \mid \theta_1) < k g(t \mid \theta_0)$,

for some $k \geq 0$, where $\alpha = \mathbb{P}_{\theta_0}(T \in S)$.

Hypotheses, such as H_0 and H_1 in the **Neyman-Pearson Lemma**, that specify only one possible distribution for the sample X are called simple hypotheses. In most realistic problems however, the hypotheses of interest specify more than one possible distribution for the sample. Such hypotheses are called composite hypotheses. Since the definition of UMP requires the test to be most powerful against each individual $\theta \in \Theta_0^c$, the **Neyman-Pearson Lemma** can be used to find UMP tests in problems involving composite hypotheses.

In particular, hypotheses that assert that a univariate parameter is large, for example, $H : \theta \geq \theta_0$, or small, e.g. $H : \theta < \theta_0$, are called one-sided hypotheses. Hypotheses that assert that a parameter is either large or small, e.g. $H : \theta \neq \theta_0$, are called two-sided hypotheses. A large class of problems that admit UMP level α test involve one-sided hypotheses and pdfs or pmfs with the monotone likelihood raito property, which is given below.

Definition: Monotone Ratio Likelihood Ratio (MLR)

> A family of pdfs or pmfs $\{g(t \mid \theta) \mid \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a monotone likelihood ratio (MLR) if, for every $\theta_2 > \theta_1$, $g(t \mid \theta_2)/g(t \mid \theta_1)$ is monotone (nonincreasing or nondecreasing) function of t on $\{t \mid g(t \mid \theta_1) > 0 \text{ or } g(t \mid \theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $0 < c$.

Many common families of distributions have an MLR. For example, the normal (known variance, unknown mean), the Poisson, and binomial all have an MLR. Indeed, any regular exponential family with $g(t \mid \theta) = h(t)c(\theta)e^{w(\theta)t}$ has an MLR if $w(\theta)$ is a nondecreasing function.

Theorem 8.5: Karlin-Rubin

> Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t \mid \theta) \mid \theta \in \Theta\}$ of T has an MLR then for any t_0, the test that rejects $H_0 \Leftrightarrow T > t_0$ is a UMP level α test where $\alpha = \mathbb{P}_{\theta_0}(T > t_0)$.

By an analogous argument, it can be shown that under the conditons of Karlin-Rubin, the test that rejects $H_0 : \theta \geq \theta_0$ in favor of $H_1 : \theta < \theta_0 \Leftrightarrow T < t_0$ is a UMP level α test with $\alpha = \mathbb{P}_{\theta_0}(T < t_0)$.

Now we move to the third topic in this subsection. Recall that because of the simple way in which they are constructed, the sizes of union-intersection tests (UIT) and intersection-union tests (IUT) can often be bounded above by the sizes of some other tests. Such bounds are useful if a level α test is wanted, but the size of UIT or IUT is too difficult to evaluate. We now discuss these bounds.

First consider UITs. Recall that in this situation, we are testing a null hypothesis of the form $H_0 : \theta \in \Theta_0$, where $\Theta_0 := \bigcap_{\gamma \in \Gamma} \Theta_\gamma$. To be specific, let $\lambda_\gamma(x)$ be the LRT statistic for testing $H_{0\gamma} : \theta \in \Theta_\gamma$ versus $H_{1\gamma} : \theta \in \Theta_\gamma^c$, and let $\lambda(x)$ be the LRT statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. Then we have the following relationships between the overall LRT and the UIT based on $\lambda_\gamma(x)$.

Theorem 8.6:

Consider testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ where $\Theta_0 := \bigcap_{\gamma \in \Gamma} \Theta_\gamma$ and $\lambda_\gamma(x)$ is defined as above. Define $T(x) := \inf_{\gamma \in \Gamma} \lambda_\gamma(x)$, and form the UIT with rejection region $\{x \mid \lambda_\gamma(x) < c \text{ for some } \gamma \in \Gamma\} = \{x \mid T(x) < c\}$. Also consider the usual LRT with rejection region $\{x \mid \lambda(x) < c\}$. Then

(a) $T(x) \geq \lambda(x)$ for all x.

(b) If $\beta_T(x)$ and $\beta_\lambda(x)$ are the power functions for the tests based on T and λ, respectively, then $\beta_T(\theta) \leq \beta_\lambda(\theta)$ for every $\theta \in \Theta$.

(c) If the LRT is a level α test, then the UIT is a level α test.

Since the LRT is uniformly more powerful in the above theorem than UIT, we might ask why we should use the UIT. One reason is that UIT has a smaller Type I Error probability for every $\theta \in \Theta_0$. Moreover, if H_0 is rejected, we may wish to look at the individual tests of $H_{0\gamma}$ to see why, for which UIT provides us an access.

We now investigate the sizes of IUTs. A simple bound for the size of an IUT is related to the sizes of the individual tests that are used to define the IUT. Recall that in this situation the null hypothesis is expressible as a union, i.e. we are testing

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_0^c, \text{ where } \Theta_0 = \bigcup_{\gamma \in \Gamma} \Theta_\gamma.$$

An IUT has a rejection region of the form $R = \bigcap_{\gamma \in \Gamma} R_\gamma$ where R_γ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$.

Theorem 8.7:

Let α_γ be the size of the test of $H_{0\gamma}$ with rejection region R_γ. Then the IUT with rejection region $R = \bigcap_{\gamma \in \Gamma} R_\gamma$ is a level $\alpha = \sup_{\gamma \in \Gamma} \alpha_\gamma$ test.

Typically, the individual rejection regions R_γ are chosen so that $\alpha_\gamma = \alpha \, \forall \gamma$. In such a case, **Theorem 8.7** states that the resulting IUT is a level α test. Moreover, this theorem provides an upper bound for the size of an IUT, is somewhat more useful than **Theorem 8.6**, which provides an upper bound for the size of a UIT.

Remark:

Theorem 8.6 applied only to UITs constructed from LRTs while **Theorem 8.7** applies to any IUT. ||

The bound in **Theorem 8.6** is the size of the LRT, which, in a complicated problem, may be difficult to compute. In **Theorem 8.7** however, the LRT need not be used to obtain the upper bound. Any test $H_{0\gamma}$ with unknown size α_γ can be used, and then the upper bound on the size of the IUT is given in terms of the known sizes $\alpha_\gamma, \gamma \in \Gamma$.

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_0 or accept H_0. The size of the test carrise important information. If α is small, the decision to reject H_0 is

fairly convincing, but if α is large, the decision to reject H_0 is not very convincing since the test has a large probability of incorrectly making that decision. Another way of reporting the results of a hypothesis test is to report the value of a certain kind of test statistic called a p-value.

Definition: p-Value

A p-value $p(X)$ is a test statistic satisfying $0 \leq p(x) \leq 1$ for every sample point x. Small values of $p(X)$ give evidence that H_1 is true. A p-value is valid if $\forall \theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$, $\mathbb{P}_\theta(p(X) \leq \alpha) \leq \alpha$.

If $p(X)$ is valid it is then easy to construct a level α test based on $p(X)$. The test that rejects H_0 if and only if $p(X) \leq \alpha$ is a level α test. An advantage to reporting a test result via a p-value is that each reader can choose the α and then can compare the reported $p(x)$ to α and know whether these data lead to acceptance or rejection of H_0. Morover, the smaller the p-value, the stronger the evidence for rejecting H_0. Hence, a p-value reports the results of a test on a more continuous scale, rather than just accepting H_0 or Rejecting H_0.

The most common way to define a valid p-value is given by the following result.

Theorem 8.8:

Let $W(X)$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point x, define $p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(X) \geq W(x))$. Then, $p(X)$ is valid.

9.1 Methods of Finding Interval Estimators

We have seen in Section 7 for the inference of a single value as the value of θ. In this subsection we focus on extending this concept to an interval. As before, this section is divided into two parts, in Section 9.1 we introduce the methods of finding interval estimators and in Section 9.2 we shall talk about the methods in evaluating them.

Definition: Interval Estimate, Interval Estimator

An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \cdots, x_n)$ and $U(x_1, \cdots, x_n)$, of a sample that satisfy $L(x) \leq U(x)$ $\forall x \in \Omega_X$. If $X = x$ is observed, the inference $L(x) \leq \theta \leq U(x)$ is made. The random interval $\big(L(X), U(X)\big)$ is called an interval estimator.

The purpose of using an interval estimator rather than a point estimator is to have some guarantee of capturing the parameter of interest. The certainty of this guarantee is quantified in the following definitions.

Definition: Coverage Probability

For an interval estimator $\big(L(X), U(X)\big)$ of a parameter θ, the coverage probability of $\big(L(X), U(X)\big)$ is the probability that the random interval $\big(L(X), U(X)\big)$ covers the true parameter θ. In symbols, it is denoted by either $\mathbb{P}_\theta\Big(\theta \in \big(L(X), U(X)\big)\Big)$ or $\mathbb{P}\Big(\theta \in \big(L(X), U(X)\big) \Big| \theta\Big)$.

Definition: Confidence Coefficient

For an interval estimator $\big(L(X), U(X)\big)$ of a parameter θ, the confidence coefficient of $\big(L(X), U(X)\big)$ is the infimum of the coverage probability, i.e. $\inf_{\theta} \mathbb{P}_{\theta}\Big(\big(L(X), U(X)\big)\Big)$.

Interval estimators together with a measure of confidence (usually a confidence coefficient) are sometimes called confidence intervals. A confidence set with confidence coefficient equal to some value, say $1-\alpha$, is simply called a $1-\alpha$ confidence set.

There is a very strong correspondence between hypothesis testing and interval estimation. In fact, we can say in general that every confidence set corresponds to a test and vice versa.

The hypothesis test fixes the parameter and asks what sample values (the acceptance region) are consistent with the fixed value. The confidence set fixes the sample and asks what parameter values (the confidence interval) make this sample value most plausible. This correspondence between acceptance region and confidence intervals hold in general. We state it in the following theorem.

Theorem 9.1:

For each $\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $x \in \Omega_X$ define $C(x) := \{\theta_0 \,|\, \theta_0 \in A(x)\}$. Then the random set $C(X)$ is a $1-\alpha$ confidence set. Conversely, let $C(X)$ be a $1-\alpha$ confidence set. For any $\theta_0 \in \Theta$ define $A(\theta_0) = \{x \,|\, \theta_0 \in C(x)\}$. Then $A(\theta_0$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Note that the coverage probability for $\{aX, bX\}$ and $\{X+c, X+d\}$ are different for $a, b, c,$ and d constants. One important difference is that the coverage probability of the interval $\{aX, bX\}$ could be expressed by the quantity X/θ, a random variable whose distribution does not depend on the parameter, while $\{X+c, X+d\}$ depends on θ. The quantity X/θ is known as a pivotal quantity, or simply pivot.

Definition: Pivot

A random variable $Q(X, \theta) = Q(X_1, \cdots, X_n, \theta)$ is a pivot if the distribution of $Q(X, \theta)$ is dependent on all parameters. That is, if $X \sim F(x \,|\, \theta)$ then $Q(X, \theta)$ has the same distribution for all values of θ.

Theorem 9.2: Pivoting a Continuous CDF

Let T be a statistic with continuous cdf $F_T(t \,|\, \theta)$. Let $\alpha_1 + \alpha_2 =: \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows:

(i) If $F_T(t \,|\, \theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $F_T(t \,|\, \theta_U(t)) = \alpha_1$ and $F_T(t \,|\, \theta_L(t)) = 1 - \alpha_2$.

(ii) If $F_T(t \,|\, \theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $F_T(t \,|\, \theta_U(t)) = 1 - \alpha_2$ and $F_T(t \,|\, \theta_L(t)) = \alpha_1$.

Then the random inverval $\big(\theta_L(T), \theta_U(T)\big)$ is a $1-\alpha$ confidence interval for θ.

Theorem 9.3: Pivoting a Discrete CDF

Let T be a discrete statistic with cdf $F_T(t \,|\, \theta) = \mathbb{P}(T \le t \,|\, \theta)$. Let $\alpha_1 + \alpha_2 =: \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathcal{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows:

(i) If $F_T(t\,|\,\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $\mathbb{P}(T \le t\,|\,\theta_U(t)) = \alpha_1$ and $\mathbb{P}(T \ge t\,|\,\theta_L(t)) = \alpha_2$.

(ii) If $F_T(t\,|\,\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $\mathbb{P}(T \ge t\,|\,\theta_U(t)) = \alpha_1$ and $\mathbb{P}(T \le t\,|\,\theta_L(t)) = \alpha_2$.

Then the random inverval $\big(\theta_L(T), \theta_U(T)\big)$ is a $1-\alpha$ confidence interval for θ.

9.2 Methods of Evaluating Interval Estimators

Directly from the definition of the interval estimator, we could tell that with the smaller the "length" is, we have a better estimator; on the other hand, if the interval covers the parameter with high probability, we can say the estimator is good. Therefore there are two scales to describe the performance of the estimators.

Definition: Unimodal

A pdf $f(x)$ is unimodal if there exists x^* such that $f(x)$ is nondecreasing for $x \le x^*$ and $f(x)$ is nonincreasing for $x \ge x^*$.

Theorem 9.4:

Let $f(x)$ be a unimodal pdf. If the interval $[a, b]$ satisfies

(i) $\int_a^b f(x)dx = 1-\alpha$.

(ii) $f(a) = f(b) > 0$.

(iii) $a \le x^* \le b$, where x^* is a mode of $f(x)$.

Then $[a, b]$ is the shortest among all intervals that satisfies (i).

In some cases, especially when working outside of the location problem, we must be careful in the application of this theorem. In scale cases in particular, the theorem may not directly applicable, but a variant may be.

Since there is a one-to-one correspondence between confidence sets and tests of hypothese, there is some correspondence between optimality of tests and optimality of confidence sets. Usually, test-related optimality properties of confidence sets do not directly relate to the size of the set but rather to the probability of the set covering false values.

The probability of covering false values, or the probability of false coverage, indirectly measures the size of a confidence set. Intuitively, smaller sets cover fewer values and, hence, are less likely to cover false values.

Definition: Uniformly Most Accurate (UMA) Confidence Set

A $1-\alpha$ confidence set that minimizes the probability of false coverage over a class of $1-\alpha$ confidence set is called a uniformly most accurate (UMA) confidence set.

Theorem 9.5: UMA Lower Confidence Bound

Let $X \sim f(x\,|\,\theta)$ where θ is a real-valued parameter. For each $\theta_0 \in \Theta$, let $A^*(\theta_0)$ be the UMP level α acceptance region of a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Let $C^*(x)$ be the $1-\alpha$ confidence set formed by inverting the UMP acceptance regions. Then for any other $1-\alpha$ confidence set C, $\mathbb{P}_\theta(\theta' \in C^*(X)) \le \mathbb{P}_\theta(\theta' \in C(X))$ for all $\theta' < \theta$.

Definition: Unbiased

A $1-\alpha$ confidence set $C(x)$ is unbiased if $\mathbb{P}_\theta(\theta' \in C(X)) \leq 1-\alpha \ \forall \theta \neq \theta'$.

Sets that minimize the probability of false coverage are called Neyman-shortest. The fact that there is a length connotation to this name is somewhat justified by the following theorem.

Theorem 9.6: Pratt

Let X be a real-valued random variable with $X \sim f(x \mid \theta)$ where θ is a real-valued parameter. Let $C(x) = \big(L(X), U(X)\big)$ be a confidence interval for θ. If $L(x)$ and $U(x)$ are both increasing functions of x, then for any value θ^*,

$$\mathbb{E}_{\theta^*}\big(\text{Length}(C(X))\big) = \int_{\theta \neq \theta^*} \mathbb{P}_{\theta^*}(\theta \in C(X))d\theta.$$

The result is that the expected length of $C(x)$ is equal to a sum (integral) of the probabilities of the false coverage, the integral being taken over all false values of the parameter θ.

The goal of obtaining a smallest confidence set with a specified coverage probability can also be attained using Bayesian criteria. If we have a posterior distribution $\pi(\theta \mid x)$ the posterior distribution of θ given $X = x$, we would like to find the set $C(x)$ that satisfy

(i) $\displaystyle\int_{C(x)} \pi(\theta \mid x)dx = 1-\alpha,$

(ii) $\text{Size}(C(x)) \leq \text{Size}(C'(x)),$

for any set $C'(x)$ satisfying $\displaystyle\int_{C'(x)} \pi(\theta \mid x)dx \geq 1-\alpha$. If we take our measure of size to be length, then we apply **Theorem 9.4** and obtain the following result.

Corollary 9.7:

If the posterior density $\pi(\theta \mid x)$ is unimodal, then for a given value of α, the shortest credible interval for θ is given by

$$\{\theta \mid \pi(\theta \mid x) \geq k\} \text{ where } \int_{\{\theta \mid \pi(\theta \mid x) \geq k\}} \pi(\theta \mid x)d\theta = 1-\alpha.$$

The credible set in this corollary is called a highest posterior density (HPD) region, as it consists of the values of the parameter for which the posterior density is highest. Notice the similarity in form between the HPD region and the likelihood region.

Reference:

[1]: George Casella, Roger L. Berger, *Statistical Inference, Second Edition*, Duxbury Thomson Learning.

[2]: Fan Yang, *Lecture Notes on Statistical Theory,* Lecture Given to Tsinghua University at Fall 2022, available online.

[3]: Tianyu Zhang, *Lecture Notes on Probability on Banach spaces*, unpublished version, available online.

[4]: Xiangdong Li, *Lecture Notes on Optimal Transportation Problems*, Lecture Given to Tsinghua University at Spring 2023, available online.

[5]: Shu Cherng Fang, Sarat Puthenpura, *Linear Optimization and Extensions: Theory and Algorithms*.

[6]: Gordan Zitkovic, *Theory of Probability I — Weak Convergence*, available online.