# Statistical Theory of Economists

Notes from ECON3320

Mark Chiu Chong

Last updated: Semester 1, 2021

# Contents

# 1 Overview of Probability Theory

The basic idea of *classical statistics* is that by studying a sample from a population, we can infer (estimate and test hypotheses) about various properties of this population. We can make inferences about:

- Student populations (e.g., are this year's students better than last year's students?).
- Australia's population (e.g., average heights and weights of adult males aged 20 to 25).
- Inferences on world population (e.g., how unequal is the distribution of income in the world? Is inequality increasing over time?).

### What is Probability?

Probability is a commonly used term to express the *degree of belief* associated with a given event. It is measured on a scale of 0 to 1 - low probability events are assigned probabilities close to zero and high probability to events considered certain are assigned probabilities close to 1.

How do we assign probabilities? In classical statistics, an objective process underpins the assignment of probabilities. The probability of an event is understood as the *limit of the empirical frequency of the event* if it were observed (hypothetically) infinitely often, e.g.,

$$\lim_{n\to\infty} \frac{\#\text{ of time } H \text{ appeared in } n \text{ trials}}{n}$$

### A Framework to Study Probability

The following framework is used in a formal approach to study and use the notion of probability of events:

- **Experiment:** The process by which an observation of data is made. tossing a coin twice and observe the outcomes; rolling dice are standard examples.
- **Sample space:** A collection of all possible outcomes of a given experiment. E.g. tossing a coin twice $S = \{HH, HT, TH, TT\}$.
- **Event:** A subset of the sample space. Note that it can be an empty subset or it could be the whole sample space.
- **Simple event:** An event from an experiment that *cannot be decomposed into several events.* Each simple event corresponds to only one outcome of the experiment.
- **Mutually exclusive or disjoint events:** Two events $S_i$ and $S_j$ are said to be pairwise mutually exclusive or disjoint events in $S$ if and only if $S_i \cap S_j = \emptyset$. E.g. events that no $T$ appears and that at least one $T$ appears are mutually exclusive.

## 1.1 Probability Measure

> **Definition (Probability Measure):** Given a sample $S$ from an experiment, the *probability measure* is a real valued function that assigns values in the range $[0, 1]$ for each subset $S_i \subseteq S$ satisfying the following axioms:
>
> 1. $P(S_i) \geq 0, \ \forall \, S_i \subseteq S$;
> 2. $P(S) = 1$, where $S$ is the sample space;
> 3. $P\left(\bigcup\limits_{i=1}^{\infty} S_i\right) = \sum\limits_{i=1}^{\infty} P(S_i), \text{ if } S_i \cap S_j = \emptyset, \ \forall \, i, j \neq i.$

*Remark.* This definition only implies restrictions on the probability measure but does not tell us how it can be actually assigned. The exact value of the probability of a random event depends on the random nature of the event (stochastic process).

> **Definition:** A sequence of events $S_1, S_2, \ldots, S_n \subseteq S$ are called *exhaustive* if $P(S_1 \cup S_2 \cup \cdots \cup S_n) = 1$ and *non-exhaustive* otherwise.

### Properties of Probability Measures

The following properties can be derived from the three axioms of probability measures:

1. $P(\emptyset) = 0$.
2. $P(S_i) \leq 1, \ \forall \, S_i \subseteq S$.
3. $P(S_i^c) = 1 - P(S_i), \ \forall \, S_i \subseteq S$, where $S_i^c$ is the complement of $S_i$ in $S$ ($S_i^c = S \setminus S_i$).
4. $S_i \subseteq S_j \Rightarrow P(S_i) \leq P(S_j), \ \forall \, S_i, S_j \subseteq S$.
5. $P(S_i \cup S_j) = P(S_i) + P(S_j) - P(S_i \cap S_j), \ \forall \, S_i, S_j \subseteq S$.
6. $P(S_i \cup S_j) \leq P(S_i) + P(S_j), \ \forall \, S_i, S_j \subseteq S$.
7. $P\left(\bigcup_i S_i\right) \leq \sum_i P(S_i), \ \forall \, S_i \subseteq S$ (Boole's inequality).
8. If $S_i$ and $S_j$ are disjoint for all $i, j \neq i$ then $P(\cup_i) = \sum_i P(S_i), \ \forall \, S_i \subseteq S$.

## 1.2 Law of Total Probability

As a convention, the probability of a single event is called marginal probability. Probability of several events happening jointly is called joint probability. E.g., $P(S_i \cap S_j)$ denotes the joint probability of events $S_i$ and $S_j$.

Let $S_1, S_2, ..., S_n$ be disjoint and exhaustive events in $S$, and let $A$ be some event of interest, then

$$P(A) = \sum_{i=1}^{n} P(A \cap S_i),$$

i.e., the probability of any event can be represented as the sum of probabilities of this event jointly occurring with all other disjoint events that exhaustively partition $S$. The following result is commonly used:

**Result 1.1.** *Let $A, B \subseteq S$ and $B^c$ be the complement of $B$. Then,*

$$P(A) = P(A \cap B) + P(A \cap B^c).$$

### 1.3 Conditional Probability

> **Definition (Conditional Probability):** The *conditional probability* of an event $S_i$, given a probable event $S_j$ (i.e., $P(S_j) > 0$) is defined as:
>
> $$P(S_i \,|\, S_j) = \frac{P(S_i \cap S_j)}{P(S_j)} \text{ if } P(S_j) > 0.$$

> **Definition (Law of total conditional probabilities):** Let $S_1, S_2, ..., S_n$ be exhaustive, pairwise disjoint and probable events in $S$. Let $A$ be some event of interest in $S$, then
>
> $$P(A) = \sum_{i=1}^{n} P(A \cap S_i) = \sum_{i=1}^{n} P(A \,|\, S_i) P(S_i).$$

This definition is useful in evaluating probabilities for the whole population when information is available for sub-populations. E.g., suppose risk of cancer of a particular type is known for Australian born and overseas born individuals, we can work out the risk of cancer for the Australian population as a whole.

### 1.4 Bayes' Theorem

Bayes theorem provides answers to questions that are reverse to total probability.

> **Definition (Bayes' theorem):** Let $A, B \subseteq S$ be two probable events, then
>
> $$P(A \,|\, B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \,|\, A) P(A)}{P(B)}.$$
>
> Bayes theorem is the cornerstone of Bayesian Econometrics and it provides a mechanism of updating information from observed events.

> **Definition (Generalisation of Bayes' theorem):** Let $S_1, S_2, ..., S_n \subseteq S$ be mutually exclusive, probable and exhaustive events. Let $B$ be an arbitrary probable event, then
> $$P(S_i \,|\, B) = \frac{P(B \,|\, S_i)P(S_i)}{\sum_{j=1}^{n} P(B \,|\, S_j)P(S_j)}.$$

## 1.5 Independence of Events

> **Definition (Independence):** Two events $S_i$ and $S_j$ are said to be *independent* if and only if:
> $$P(S_i \cap S_j) = P(S_i)P(S_j).$$

*Remark.* If events $S_i$ and $S_j$ are independent and if $S_j$ is probable, then

$$P(S_i \,|\, S_j) = P(S_i),$$

i.e., the probability of one does not effect the probability of the other.

*Remark.* Finally, a collection of events is *mutually independent* if and only if for every finite subset $\{S_1, S_2, ..., S_j\}$ of this collection, it is true that

$$P\left(\bigcap_{j=1}^{J} S_j\right) = \prod_{j=1}^{J} P(S_j).$$

## 1.6 Random Variables

In real life it is difficult to identify events in a formal fashion as we have done so far – it is difficult to know the experiment, sample space and events, etc. Often we are interested in the probability of events that are simply described:

- What is the probability of getting a total of 10 when two dice are rolled?
- What is the probability that it will rain today? What is the probability of a downpour today?
- What is the probability that it will take me more than 34 minutes to reach UQ from my home?

For a given experiment, and resulting sample space $S$, suppose that we are able to assign a real number to each simple event. Such a real number is referred to as a *random variable.*

> **Definition (Random Variable):** Formally, a *random variable* is a measurable function defined on a probability space that maps from the sample space to the real numbers. There are two types of random variables – discrete and continuous.
>
> - A random variable is said to be *discrete* if it can take a finite or countable number of values.

> - A random variable is said to be *continuous* if it can assume all real values in some interval or intervals.

*Remark.* Upper case letters $(X, Y, \dots)$ are used to denote random variables, and lower case letters $(x, y, \dots)$ are used to denote fixed values a random variable can take. Lower case letters with subscripts $(x_i, y_i, \dots)$ are usually used to denote data points or observations which are usually collected using a random sampling design.

## 1.7 Functions of Random Variables

In order to be able to make statements about a random variable, it is necessary to know the following information:

- The range: finite or countably infinite for discrete variables or an interval for continuous variables.
- Probability information: necessary to make probabilistic statements regarding random variables. This is in the form of cumulative distribution functions (CDF) for both discrete and continuous random variables, probability mass functions (PMF) for discrete random variables, and probability density functions (PDF) for continuous random variables.

> **Definition (Cumulative Distribution Function):** The distribution function or the *cumulative* distribution function (CDF) of a random variable at a point $y$, denoted by $F_Y(y)$, is a real-valued function mapping from the sample space $S$ such that:
> $$F_Y(y) = P(Y \leq y),$$
> which shows the probability that the random variable $Y$ takes a value less than or equal to $y$.

### Properties of Distribution Functions

From the definition, the following properties of distribution functions can derived:

- $F_Y(y)$ is defined for all values of $y$ in the range $(-\infty, +\infty)$.
- Lower limit is zero. That is $\lim\limits_{y \to -\infty} F_Y(y) = 0$.
- Upper limit is 1. That is $\lim\limits_{y \to \infty} F_Y(y) = 1$.
- $F_Y(y)$ is non-decreasing.
- $F_Y(y)$ is continuous from the right. That is, $F_Y(y) = \lim\limits_{\Delta \to 0} F_Y(y + \Delta)$ exists.

### Continuity of Probability

> **Definition (Continuity of Probability):** If $\{A_j\}_{j=1}^{\infty}$ is a sequence of events, then we say that it increases to $A$ if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$ and $\bigcup\limits_{j=1}^{\infty} A_j = A$ and decreases to $A$ if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$ and $\bigcap\limits_{j=1}^{\infty} = A_j = A$.

> **Theorem (Continuity of Probability):** If $\{A_j\}_{j=1}^{\infty}$ is a sequence of events increasing to $A$ then:
>
> $$\lim_{j \to \infty} \text{Prob}(A_j) = \text{Prob}(A).$$
>
> If $\{A_j\}_{j=1}^{\infty}$ is a sequence of events decreasing to $A$ then:
>
> $$\lim_{j \to \infty} \text{Prob}(A_j) = \text{Prob}(A).$$

## 1.8 Probability Functions for Discrete Random Variables

Consider a discrete random variable $Y$, and let $y_1, y_2, \dots, y_k$ represent the values taken by the random variable. The probability function shows the probability that $Y$ takes a given value $y$. It is given by

$$p(y_i) = P(\{Y = y_1\}) = P(Y = y_i),$$

and zero at all other points. $p(y)$ is the sum of all probabilities attached to all the sample points in the sample space that are mapped to the value $y$. The relationship between probability (PDF) and distribution (CDF) functions is given by

$$F(y) = \sum_{\forall y_i \leq y} P[Y = y_i].$$

## 1.9 Density Functions for Continuous Random Variables

Let $Y$ be a continuous random variable taking values in the interval $(a, b)$. The density function of $Y$, denoted by $f_Y(y)$, is a non-negative function such that $f_Y(y)$ is zero for values outside the range $(a, b)$ and

$$P(Y \leq y) = F_Y(y) = \int_{\infty}^{y} f_Y(y)\ dy = \int_{a}^{y} f_Y(y)\ dy.$$

If the distribution function $F_Y(y)$ of $Y$ is differentiable, then the density function of $Y$ can be derived using the relationship

$$f_Y(y) = F_Y^{'}(y) = \frac{dF_Y(y)}{dy}.$$

**Note:**

1. The CDF and PDF are *equivalent* characterisations of a probability distribution or random variable.
2. CDFs exists and are well defined for *any* random variable; and
3. PDFs exists *only* for continuous random variables but may not exist for *all* continuous random variables – it requires differentiability of the CDF.

## 1.10 Jointly Distributed Random Variables

Consider the case where several random variables are defined in a given sample space:

- In the bivariate case we consider two random variables $X$ and $Y$.
- In the multivariate case, we have several $(n)$ random variables $X_1, X_2, \ldots, X_n$.

The *cumulative distribution function* (CDF) and *probability density function* (PDF) of each of the random variables are referred to as *marginal* distribution function and *marginal* densities.

In general the range for each of the random variables is specified. Where it is not specified, we simply use $(-\infty, +\infty)$ as the widest possible range. In the case of discrete random variables the range is given by the range of each of the random variables.

The distribution and density functions for bivariate/multivariate random variables are associated with the joint distribution of the random variables.

### Multivariate Discrete Random Variables

For simplicity, we will focus on the bivariate case where we have two random variables $(X, Y)$. The range of the bivariate random variables is given by

$$\{(x_1, x_2, \ldots, x_n) \ldots (y_1, y_2, \ldots, y_n)\}.$$

Here for simplicity both random variables assume a finite set of values. When we observe these two random variables we may observe all possible combinations of the values of $X$ and $Y$. Probability information on $\{X, Y\}$ is given by the probability function for all possible pairs of values of $X$ and $Y$.

### Multivariate Continuous Random Variables

For simplicity, we will focus on the bivariate case where we have two random variables $(X, Y)$. The range for each of the random variables needs to be specified – for example $(-\infty, \infty)$ or $[0, 1]$, etc.

- The joint CDF of $(X, Y)$ is defined as: $F_{XY}(x, y) = P(X \leq x, Y \leq y)$.

- If $F_{XY}(xy)$ is differentiable, then the joint density function (pdf) of the pair of random variables $(X, Y)$ is given by

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}.$$

- the CDF and pdf are related as:

$$F_{XY}(x,y) = P(X \leq x, Y \leq y) = \int\limits_{-\infty}^{y} \int\limits_{-\infty}^{x} f_{XY}(u,v)\ du\,dv, \text{ and}$$

$$\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f_{XY}(u,v)\ du\,dv = 1.$$

## 1.11 Marginal Distributions

Within the context of multivariate random variables, we would still be interested in the distribution of each of the random variables in the list. For example, in the bivariate case $(X, Y)$ with pdf $f_{XY}(x, y)$ we could be interested in the distribution of $X$ and its properties. Here we use the marginal distribution of $X$. The following theorem is useful in deriving the marginal distribution.

**Theorem:** Let $f_{XY}(x, y)$ be the joint density of $X$ and $Y$ and let $f_X(x)$ denote the marginal density function of $X$. Then,

$$f_X(x) = \int\limits_{-\infty}^{\infty} f_{XY}(x,y) dy.$$

**Definition (Independence – Bivariate case):** Two variables $X$ and $Y$ are said to be independently distributed if and only if their joint distribution function (and joint density if it exists) equals the product of marginal distributions (density functions),

$$\begin{aligned} F_{XY}(x,y) &= F_X(x) \cdot F_Y(y) \quad && \forall x, y \text{ or equivalently,} \\ f_{XY}(x,y) &= f_x(x) \cdot f_Y(y) \quad && \forall x, y \text{ if the pdf exists.} \end{aligned}$$

If in addition to being independent, random variables also have the exact same distribution, then they are called *independently and identically distributed* or i.i.d.

**Definition (Conditional Density Function):** Here we consider the density function of $X$ conditional on a given event that $(X, Y) \in S$ where $S$ is a subset of the plane and $P[(X, Y) \in S] > 0$, then the conditional density is given by:

$$f_{XY}(x, y \,|\, S) = \frac{f_{XY}(x,y)}{P[(X,Y) \in S]} \text{ for } (x,y) \in S \text{ and 0 otherwise.}$$

> A special case of this is the conditional density function for $y$ given that $X = x$. This is given by:
>
> $$f_{Y|X}(y\,|\,x) = \frac{f_{XY}(x, y)}{f_X(x)} \text{ provided } f_X(x) > 0.$$

### Multivariate Continuous Random Variables: General Case

Here we consider the general case with a vector of $n$ random variables:

$$\boldsymbol{X} = (X^1, X^2, \ldots, X^n).$$

- The joint CDF of $X$ at any given $x$ is given by:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{X}}(x^1, x^2, \ldots, x^n) = P[X^1 \leq x^1, X^2 \leq x^2, \ldots, X^n \leq x^n]$$

- The joint PDF of $X$ at a given $x$ is defined as:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{X}}(x^1, x^2, \ldots, x^n) = \frac{\partial^n F_{\boldsymbol{X}}(x^1, x^2, \ldots, x^n)}{\partial x^1 \partial x^2 \ldots \partial x^n},$$

  provided the derivatives exist.

- The marginal density of $X^j$ is given by:

$$f_{\boldsymbol{X}}(x^j) = \int \int \cdots \int f_{\boldsymbol{X}}(x^1, x^2, x^{j-1}, \ldots, x^n) dx^1 dx^2 dx^{j-1} dx^{j+1} \ldots dx^n.$$

- The vector of random variables $\boldsymbol{X} = (X^1, X^2, \ldots, X^n)$ are said to be independently distributed if the joint density is equal to the product of the marginal density functions.

## 1.12 Transformations of Random Variables

Suppose we have a random variable $X$ with a given density function $f_X(x)$. Suppose we are interested in a random variable $Y$ which is a function of $X$. That is, $Y = \phi(X)$. How do we find the distribution of $Y$? There are three methods that can be used:

1. Method of distribution functions
2. Method of transformations
3. Method of moment generating functions

### Distribution Function Method

Consider the general case of $n$ random variables: $X_1, X_2, \ldots, X_n$ with density function $f(x_1, x_2, \ldots, x_n)$. Let $Y = g(X_1, X_2, \ldots, X_n)$ be a function of interest. Then the *distribution function* approach involves the following steps:

- For a given $y$, a value of $Y$, find the region $R_y$ in $x_1, x_2, \ldots, x_n$ space such that $g(x_1, x_2, \ldots, x_n) \leq y$.

- Find the distribution function value at $y$:

$$F_Y(y) = P(Y \leq y) = \int\int \cdots \int_{\{x_1,x_2,\ldots,x_n\} \in R(y)} f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n.$$

- The density function of $Y$ can be obtained by differentiating $F_Y(y)$ with respect to $y$.

**Examples**

1. Let $Y$ be such that $f_Y(y) = 2y$ if $0 \leq y \leq 1$ and 0 otherwise. Derive the density function of $Z = 3Y - 1$.
2. Let $X, Y$ have density $f_{XY}(x, y) = xe^{-x(1+y)}$ if $x > 0,\ y > 0$ and 0 otherwise. Derive the density function of $Z = XY$.

**Transformation Method**

The following theorem provides the method used in deriving density functions of functions of random variables.

> **Theorem (Page 75, textbook):** Let $f(x)$ be the density function $X$ and let $Y = \phi(X)$ be the function of interest where $\phi$ is monotonic differentiable function. Then the density function $g(y)$ of $Y$ is given by:
>
> $$g(y) = f[\phi^{-1}(y)] \left| \frac{d\phi^{-1}}{dy} \right| \text{ where } \phi^{-1} \text{ is the inverse function of } \phi.$$

**Note:** Monotonicity of $\phi$ is critical. Monotonicity of $\phi$ ensures that there is a one-to-one mapping from $X$ to $Y$.This theorem is easily extends to the case where several $x_i$ values are mapped to $y$.

> **Theorem (Page 75, textbook):** Further to notation above, suppose the inverse of $y = \phi(x)$ is multivalued such that
>
> $$x_i = \psi_i(y) \quad i = 1, 2, \ldots, n_y$$
>
> where $n_y$ indicates that the number of values of $x$ can vary with $y$. Then the density function $g(y)$ is given by
>
> $$g(y) = \sum_{i=1}^{n_y} \frac{f[\psi_i(y)]}{|\phi'[\psi_i(y)]|} \text{ where } \phi' \text{ is the derivative of } \phi$$

### Linear Transformations of Bivariate Random Variables

> **Theorem (Page 76, Textbook):** Let $f_{X1,X2}(x_1, x_2)$ be the joint density of a bi-variate random variable $(X_1, X_2)$. Let us consider two random variables $(Y_1, Y_2)$ defined by the linear transformation:
>
> $$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 \\ Y_2 &= a_{21}X_1 + a_{22}X_2 \end{aligned}$$
>
> such that $a_{11}a_{22} - a_{12}a_{21} \neq 0$, so that the equations can be solved for $X_1$ and $X_2$ in terms of $(Y_1, Y_2)$. Let the solution be given by:
>
> $$\begin{aligned} X_1 &= b_{11}Y_1 + b_{12}Y_2 \\ X_2 &= b_{21}Y_1 + b_{22}Y_2 \end{aligned}$$
>
> Then the joint density is given by:
>
> $$f_{Y_1,Y_2}(y_1, y_2) = \frac{f_{X_1,X_2}(b_{11}y_1 + b_{12}y_2, b_{21}y_1 + b_{22}y_2)}{|a_{11}a_{22} - a_{12}a_{21}|}$$
>
> We note that: $|a_{11}a_{22} - a_{12}a_{21}| = |b_{11}b_{22} - b_{12}b_{21}|^{-1}$

### General Transformations of Bivariate Random Variables

In general, consider the random variables $(X, Y)$ with joint distribution $f_{XY}(x, y)$ and suppose that $U = U(X, Y)$ and $V = V(X, Y)$. First, we solve to obtain $X = X(U, V),\ Y = Y(U, V)$. Next, we compute the Jacobian:

$$\begin{pmatrix} \dfrac{\partial X(u,v)}{\partial u} & \dfrac{\partial X(u,v)}{\partial v} \\ \dfrac{\partial Y(u,v)}{\partial u} & \dfrac{\partial Y(u,v)}{\partial v} \end{pmatrix},$$

and its determinant,

$$|J(u,v)| = \left| \left(\frac{\partial X(u,v)}{\partial u}\right)\left(\frac{\partial Y(u,v)}{\partial v}\right) - \left(\frac{\partial X(u,v)}{\partial v}\right)\left(\frac{\partial Y(u,v)}{\partial u}\right) \right|.$$

Finally, we obtain

$$f_{UV}(u,v) = |J(u,v)| f_{XY}(X(u,v), Y(u,v)).$$

## 1.13 Expected Value

> **Definition (Expected Value – Continuous case):** Let $X$ be a continuous random variable with density function $f(x)$. Then, the *expected value* of $X$, denoted by $\mathbb{E}(X)$ is defined as:
>
> $$\mathbb{E}(X) = \int\limits_{-\infty}^{\infty} x\, dF_X(x) = \int\limits_{-\infty}^{\infty} x\, f_X(x)\ dx,$$
>
> if the density function exists and the integral is absolutely convergent. This condition for existence of mean is stated as the condition that:
>
> $$\int |x|\ dF_X(x) < \infty \text{ or } \int |x|\ f_X(x) d(x)$$

> **Definition (Expected Value – Discrete case):** Let the probability function that a discrete random variable $X$ takes a particular value $x_i$ is given by $P(X = x_i)$, then the *expected value* of $X$ is given by
>
> $$\mathbb{E}(X) = \sum_i x_i P(X = x_i) \text{ provided the sum is finite}$$

> **Definition (Expected Value – Mixture):** Let $X$ be a *mixture* random variable with CDF $F_X(x)$, such that $X$ takes discrete values $x_i$ with probability $p_i$, for $i = 1, 2, \dots, n$; and a continuum of values in an interval $[a, b]$ with a density function $f_X(x)$, then the *expected value* is given by:
>
> $$\mathbb{E}(X) = \int\limits_{-\infty}^{\infty} x f_X(x)\ dx = \sum_{i=1}^{n} x_i p_i + \int\limits_{a}^{b} x f_X(x)\ dx$$
>
> provided $\sum\limits_{i=1}^{n} p_i + \int\limits_{a}^{b} f_X(x)\, dx = 1$.

### Properties of Expectation

The following properties of expectation can be easily proven from the definition. Let $a$, $b$ and $c$ be constants and $X$ and $Y$ be some random variables, then:

1. $\mathbb{E}(a) = a$ but $\text{var}(a) \neq a$.
2. $\mathbb{E}(X \pm Y) = \mathbb{E}(X) \pm \mathbb{E}(Y)$.
3. $\mathbb{E}(X)$ is a constant. So $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$.
4. $\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$.

5. $\mathbb{E}(g(X)) = \int g(x)f_X(x)\ dx$, provided the integral exists.
6. $P(X \leq Y) = 1$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.
7. If $a \leq X \leq b$, then $a \leq \mathbb{E}(X) \leq b$.

**Other Measures of Location**

> **Definition (Median):** The *median*, $m$, of a random variable $X$ is defined as the solution to
>
> $$P(X \leq m) = 0.5.$$
>
> The median divides the population into two parts each having probability of 0.5.

> **Definition (Mode):** The *mode* of a random variable $X$ is defined as the solution to
>
> - $\max_{x_i} = \{P(X = x_i)\}$, if $X$ is a discrete random variable, and to
> - $\max_{x_i} = \{f_X(x)\}$, if $X$ is a continuous random variable.
> - If the distribution is symmetric and unimodal, then $\mathbb{E}(X)$ = median = mode.

**Note:** In general, $\mathbb{E}(g(X))$ is *not* equal to $g(\mathbb{E}(X))$. They are equal when $g$ is linear.

## 1.14 Important Inequalities

> **Lemma (Jensen's inequality):** Let $G$ be a continuous function and $X$ be a random variable whose distributions function is given by $F_X(x)$. Then,
>
> $$\mathbb{E}(G(X)) \geq G(\mathbb{E}(X)), \text{ if } G(x) \text{ is a convex function in } x,$$
>
> and
>
> $$\mathbb{E}(G(X)) \leq G(\mathbb{E}(X)), \text{ if } G(x) \text{ is a concave function in } x.$$

> **Definition (Markov's Inequality):** Let $X$ be any non-negative random variable and let $a > 0$. Then,
>
> $$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

> **Definition (Chebyshev's Inequality):** The general form of Chebyshev's inequality is given by
>
> $$P\left(|X - \mathbb{E}(X)| \geq a\right) \leq \frac{\text{var}(X)}{a^2}.$$

> In standard Introductory Statistics courses, this inequality is stated as, for $k > 0$,
>
> $$P\left(|X - \mathbb{E}(X)| \geq k\sigma\right) \leq \frac{1}{k^2}.$$
>
> This form is a special case when $a = k\sigma$.

**Mean of a Function of a Bivariate Random Variable**

> **Definition (Expected Value – Bivariate Case):** Let $(X, Y)$ be a *bivariate* random variable with cdf given by $F_{XY}(x, y)$ and let $g(X, Y)$ be an arbitrary function. Then, the *expected value* of $g(\cdot)$ is given by
>
> $$\begin{aligned}\mathbb{E}(g(X,Y)) &= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} g(x,y)\ d_x\, d_y\ F_{XY}(x,y), \\ &= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} g(x,y)\ g(x,y) f_{XY}(x,y)\ dx\, dy,\end{aligned}$$
>
> provided the density function exists.

> **Definition (Covariance):** The *covariance* between two random variables $X, Y$ is a measure of statistical linear dependency between $X$ and $Y$, given by
>
> $$\begin{aligned}\text{cov}[X,Y] &= \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))], \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y), \\ &= \sigma_{XY}.\end{aligned}$$

> **Definition (Cauchy-Schwarz Inequality):** For random variables $X, Y$ we have
>
> $$[\text{cov}(X,Y)]^2 \leq \text{var}(X) \cdot \text{var}(Y).$$

## 1.15 Correlation Between Random Variables

> **Definition (Correlation):** The *correlation* between two variables is a measure of the *linear dependency* between them. The correlation coefficient between two variables $(X, Y)$ is given by
>
> $$\rho_{XY} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

*Remark.* For any pair of random variables $(X, Y)$, $-1 \leq \rho_{XY} \leq 1$. If $\rho_{XY}$ is close to 1 then there is a perfect positive linear relationship between $(X, Y)$. Conversely, if $\rho_{XY}$ is close to -1, then there is a perfect negative correlation.

> **Definition (Linear Independence):** A pair of variables $(X, Y)$ are *linearly independent* if and only if $\mathrm{cov}(X, Y) = 0$, which is equivalent to $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

*Remark.* Independence implies linear independence, but the converse is not true.

### Conditional Mean and Variance of a Random Variable

Conditional expectation or conditional expected value (or simply conditional mean) is the *expected value* of a random variable with respect to the conditional probability of the random variable.

> **Definition (Conditional Expectation):** Let $f_{Y|X}(y \,|\, X = x)$ be the conditional density function of $Y$ given $X = x$. Then, the *conditional expectation* is defined as
>
> $$\mathbb{E}_{Y|X}(Y \,|\, X = X) = \int_{-\infty}^{\infty} y \; f_{Y|X}(y \,|\, x) \; dy.$$

> **Definition (Conditional Variance):** The *conditional variance* is defined by:
>
> $$\begin{aligned} \mathrm{var}_{\,Y|X}(Y \,|\, X = x) &= \mathbb{E}_{Y\,|\,X}((Y - \mathbb{E}(Y))^2 | X = x) \\ &= \mathbb{E}_{Y\,|\,X}(Y^2 \,|\, X = x) - \left[\mathbb{E}_{Y|X}(Y \,|\, X = x)\right]^2. \end{aligned}$$

> **Definition (Law of Iterated Expectation and Variance):** The law of iterated
>
> expectation: $\mathbb{E}_Y = \mathbb{E}_X \left[\mathbb{E}_{Y|X}(Y \,|\, X = x)\right]$.
>
> variance: $\mathrm{var}_{\,Y}(Y) = \mathbb{E}_X \left[\mathrm{var}_{\,Y\,|\,X}(Y \,|\, X = x)\right] + \mathrm{var}_{\,X}[\mathbb{E}_{Y|X}(Y \,|\, X = x)\right]$.
>
> These expressions can be used for general functions $g(X, Y)$.

### Higher moments of distributions of random variables

The mean and variance represent the first and second moments of the distribution of a given random variable. For many random variables, these two moments characterise the distribution. However, higher order moments may be necessary for other random variables. Useful insights can be obtained from higher order moments (skewness).

We consider two types of moments – moments around the origin and moments around the mean.

- The $r^{th}$ moment of $Y$ around the origin (zero) is defined as: $\mu_r = \mathbb{E}(Y^r)$ for $r = 1, 2, \ldots,$
- The $r^{th}$ moment of $Y$ around the mean, or the $r^{th}$ central moment is given by: $\mu_r = \mathbb{E}((Y - \mathbb{E}(Y))^r)$.

## 1.16 Moment Generating Functions

Moment generating functions (MGF) are used to generate higher order moments around zero without having to go through the process each time.

> **Definition (Moment Generating Function):** Let $X$ be a random variable with $F_X(x)$ as its distribution function (PDF), then the *moment generating function* (MGF) of $X$ is
> $$m_X(t) = \mathbb{E}(e^{tX})$$
> From the definition of expectations, we have:
> $$m_X(t) = \begin{cases} \int\limits_{-\infty}^{\infty} e^{tu} \ f_X(u) \ du \text{ if } X \text{ is continuous,} \\ \sum\limits_{X} e^{tx} P(X = x) \text{ if } X \text{ is discrete.} \end{cases}$$

We can expand this using a series expansion for the exponential term.

> **Theorem (Moss pg 105):** If $m_X(t)$ is the moment generating function of a random variable $X$ then:
> $$\left. \frac{d^r \, m_X(t)}{dt^r} \right|_{t=0} = \mu'_r$$
> which is moment $r$ around 0.

### Properties of Moment Generating Functions

> **Definition (MGF of Function of a Random Variable):** Let $G$ be a continuous single-valued function and $X$ be a random variable with distribution function

> $F_X(x)$, then the MGF of $G(X)$ is given by:
>
> $$m_{G(X)}(t) = \mathbb{E}\left[e^{tG(X)}\right] = \int\limits_{-\infty}^{\infty} e^{tG(X)}\, dF_X(x).$$
>
> This follows from the definition of MGF.

> **Definition (MGF of a Sum of Independent Random Variables):** Let $X_1, X_2, \ldots, X_n$ be $n$ independent random variables, each having MGF
>
> $$m_{X_1}(t),\ m_{X_2}(t), \ldots, m_{X_n}(t),$$
>
> respectively. Then, the MGF of the *sum* of these variables $U = X_1 + X_2 + \cdots + X_n$ is:
>
> $$m_U(t) = m_{X_1}(t) + m_{X_2}(t) + \cdots + m_{X_n}(t).$$

**Result 1.2.** *(Equivalence of Distributions and MGFs). Let $X$ and $Y$ be two random variables with MGFs $m_X(t)$ and $m_Y(t)$ respectively. Then $X$ and $Y$ have the same probability distributions if*

$$m_X(t) = m_Y(t)\ \forall\, t.$$

The MGF, if it exists, contains all the information about the moments of the associated distribution – it completely characterises the distribution.

### Moment Generating Functions of Bivariate Random Variables

Suppose that $X, Y$ are bivariate random variables. The bivariate moment generating function is:

$$m_{X,Y}(s,t) = \mathbb{E}(e^{sX+tY}) = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} e^{sx+ty} f_{X,Y}(x,y)\ dxdy$$

> **Theorem:** Let $a$ and $b$ be non-negative integers. Then,
>
> $$\mathbb{E}(X^aY^b) = \left.\frac{\partial^{a+b}\ m_{X,Y}(s,t)}{\partial s^a\ \partial t^b}\right|_{s=0,t=0}$$
>
> **Examples:**
>
> $$\mathbb{E}(X) = \left.\frac{\partial\, m_{X,Y}(s,t)}{\partial\, s}\right|_{s=0,t=0},\ \mathbb{E}(XY) = \left.\frac{\partial^2\, m_{X,Y}(s,t)}{\partial s\, \partial t}\right|_{s=0,t=0},$$
>
> $$\mathbb{E}(X^2) = \left.\frac{\partial^2\, m_{X,Y}(s,t)}{\partial^2 s}\right|_{s=0,t=0}.$$

# 2 Probability Distributions

## 2.1 Bernoulli Distribution

Consider an experiment with only two outcomes: {Success, Failure}. Further, let the probability of Success $= p$. Then probability of Failure is $1 - p = q$. Let $X$ be a random variable which takes value 1 for Success and 0 for Failure. This is a discrete random variable with:

$$P(X = 1) = p;\ P(X = 0) = 1 - p = q,$$
$$\mathbb{E}(X) = p,\ \text{var}(X) = pq,\ m_x(t) = \mathbb{E}(e^{tX}) = pe^t + q.$$

The commonly used binomial distribution relates to the Bernoulli distribution as it relates to the number of successes in $n$ independent trials.

## 2.2 Binomial Distribution

Let $Y = \textit{No. of successes}$ with $p$ as probability of success. Then the range of values $Y$ can take are $\{0, 1, 2, \ldots, n\}$. The probability function of $Y$ is given by

$$P(Y = y) = \binom{n}{y} p^y\ q^{n-y} \quad p, q \in [0, 1];\ y = 0, 1, 2, \ldots, n.$$

Where $\binom{n}{y} = \dfrac{n!}{y!(n-y)!}$ (Hint: $Y$ can be written as a sum).

## 2.3 Poisson Distribution

Let $Y$ be a random variable taking values $0, 1, 2, \ldots$ (example: no. of cars sold in a month). Then it is said to follow a Poisson distribution if its probability function is given by

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \quad \lambda > 0,\ y = 0, 1, 2, \ldots$$

We have

$$\mathbb{E}(Y) = \lambda,\ \text{var}(Y) = \lambda,\ m_Y(t) = e^{\lambda(e^t - 1)}.$$

It can be proven that Poisson distribution is the limit of the Binomial distribution as the number of trials $n$ goes to infinity.

## 2.4 Uniform Distribution

Consider a random variable $Y$ which can take values in the interval $[a, b] \subset \mathbb{R}$. $Y$ follows a *uniform distribution* if and only if the density function is given by:

$$f_Y(y) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq Y \leq b \\ 0, & \text{otherwise} \end{cases}$$

It is usually denoted as $Y \sim Uniform(a, b)$ or $Y \sim U(a, b)$. If $Y \sim Uniform(a, b)$, then $U = \frac{Y-a}{b-a} \sim Uniform(0, 1)$ is called a standardised uniform random variable.

## 2.5 Exponential Distribution

Poisson distribution relates to the random variable $X$: Number of occurrences of an event of interest in a unit length of time (e.g. day, week, year, etc.). Let $\lambda$ be the parameter of the Poisson distribution. Consider the random variable $Y$ = time between occurrence of two events. Then $Y$ is a continuous random variable in the range $[0, \infty)$. It follows an *exponential distribution* with parameter $\beta$. Its density and distribution function respectively are given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-\frac{1}{\beta} y} & \text{for } 0 \leq y < \infty \\ 0, & \text{otherwise} \end{cases}$$

$$F_Y(y) = \begin{cases} 0 & \text{for } y < 0 \\ 1 - e^{-\frac{y}{\beta}} & \text{for } 0 \leq y < \infty \end{cases}$$

It can be shown that $\beta = 1/\lambda$. We have

$$\mathbb{E}(Y) = \beta,\ \text{var}(Y) = \beta,\ m_Y(t) = \frac{1}{1 - \beta t} \text{ for all } t < 1/\beta.$$

We also note that exponential distribution belongs to the family of the Gamma distribution. It is Gamma with $\alpha = 1$. The Exponential distribution is closely related to Poisson distribution.

## 2.6 Normal Distribution

Let $X$ be a continuous random variable following a normal probability distribution. Then, its density function is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\} \text{ for} - \infty < x < \infty,\ -\infty < \mu < \infty,\ \sigma > 0$$

Let $Y$ be a random variable following normal distribution with parameters $\mu$ and $\sigma^2$, then

$$\mathbb{E}(Y) = \mu,\ \text{var}(Y) = \sigma^2,\ m_Y(t) = \exp\left( \mu t + \frac{t^2\sigma^2}{2} \right).$$

**Result 2.1.** *Using the MGF of the normal distribution, we can show that if $Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$, then $W = a + bY$ is normally distributed with mean $a + b\mu$ and variance $b^2\sigma^2$.*

### Standard Normal Distribution

Consider the special case of $W = a + bY$ where $Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$ with $a = \mu/\sigma$ and $b = -1/\sigma$. Then we have:

$$\mathbb{E}(W) = 0, \quad \text{and} \quad \text{var}(W) = 1$$

Given that this is a standardised random variable, this particular case is referred to as *standard normal random variable* and is usually denoted by $Z$. This result is often stated as:

$$\text{If } X \sim N(\mu, \sigma^2), \text{ then } Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Conversely,

$$\text{If } Z \sim N(0, 1), \text{ then } X = \mu + \sigma Z \sim N(\mu, \sigma).$$

These two results can be established using the transformation technique used in determining distribution of functions of random variables.

## 2.7 Gamma Distribution

A continuous random variable $Y$ follows a Gamma distribution with parameters $(\alpha, \beta)$ if its density function is given by

$$f_Y(y) = \begin{cases} \dfrac{y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right)}{\beta^\alpha \Gamma(\alpha)}, & 0 \leq y < \infty \\ 0 & \text{otherwise} \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} \exp(-u) \, du.$$

The first parameter, $\alpha$, is the shape parameter. The larger it is the less skewed it is. The second parameter, $\beta$, is the scale-parameter. It scales up and down. The following properties of the Gamma function are useful:

- $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$,
- $\Gamma(z + 1) = z\Gamma(z)$,
- $\Gamma(n) = (n - 1)!$, if $n$ is an integer.
- Let $Y$ follow a Gamma$(\alpha, \beta)$, then

$$\mathbb{E}(Y) = \alpha\beta, \quad \text{var}(Y) = \alpha\beta^2, \; m_Y(t) = (1 - \beta t)^{-\alpha} \text{ for } t < 1/\beta.$$

- If $Y$ follows Gamma$(\alpha, \beta)$, then for any $\delta > 0$

$$\delta Y \sim \Gamma(\alpha, \delta\beta).$$

- Let $Y_i$ $(i = 1, 2, \dots, n)$ are independently distributed Gamma random variables such that $Y_i \sim Gamma(\alpha_i, \beta)$ then $\sum_i Y_i \sim Gamma(\sum_i \alpha_i, \beta)$.

If $\alpha$ is an integer then the Gamma distribution is sometimes called the Erlang or Generalised chi-square distribution.

## 2.8 Chi-square Distribution

Chi-square distribution is a special case of Gamma distribution and it is very useful is statistics and econometrics.

A continuous random variable $Y$ follows a chi-square distribution with $d$ (a positive integer) degrees of freedom if and only if

$$Y \sim Gamma(\alpha, \beta) \text{ with } \alpha = \frac{d}{2} \text{ and } \beta = 2.$$

From the properties of the Gamma distribution, If $Y \sim \chi^2_d$, then

$$\mathbb{E}(Y) = d, \ \operatorname{var}(Y) = 2d, \ m_Y(t) = (1-2t)^{-d/2} \text{ for } t < 1/2.$$

As the degrees of freedom increase, the Chi-square distribution becomes more symmetric and it converges to a normal distribution as the degrees of freedom increase to infinity. Some useful results:

1. If $Y \sim \chi^2_d$ then for any $\delta > 0, \delta Y \sim Gamma(\frac{d}{2}, 2\delta)$.
2. If $Y \sim \chi^2_{d_i}$ for $i = 1, \ldots, k$ are independently distributed, then
$$\sum_i Y_i \sim \chi^2_d \text{ where } d = \sum_i d_i.$$
3. If $Z$ follows the standard normal and $Y = Z^2$ then $Y$ follows a $\chi^2_1$ distribution. (If $W \sim N(\mu, \sigma^2)$ then $(\frac{W-\mu}{\sigma})^2$ follows a $\chi^2_1$ distribution).

## 2.9 Student's T Distribution

The density function of Student's $t$-distribution is given by:

$$f(t) = \frac{\Gamma\left(\dfrac{d+1}{2}\right)}{\Gamma\left(\dfrac{d}{2}\right)\sqrt{\pi d}} \left(1 + \frac{t^2}{d}\right)^{-\frac{d+1}{2}} \quad \text{for } -\infty < t < \infty$$

$$\mathbb{E}(T) = 0 \text{ but exists only when } d > 1,$$
$$\operatorname{var}(T) = \frac{d}{(d-1)} \text{ but exists only when } d > 2.$$

The $t$-distribution has a symmetric density function. It has fatter tails than a standard normal distribution. '$d$' is known as the degrees of freedom. As $d$ increases to infinity $t$-distribution tends towards standard normal distribution.

Student's $t$-distribution arises frequently in testing significance of regression coefficients. This distribution is relevant when the ratio of standard normal and square root of Chi-square random variable is used.

**Result 2.2.** *Let $Z \sim N(0,1)$, $W \sim \chi^2_d$ be independent. Then:*

$$T = \frac{Z}{\sqrt{\dfrac{W}{d}}} \sim t_d$$

*(Proof by method of transformations)*

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean $\mu$ and variance $\sigma^2$. Let $s = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n-1}}$, then

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

follows a t-distribution with $n-1$ degrees of freedom.

## 2.10 Snedecor's F Distribution

Let $J_1$ and $J_2$ be two random variables that are independently distributed with chi-square distribution with parameters $d_1$ and $d_2$. Then, the ratio of these two random variables is distributed as an $F - distribution$. That is,

$$F = \frac{(J_1/d_1)}{(J_2/d_2)} \sim F_{(d_1,d_2)}.$$

Where $F_{(d_1,d_2)}$ is an F-distribution with $(d_1, d_2)$ degrees of freedom. Some notes on the $F$-distribution:

- "Degrees of freedom" are the only parameters for this distribution.
- Probabilities and critical values can be found in $F$-tables.
- As degrees of freedom increase, the distribution becomes more symmetric.

Some useful results for the $F$-distribution:

**Result 2.3.** *If $Y \sim F_{(d_1,d_2)}$, then $\frac{1}{Y} \sim F_{(d_2,d_1)}$.*

**Result 2.4.** *If $Y \sim t_d$, then $Y^2 \sim F_{(1,d)}$ and $Y^{-1} \sim F_{(d,1)}$.*

These two results are quite important in regression analysis. This connects the $t$-tests used for testing significance of single parameters in regression with the $F$-test used in testing hypotheses on several parameters.

## 2.11 Density Function for a Truncated Distribution

In econometrics, the truncated normal distribution is commonly used. The density function of a normal random variable truncated from the left at $c \in \mathbb{R}$ is given by:

$$f_Y(y \,|\, Y \geq c) = \frac{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)}{P(Y \geq c)} = \frac{\phi(z)/\sigma}{1 - \Phi\left(\frac{c-\mu}{\sigma}\right)},$$

where $c \leq y < \infty$ and 0 otherwise. A particular case of interest is the *half-normal distribution* which is a normal random variable with mean equal to zero ($\mu = 0$) truncated at $c = 0$. The density function is given by:

$$f_Y(y) = \frac{1/\sigma \; \phi\left(\frac{y}{\sigma}\right)}{1 - \Phi(0)} = \frac{2}{\sigma}\phi\left(\frac{y}{\sigma}\right),$$

for $0 \leq y < \infty$ and 0 otherwise. The half-normal distribution is denoted as $N^+(0, \sigma^2)$ or $|N(0, \sigma^2)|$.

## 2.12 Multivariate-Normal Distribution

> **Definition (Multivariate Normal Distribution):** Consider a vector of $J$ random variables with $X = (X_1, X_2, \ldots, X_J)' \in \mathbb{R}^J$. Then $X$ follows a multivariate normal distribution if at any $x = (x_1, x_2, \ldots, x_J)' \in \mathbb{R}^J$ with pdf
>
> $$f_X(x) = (2\pi)^{-J/2}|\Omega|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)'\Omega^{-1}(x-\mu)\right\}.$$
>
> The mean and the covariance matrix of $X$ are given by:
>
> $$\mu = (\mu_1\mu_2\ldots\mu_J)' = \mathbb{E}(X_1X_2\ldots X_J)',$$
> $$\Omega = \begin{pmatrix} \text{var}(X_1) & & \text{cov}(X_1, X_2) & \ldots & \text{cov}(X_1, X_J) \\ \vdots & \ddots & & & \\ \vdots & & & \ddots & \\ \text{cov}(X_J, X_1) & & \ldots & & \text{var}(X_J) \end{pmatrix} = \mathbb{E}(X-\mu)(X-\mu)'.$$
>
> We denote multivariate normal distributions as: $Z \sim N_J(\mu, \Omega)$.

**Result 2.5.** *Let $Z \sim N_J(\mu, \Omega)$ and let $\boldsymbol{A}$ be a $K \times J$ matrix of constants such that $K \leq J$. If the rows of $\boldsymbol{A}$ are linearly independent, then*

$$\boldsymbol{A}Z \sim N_K(\boldsymbol{A}\mu, \boldsymbol{A}\Omega\boldsymbol{A}').$$

### Parameters of a Distribution

Let $X$ be a random variable of interest. Properties of the random variable are usually determined by a few constants which are usually not known in practice. These unknown constants are referred to as parameters of the distribution.

Let $\boldsymbol{\theta}$ be a parameter (or a vector of parameters) of the distribution of $X$. Then the density function, distribution and all the moments are completely determined by $\boldsymbol{\theta}$.

**Example:**

| | | |
|---|---|---|
| $X \sim Bernoulli(p)$ | one parameter | $\theta = p$ |
| $X \sim Binomial(n, p)$ | two parameters | $\theta_1 = n$; $\theta_2 = p$ |
| $X \sim Normal(\mu, \sigma^2)$ | two parameters | $\theta_1 = \mu$; $\theta_2 = \sigma$ |

*Remark.* Typically, numerical values of the parameters are not known in advance. These are usually estimated using sample data collected from the population or from the distribution.

### Random Sample from a Population

Let $X$ be a random variable of interest with density function $f_X(x : \theta)$. Suppose we draw a random sample of size $n$ from $X$, represented by $\{X_1, X_2, \ldots, X_n\}$. Then for each $i$, $X_i$ has the same distribution as $X$. Further, $\{X_1, X_2, \ldots, X_n\}$ are independently distributed.

Therefore, a random sample is represented by $n$ random variables $\{X_1, X_2, \ldots, X_n\}$ are independently and identically distributed random variables (iid). The joint density function of $\{X_1, X_2, \ldots, X_n\}$ is then the product of the marginal density functions $X_i$.

### What is a Statistic?

A statistic is a real valued function of the sample and does not depend on any of the unknown parameters. Statistics as a discipline is devoted to study the nature and properties of *statistics*.

*Remark.* Each statistic is itself a random variable and, therefore, has a statistical distribution.

> **Definition (Sampling Distribution):** The *sampling distribution* of a statistic is the distribution of a given sample statistic for a given sample of size $n$. This is often referred to as the "small sample" distribution of the statistic.

> **Definition (Asymptotic Distribution):** The *asymptotic distribution* of a statistic is the distribution of the statistic when the sample size increases without bound.

The following results are extremely important.

**Result 2.6.** *Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a random variable with mean $\mu$ and variance $\sigma^2$. Then the sample mean and sample variance are random variables with the following properties:*

$$\mathbb{E}(\bar{X}) = \mu \text{ and } \operatorname{var}(\bar{X}) = \frac{\sigma^2}{n}, \text{ irrespective of the distribution of the r.v. } X.$$

$$\mathbb{E}(s^2) = \sigma^2$$

### Properties of Sampling Distributions

Sampling from a normal distribution: Suppose $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a random variable which has a normal distribution with mean $\mu$ and variance $\sigma^2$, then we can derive the distribution of a number of statistics of importance.

1. If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, then the sample mean $\bar{X}$ has the distribution $N(\mu, \sigma^2/n)$.

2. If $Z_1, Z_2, \ldots, Z_n$ be a random sample from a standard normal distribution (i.e., these are independently and identically distributed random variables), then show that

   (a) $\bar{Z} = \frac{\sum_{i=1}^{n} Z_i}{n}$ has a normal distribution with mean 0 and variance $1/n$.

   (b) $\bar{Z} = \sum\limits_{i=1}^{n}(Z_i - \bar{Z})^2$ are independently distributed.

   (c) $\sum\limits_{i=1}^{n}(Z_i - \bar{Z})^2$ has a Chi-square distribution with $n-1$ degrees of freedom.

## Proofs

The moment generating function of a normal random variable:

**Proof:**

$$
\begin{aligned}
m_Y(t) &= \mathbb{E}[e^{tY}] \\
&= \int_{-\infty}^{\infty} e^{ty} f_Y(y) dy, \\
&= \int_{-\infty}^{\infty} e^{ty} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) dy, \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[ty - \frac{1}{2}\left(\frac{y-\mu}{\sigma^2}\right)^2\right] dy, \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2\sigma^2}[(y-\mu)^2 - 2t\sigma^2(y-\mu) + (t\sigma^2)^2] + t\mu + \frac{1}{2}t^2\sigma^2\right\} dy, \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2\sigma^2}(y-\mu-t\sigma^2)^2\right] \cdot \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) dy, \\
&= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).
\end{aligned}
$$

The last step is because that for $X \sim N(\mu + t\sigma^2, \sigma^2)$, the integration of its density function over $\mathbb{R}$ is:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left[-\frac{1}{2\sigma^2}(x-\mu-t\sigma^2)^2\right] dx = 1.$$

□

**Lemma:** Let $\{x_1, x_2, \ldots, x_n\}$ be a random sample from a random variable with mean $\mu$ and variance $\sigma^2$. Then the sample mean and sample variance are random variables with the following properties:

1. $\mathbb{E}(\bar{X}) = \mu$ and $\mathrm{var}(\bar{X}) = \sigma^2/n$, irrespective of the distribution of the random variable $X$.

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

**Proof:**

$$\begin{aligned}
\mathbb{E}(\bar{X}) &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \;=\; \frac{1}{n}(n\mu) = \mu \\
\mathrm{var}(\bar{X}) &= \mathrm{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \;=\; \frac{1}{n^2}\,\mathrm{var}\left[\sum_{i=1}^{n} X_i\right] \\
\mathrm{var}(\bar{X}) &= \frac{1}{n^2}\left[\underbrace{\sum_{i=1}^{n}\mathrm{var}(X_i)}_{\text{var for each } X_i = \sigma^2} \;+\; \overbrace{\sum_{i\neq j}\mathrm{cov}(X_i, X_j)}^{= 0, \text{ since } X_i\text{'s are indpendent}}\right] \\
\mathrm{var}(\bar{X}) &= \frac{1}{n^2}(n\sigma^2) = \sigma^2/n \\
s^2 &= \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
s^2 &= \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\
s^2 &= \frac{1}{(n-1)}\left[\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right] \\
s^2 &= \frac{1}{(n-1)}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] \\
\mathbb{E}(X_i^2) &= \sigma^2 + \mu^2, \;\; \mathbb{E}(\bar{X}^2) = \sigma^2/n + \mu^2 \\
\therefore \quad \mathbb{E}(s^2) &= \frac{1}{(n-1)}\left[n\mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}^2)\right] \\
\mathbb{E}(X_i^2) &= \frac{n}{(n-1)}\;\left[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} + \mu^2\right] = \sigma^2.
\end{aligned}$$

□

# 3 Point Estimation

## 3.1 Inference on Parameters

Inference involves estimation of parameters (point and interval estimation – confidence intervals), and testing hypotheses on parameters.

We denote parameters by $\boldsymbol{\theta}$, which can be a scalar or a vector if there are two or more parameters. The parameter space is the set of all possible values a parameter or a vector of parameters can take.

**Example:**

1. In the case of a Binomial distribution, $p$ is the parameter of interest. Here, the parameter space is given by the interval $[0, 1]$.
2. In the case of a normal distribution, the parameters are $\mu$ and $\sigma^2$. Here, the parameter space is $(-\infty, \infty)$ for $\mu$ and $(0, \infty)$ for the parameter $\sigma^2$.

## 3.2 Estimation of Parameters

> **Definition (Estimator):** An estimator is any statistic or known function of a random sample drawn from the random variable of interest, whose values are used to estimate an unknown parameter $\theta$ or a function $\theta$.

By definition, an estimator of an unknown parameter $\theta$ is itself a random variable. Its value depends upon the particular sample drawn. An *estimate* of $\theta$ is a realised value of the *estimator* based on the actual sample drawn. We denote $\hat{\theta}$ as an estimator of $\theta$.

### Properties of Estimators

In assessing the merits of an estimator, we consider finite sample properties as well as asymptotic properties of the estimator as the sample size increases to infinity. Some importance finite sample properties are:

1. **Unbiasedness:** $\mathbb{E}(\hat{\theta}) = \theta$ i.e., the mean of the estimator is equal to the parameter of interest.
2. **Efficiency:** $\hat{\theta}$ has the smallest variance of any estimator.
3. Possessing a *known distribution* which is useful for the purpose of constructing confidence interval and for hypothesis testing.

Similarly for asymptotic or large samples:

1. **Asymptotic unbiasedness:** $\mathbb{E}\left(|\hat{\theta} - \theta|\right) \to 0$ as sample size $n \to \infty$ or $\lim\limits_{n\to\infty} \mathbb{E}(\hat{\theta}) = \theta$.
2. **Consistency:** $\hat{\theta} \xrightarrow{\text{p}} \theta$ as $n \to \infty$ (convergence in probability). This means that $\hat{\theta}$ gets close to $\theta$ as sample size increases.

3. **Asymptotic efficiency:** $Asy.Var(\hat{\theta}) \leq Asy.Var(\bar{\theta})$. Asymptotically $\hat{\theta}$ has smaller variance than any other estimator $\bar{\theta}$.

4. Convergence to a known distribution as sample size increases. In many cases the estimators have asymptotic normal distribution (asymptotic normality). In some cases, the asymptotic distributions can be Chi-square.

5. Fast rate of convergence in probability or distribution.

**Definition (Mean Squre Error):** When comparing unbiased estimators it is natural to use the variance. However, when comparing biased estimators, we may refer to the *mean squared error*. Its definition is as follows

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \qquad \text{(finite sample)}$$
$$\lim_{n\to\infty} MSE(\hat{\theta}) = \lim_{n\to\infty} \mathbb{E}\left[(\hat{\theta} - \theta)^2\ \right] \text{(asymptotic)}$$

**Lemma:**

$$MSE(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})$$

**Proof:**

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \text{var}(\hat{\theta} - \theta) + \mathbb{E}(\hat{\theta} - \theta)^2 \\ &= \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2 \end{aligned}$$

## 3.3 Estimation Strategies

There are a number of estimation strategies used in statistics and econometrics. A few other these are:

1. Method of Moments (MoM),
2. Generalised Method of Moments (GMM),
3. Maximum Likelihood Estimation (MLE),
4. Method of Least squares (OLS),
5. Bayesian estimation.

### Method of Moments

Suppose we are interested in estimating parameters, $\theta$, of a random variable with density function which can be expressed as $f_X(x; \theta)$.

1. The method of moments estimator first recognises that moments of the distribution depend upon the parameters $\theta$.

2. The parameters are estimated by equation the sample moments with the population moments and derive estimates of unknown parameters.

3. The main basis for the method is the fact that the sample moments converge towards the population moments as sample size increases.

4. The number of moments used is equal to the number of parameters to be estimated.

5. If the number of moment conditions are more than the number of parameters, we need to use GMM.

**Examples**

1. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a $Gamma(\alpha, \beta)$ distribution. Find MoM of $\alpha, \beta$.

2. Let $\{X_1, X_2, \ldots, X_m\}$ be a random sample from a $Poisson(\lambda)$ distribution. Find MoM of $\lambda$.

3. Let $\{X_1, X_2, \ldots, X_m\}$ be a random sample from a distribution with the following density function:

$$f_X(x : \theta) = \begin{cases} \theta e^{-\theta x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   Find the method of moments estimate of $\theta$.

### Method of Least Squares

The method of least squares is useful in cases where we can set up a regression model. We have seen in the bivariate normal distribution case, the conditional distribution of $Y \mid X = x$ has a mean which is a linear function of $X$.

Here we start with a simple case of a random variable $Y$ which has mean $= \mu$ and variance $= \sigma^2$. In this case we can set up a simple regression model:

$$Y_i = \mu + e_i \quad \text{where } (Y_1, Y_2, \ldots, Y_n) \text{ is a random sample from } Y.$$

We can establish the properties of the random disturbance term, $e_i$. The $e_i$'s have mean equal to zero with variance $\sigma^2$ and they are all independently distributed. In this case, the method of least squares minimises the following function with respect to the unknown $\mu$,

$$\min \sum_{i=1}^{n} (Y_i - \mu)^2 \text{ with respect to } \mu.$$

The solution to this minimisation obtained by equating the first derivative with respect to $\mu = 0$ is given by the OLS estimator of $\mu$:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \hat{Y}.$$

> **Definition (BLUE):** The best linear unbiased estimator of a parameter is the unbiased linear estimator (i.e. a linear function of the data) which attains the smallest possible variance.

### Simple Regression Model

Let us start with a simple regression model

$$y_i = \beta_1 + \beta_2 x_i + e_i \quad i = 1, 2, \ldots, n$$

with the following properties:

$$\mathbb{E}[e_i \mid x_i] = 0, \ \operatorname{var}[e_i \mid x_i] = \sigma^2, \ \mathbb{E}[e_i e_j] = 0 \text{ for } i \neq j. \tag{1}$$

Which, by iterated expectations imply:

$$\mathbb{E}[x_i e_i] = 0, \ \mathbb{E}[e_i] = 0, \ \operatorname{var}[e_i] = \sigma^2.$$

The parameters of interest are $(\beta_1, \beta_2, \sigma^2)$. The least squares method requires:

$$\arg\min \sum_{i=1}^{n} (Y_i - \beta_1 - \beta_2 x_i)^2 \text{ with respect to } \beta_1 \text{ and } \beta_2.$$

The solution to this minimisation obtained by equating the first derivative with respect to the unknown parameters to zero. The OLS estimator is given by:

$$b_2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \text{ and } b_1 = \bar{y} - b_2 \bar{x}.$$

**Equivalence of Least Squares and the Method of Moments**

We first note that the least squares estimator are solved from the normal equations

$$\sum_{i=1}^{n}(y_i - b_1 - b_2x_i) = 0,$$
$$\sum_{i=1}^{n} x_i(y_i - b_1 - b_2x_i) = 0.$$

This means that the following sample moments are equal to zero:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - b_1 - b_2x_i) = 0,$$
$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - b_1 - b_2x_i) = 0.$$

The corresponding population moments are: $\mathbb{E}(e_i) = \mathbb{E}(e_ix_i) = 0$. These two equations imply that the population moments are zero. The method of moments would involve the corresponding sample moments also set to zero.

## 3.4 Maximum Likelihood Estimation

Suppose we are interested in estimating parameters, $\boldsymbol{\theta}$, of a random variable with density function which can be expressed as $f_X(x;\boldsymbol{\theta})$. Let $\{X_1, X_2, \ldots, X_n\}$ be a random iid sample from $X$. Consider the joint density function of $\{X_1, X_2, \ldots, X_n\}$. This is the product of the individual density functions and it is a function of the vector of parameters $\boldsymbol{\theta}$. The joint density is usually written as:

$$L(\boldsymbol{\theta} \,|\, x_1, x_2, \ldots, x_n) = f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta}) = \prod_{i=1}^{n} f_{X_i}(x_i \,|\, \boldsymbol{\theta}).$$

This function is referred to as the *likelihood function.*

Intuitively, the likelihood function is related to the probability of observing the data, given the accepted assumption that the data are drawn from the distribution. We wish to find the value of $\boldsymbol{\theta}$ that maximises the *likelihood* of observing the sample data.

We can derive the MLE by evaluating the likelihood function at different values of $\boldsymbol{\theta}$, and pick that value that maximises the likelihood. The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is given by:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \,|\, x_1, x_2, \ldots, x_n).$$

Usually, it is easier to maximise the log-likelihood function,

$$\mathcal{L}(\boldsymbol{\theta} \,|\, x) = \log\ \mathcal{L}(\boldsymbol{\theta} \,|\, x) = \log \prod_{i=1}^{n} f_{X_i}(x_i \,|\, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{X_i}(x_i \,|\, \boldsymbol{\theta})$$

- In many cases solution to the maximisation of the likelihood function can be found analytically.
- In some cases, it may not be possible to find a closed form solution to the optimisation problem.
- In many cases, solutions may be found via numerical optimisation.

We also note that it is standard practice to use the log of the likelihood function,

$$\ln \mathcal{L}(\boldsymbol{\theta} \mid y) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{\theta}).$$

This is for two main reasons. First, taking logs makes the function linear and it is easier to handle linear functions. Second, the log likelihood function has important information theoretic interpretation. For example, the expected value of the matrix of second order derivatives is known as the *information matrix.* Some comments about maximum likelihood estimation:

1. To obtain ML estimators we need to make an assumption about the distribution of the random variable involved.
2. However, such an assumption is not needed when we consider *least squares estimation* of the regression model.
3. However, ML estimation is a powerful technique and the resulting estimators have very useful asymptotic properties.
4. There are several econometric problems where the likelihood plays a valuable role. Examples are the limited dependent variable models such as logit, probit or count data models.

**Examples**

1. Derive the MLE of $p$ of the Bernoulli random variable based on a sample size of $n$.
2. Derive the MLE of $\mu$ and $\sigma^2$ based on the sample of size $n$ from a normal distribution with these parameters

### Maximum Likelihood Estimators

Prior to the estimation of parameters, it is important to make sure that all the parameters are identified.

> **Definition:** A parameter or parameter vector $\boldsymbol{\theta}$ is said to be identified if for any other parameter vector $\boldsymbol{\theta}^* \in \Theta$ (parameter space), then $\mathcal{L}(\boldsymbol{\theta}^*|y) \neq \mathcal{L}(\boldsymbol{\theta}|y)$. If $\boldsymbol{\theta}$ is identified, then it is estimable.

The *method of maximum likelihood* involves finding a value for $\theta$ that maximises

$$\mathcal{L}(\theta \,|\, y), \text{ or equivalently, } \ln \mathcal{L}(\theta \,|\, y).$$

The first order condition of maximisation if there is only one parameter is:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta} \,|\, y)}{\partial \boldsymbol{\theta}} = 0$$

If we have several parameters, say $\{\theta_1, \theta_2, \ldots, \theta_n\}$, then we have the following first order conditions:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\theta} \,|\, y)}{\partial \theta_1} = 0; \; \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta} \,|\, y)}{\partial \theta_2} = 0; \; \ldots; \frac{\partial \ln \mathcal{L}(\boldsymbol{\theta} \,|\, y)}{\partial \theta_k} = 0;$$

The second order condition in the case of a single parameter is given by:

$$\frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\theta} \,|\, y)}{\partial \boldsymbol{\theta}^2} < 0$$

In the case of several parameters $\{\theta_1, \theta_2, \ldots, \theta_k\}$, the second order condition is given by:

$$\frac{\partial^2 \ln \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{bmatrix} \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_1^2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_1 \partial \theta_3} & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \\ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_k \partial \theta_1} & \cdots & \cdots & \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_k^2} \end{bmatrix},$$

which is a negative semi-definite matrix.

### Properties of Maximum Likelihood Estimators

The following theorem highlights the most important properties of MLE.

**Theorem (Properties of MLE):** Suppose $\theta_0$ is the true value of the unknown parameter. Under a set of *regularity conditions*, $\hat{\theta}$, the *maximum likelihood estimator* of $\theta$ has the following asymptotic properties.

1. Consistency: $\hat{\theta} \xrightarrow{p} \theta_0$
2. Asymptotic normality: $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, nI(\theta_0)^{-1})$, where
$$I(\theta_0) = -\mathbb{E}\left[\frac{\partial^2 \ln \mathcal{L}}{\partial\theta\partial\theta'}\right].$$
3. Asymptotic efficiency: MLE $\hat{\theta}$ is asymptotically efficient which means that for any other consistent estimator $\theta^*$, we have: $Asy.Var(\hat{\theta}) \leq Asy.Var(\theta^*)$ or $[Asy.Var(\hat{\theta}) - Asy.Var(\theta^*)]$ is a negative semi-definite matrix.

**Definition (Regularity Conditions):** The regularity conditions are as follows:

1. The first three derivatives of $\ln f(y_i \,|\, \theta)$ exist and are continuous.
2. The following expectations:
$$\mathbb{E}\left[\frac{\partial \ln f(y_i \,|\, \theta)}{\partial\theta}\right] \text{ and } \mathbb{E}\left[\frac{\partial^2 \ln f(y_i \,|\, \theta)}{\partial\theta\partial\theta'}\right] \text{ exist.}$$
3. Third order derivatives of the log likelihood function are well behaved.

The consequences of the regularity conditions:

1. First we note that the following three entities are random variables:
$$\ln f(y_i \,|\, \theta);\; g_i(\theta) = \frac{\partial \ln f(y_i \,|\, \theta)}{\partial\theta};\text{ and } H_i = \frac{\partial^2 \ln f(y_i \,|\, \theta)}{\partial\theta\partial\theta'}$$
2. We have $\mathbb{E}[g_i(\theta)|_{\theta=\theta_0}] = 0$.
3. Further, we have $\text{var}[g_i(\theta)]_{\theta=\theta_0} = -\mathbb{E}[H_i(\theta)]_{\theta=\theta_0}$.

## 3.5 Maximum Likelihood - Optimisation

We note that the first order conditions for the maximisation of the log-likelihood function: $\partial \ln \mathcal{L}(\boldsymbol{\theta} \,|\, y)/\partial\boldsymbol{\theta} = 0$ can be highly non-linear in practice. This means that it may not have closed form or analytical solutions, and in practice numerical optimisation methods are used. The Newton-Raphson method is one of the standard methods used.

In the single parameter case, let the first order condition $\partial \ln \mathcal{L}(\theta \,|\, y)/\partial\theta = g(\theta) = 0$. Then, the following iterative scheme may be used:

1. Start with an initial guess of the solution, $\theta_0$.

2. Then follow the following scheme until it converges: $\theta_{n+1} = \theta_n - g(\theta_n)/g'(\theta_n)$

If the log likelihood function is approximated by a second order Taylor series expansion, then we can use the following iterative scheme which starts with an initial guess for the unknown parameter vector, $\theta_0$. Then the scheme is given by:

$$\theta_{n+1} = \theta_n - \left[ \left. \frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial\theta\partial\theta'} \right|_{\theta=\theta_n} \right]^{-1} \left. \left( \frac{\partial \ln \mathcal{L}(\theta)}{\partial\theta} \right) \right|_{\theta=\theta_0}$$

Most computer programs offer computer routines for optimisation. Things to watch out for are:

1. Local maximum vs global maximum.
2. Lack of solution if the likelihood function is flat in the parameter space where you start your process.
3. It is useful to start with meaningful initials values and preferably those obtained in earlier studies.

### Proofs

**Theorem (Law of the Unconscious Statistician (Continuous case)):** Let $X$ be a continuous random variable and $g$ denote a continuous function defined on the range of $X$. Then, if

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

**Proof:** We prove this result for the special case in which $g$ is differentiable with $g\prime(x) > 0$ for all $x$ in the range of $X$. In this case $g$ is strictly increasing and it therefore has an inverse function $g^{-1}$ mapping onto the range of $X$ and which is also differentiable with derivative given by:

$$\frac{d}{dy}\left[g^{-1}(y)\right] = \frac{1}{g'(x)} = \frac{1}{g'(g^{-1}(y))},$$

where we have set $y = g(x)$ for all $x$ is the range of $X$, or $Y = g(X)$. Assume also that the values of $X$ range from $-\infty$ to $\infty$ and those of $Y$ also range from $-\infty$ to $\infty$. Thus, using the Change of Variables Theorem, we have that

$$\int_{-\infty}^{\infty} g(x) f_X(x)\ dx = \int_{-\infty}^{\infty} y f_X(g^{-1}(y)) \cdot \frac{1}{g'(g^{-1}(y))}\ dy,$$

since $x = g^{-1}(y)$ and therefore

$$dx = \frac{d}{dy}[g^{-1}(y)]\ dy = \frac{1}{g\prime(g^{-1}(y))}\ dy.$$

On the other hand,

$$\begin{aligned} F_Y(y) &= P(Y \leq y), \\ &= P(g(X) \leq y), \\ &= P(X \leq g^{-1}(y)), \\ &= F_X(g^{-1}(y)), \end{aligned}$$

from which we obtain, by the Chain Rule, that

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{g\prime(g^{-1}(y))}$$

Consequently,

$$\int\limits_{-\infty}^{\infty} y f_Y(y)\ dy = \int\limits_{-\infty}^{\infty} g(x) f_X(x)\ dx,$$

or

$$\mathbb{E}(Y) = \int\limits_{-\infty}^{\infty} g(x) f_X(x)\ dx.$$

□

The law of the unconscious statistician also applies to functions of a discrete random variable. In this case we have:

**Theorem (Law of the Unconscious Statistician (Discrete case)):** Let $X$ be a discrete random variable with probability mass function $p_x$, and let $g$ denote a function defined on the range of $X$. Then, if

$$\sum_x |g(x)| p_x < \infty,$$

$$\mathbb{E}(g(X)) = \sum_x g(x) p_x(x).$$

**Proof:** Similar to above. □

**Theorem (BLUE):** Given that $\{Y_1, Y_2, \ldots, Y_n\}$ is a random sample from $Y$ with mean $\mu$ and variance $\sigma^2$, the OLS estimator is the *best linear unbiased estimator (BLUE)* of $\mu$.

**Proof:** We have $\hat{\mu} = \bar{Y}$. Consider a different estimator, $\tilde{\mu} = \sum_{i=1}^{n} a_i Y_i$. Then,

$$\mathbb{E}(\tilde{\mu}) = \left( \sum_{i=1}^{n} a_i \right) \mu$$

$\therefore$ to be unbiased, $\sum_{i=1}^{n} a_i = 1$.

$$\begin{aligned}
\operatorname{var}(\tilde{\mu}) = \operatorname{var}\left( \sum_{i=1}^{n} a_i Y_i \right) &= \sum_{i=1}^{n} \operatorname{var}(a_i Y_i) + \overbrace{\sum_{i \neq j} \operatorname{cov}(a_i Y_i, a_j Y_j)}^{0} \\
&= \sum_{i=1}^{n} a_i^2 \operatorname{var}(Y_i) \\
&= \sigma^2 \left( \sum_{i=1}^{n} a_i^2 \right) \qquad (1) \\
\operatorname{var}(\hat{\mu}) &= \frac{\sigma^2}{n} \qquad (2)
\end{aligned}$$

Then,

$$\begin{aligned}
\sum_{i=1}^{n} \left( a_i - \frac{1}{n} \right)^2 &= \left( \sum_{i=1}^{n} a_i - \frac{2}{n} \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} \frac{1}{n^2} \right) \\
&= \sum_{i=1}^{n} a_i^2 - \frac{1}{n} \geq 0 \\
&\therefore \sum_{i=1}^{n} a_i^2 \geq \frac{1}{n} \\
&\sigma^2 \sum_{i=1}^{n} a_i^2 \geq \frac{\sigma^2}{n} \\
&\therefore \operatorname{var}(\tilde{\mu}) \geq \operatorname{var}(\hat{\mu}) \qquad (3)
\end{aligned}$$

□

# 4 Asymptotic Theory

## 4.1 Cramer-Rao Lower Bound

Here we consider the *Cramer-Rao Lower Bound* for the variance of unbiased estimators. Let $\hat{\theta}(x_1, x_2, \ldots, x_n)$ be an unbiased estimator of $\theta$. Then under regularity conditions on the likelihood function, we have:

$$\mathrm{var}(\hat{\theta}) \geq -\frac{1}{\mathbb{E}\left[\dfrac{\partial^2 \, log\, L}{\partial\, \theta^2}\right]}$$

In the case where we have a vector of parameters, then

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) - \left\{-\mathbb{E}\left[\frac{\partial^2 \, log\, L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]\right\}^{-1}$$

is a positive semi-definite matrix.

### Some Concepts from Asymptotic Theory

Here we go through a number of concepts currently used in studying the large sample properties of estimators in econometrics.

> **Definition (Limit of a Sequence):** Suppose we have a sequence of real numbers $\{c_n\}$, then the sequence is said to converge to a constant $c$ if:
>
> $$\text{for any } \varepsilon > 0, \; \exists N \text{ such that } |c_n - c| < \varepsilon, \; \text{for all } n \geq N, \; \lim_{n\to\infty} \{c_n\} = c.$$
>
> We look at convergence of a sequence of random variables. In this case we cannot use the limit as defined above.

> **Definition (Probability Limit):** Suppose we have a sequence of random variables $\{x_n\}$, then the sequence is said to *converge in probability to a constant c* if
>
> $$\lim_{n\to\infty} P[|x_n - c| > \varepsilon] = 0 \text{ for any } \varepsilon > 0.$$
>
> Usually it is difficult to check if a sequence of random variables converge to a constant. When a sequence of random variables $\{x_n\}$ converge in probability to a constant $c$, the we denote it as:
>
> $$p\lim\{x_n\} = c$$

> **Definition (Convergence in Quadratic Mean):** If for any given sample size $n$, $x_n$ has mean $\mu_n$ and variance $\sigma_n^2$ then $x_n$ is said to converge in *quadratic mean* If
>
> $$\lim_{n\to\infty} \{\mu_n\} = c, \text{ and } \lim_{n\to\infty} \{\sigma_n^2\} = 0$$

> If $x_n$ converges in quadratic mean to a constant $c$ then $p\lim\{x_n\} = c$

> **Definition (Consistency (estimation)):** Let $\hat{\theta}_n$ be an estimator of $\theta$. Then $\hat{\theta}_n$ is a consistent estimator if and only if
>
> $$p\lim\{\hat{\theta}_n\} = \theta.$$

**Result 4.1.** *For any function $g(x)$, if $\mathbb{E}[g(x)]$ and $\mathrm{var}[g(x)]$ are finite constants, then*

$$p\lim \frac{1}{n}\sum_{i=1}^{n} g_i(x_i) = \mathbb{E}[g(x)].$$

> **Theorem (Slutsky's theorem):** For any continuous function $g(x_n)$ that is not a function of $n$, then
>
> $$p\lim g(x_n) = g(p\lim x_n).$$
>
> This is a powerful result, and very useful in practice.

**Examples**

1. Consider a random sample $(X_1, X_2, \ldots, X_n)$ drawn from $N(\mu, \sigma^2)$. Show that the MLE of $\sigma^2$ is consistent.

2. Consider a random sample $(X_1, X_2, \ldots, X_n)$ drawn from $\mathrm{iid}(\mu, \sigma^2)$. Show that $p\lim\{\bar{X}_n\} = \mu$.

**Some Rules from Asymptotic Theory**

If $x_n$ and $y_n$ are random variables with: $p\lim x_n = c$, $p\lim y_n = d$, then

(i) $p\lim(x_n + y_n) = c + d$,

(ii) $p\lim(x_n \cdot y_n) = c \cdot d$,

(iii) $p\lim \dfrac{x_n}{y_n} = \dfrac{c}{d}$ if $d \neq 0$.

If $W_n$ is a matrix of random variables and if $p\lim W_n = \Omega$ then $p\lim W_n^{-1} = \Omega^{-1}$.

## 4.2 Convergence to a Random Variable

Until now we have considered convergence to a constant. Now we look at the concept of a sequence of random variables converging to another random variable. These concepts are used in deriving asymptotic distributions of random variables.

> **Definition (Convergence in Probability):** A sequence of random variables $x_n$ converges *in probability* to a random variable $x$ if
>
> $$\lim_{n\to\infty} P[|x_n - x| > \varepsilon] = 0 \text{ for any } \varepsilon > 0.$$

> **Definition (Almost Sure Convergence):** A sequence of random variables $x_n$ converges *almost surely* to a random variable $x$ if and only if
>
> $$\lim_{n\to\infty} P[|x_i - x| > \varepsilon \; \forall\, i \geq n] = 0 \; \forall\, \varepsilon > 0.$$
>
> Almost sure convergence implies convergence in probability.

> **Definition (Convergence in r-th Mean):** A sequence of random variables $x_n$ converges *in r-th mean* to a random variable $x$ if
>
> $$\lim_{n\to\infty} \mathbb{E}[|x_n - x|^r] = 0.$$
>
> If r = 2, then the convergence is in mean square.

**Result 4.2.** *Suppose that a sequence of a random variables $x_n$ converges in r-th mean and $\mathbb{E}[|x|^r]$ is finite. Then,*

$$\lim_{n\to\infty} \mathbb{E}[|x_n|^r] = \mathbb{E}[|x|^r].$$

> **Definition (Convergence in Distribution):** A sequence of random variables $x_n$ converges in distribution to the random variable $x$ if
>
> $$\lim_{no\to\infty} |F_n(x) - F(x)| = 0 \text{ at all continuous points of } F(x).$$
>
> In this case $F(x)$ is known as the limiting distribution and $x_n \xrightarrow{d} x$.

*Remark.* In order to show convergence in distribution, it suffices to show that:

$$\lim_{n\to\infty} MGF(x_n) = MGF(x)$$

Some useful results: If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{p} c$, then

(i) $x_n y_n \xrightarrow{d} c \cdot x$,

(ii) $x_n + y_n \xrightarrow{d} c + x$,

(iii) $x_n / y_n \xrightarrow{d} x/c$ if $c \neq 0$,

(iv) If $g(\cdot)$ is continuous, then $g(x_n) \xrightarrow{d} g(x)$.

## 4.3 Central Limit Theorems

There are many results useful for asymptotic theory. Here we look at some of the main results.

> **Theorem (Weak law of large numbers):** If $\{y_i : i = 1, 2, \ldots, n\}$ is a sample of random variables with
>
> $$\mathbb{E}(y_i) = \mu_i < \infty,\ \operatorname{var}(y_i) = \sigma_i^2\ \forall\, i\ (i = 1, 2, \ldots, n),$$
>
> such that
>
> $$\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right) < \infty,\ \text{and}\ \operatorname{cov}(y_i, y_j) = 0\ \forall\, i \neq j.$$
>
> Then,
>
> $$p\lim\left[\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right] = 0$$
>
> This result is known as the weak law of large numbers. The *strong* law of large numbers refers to *almost sure convergence* in the place of *convergence of probability.* The law of large numbers (LLN) is a powerful result:
>
> - It allows random variables to have different distributions, no matter how complicated they may be;
> - It requires no knowledge on the form of the distributions as long as mean and variance exists and as long as the covariance is zero
> - The sample mean of these random variables will converge in probability to the true mean – i.e. in large samples, the sample average will be very close numerically to the average of the true expected values of the random variables.

> **Theorem (Lindberg-Levy Central Limit Theorem):** Let $\{y_i : i = 1, 2, \ldots, n\}$ be a random sample independently and identically distributed random variables and let $\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$. Furthermore, assume that
>
> $$\mathbb{E}(y_i) = \mu\ \text{(finite mean)},\ \operatorname{var}(y_i) = \sigma^2 < \infty\ \forall\, i.$$
>
> Then,
>
> $$\sqrt{n}(\bar{y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)\ \text{or}\ \sqrt{n}\left(\frac{\bar{y}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0, 1)$$

## 4.4 Delta Method

The delta method method makes use of the Taylor's series expansion to provide an asymptotic distributions of functions of random variables. This can be stated for the univariate and multi-variate cases.

> **Theorem (Delta Method):** Let $\sqrt{n}(g_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ and let $G$ be a function that is continuously differentiable (at least once), with $G'(\mu) \neq 0$. Then,
>
> $$\sqrt{n}(G(g_n) - G(\mu)) \xrightarrow{d} N\left(0, \sigma^2 \cdot (G'(\mu))^2\right).$$

**Example:** Let $\hat{\sigma}^2$ be the MLE estimator of $\sigma^2$ for a normally distributed sample with known mean. Show that $\sqrt{n}(\hat{\sigma} - \sigma) \xrightarrow{d} N(0, \sigma^2/2)$.

## 4.5 Asymptotic Distribution

> **Definition (Asymptotic Distribution):** Let $\hat{\theta}$ be an estimator of $\theta$. If
>
> $$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma_\theta^2),$$
>
> then we say that the *asymptotic distribution* of $\hat{\theta}$ is
>
> $$\hat{\theta} \sim N(\theta, \sigma_\theta^2/n).$$
>
> The asymptotic distribution is used to approximate the distribution of $\hat{\theta}$ for purposes of confidence intervals/hypothesis tests.

**Example:** Consider the sample mean of iid $(\mu, \sigma^2)$ random variables $(X_1, \ldots, X_n)$ given by $\hat{\mu} = \bar{X}$. Then, by the LLCLT we have

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2),$$

and the asymptotic distribution of $\hat{\mu}$ is

$$\hat{\mu} \sim N(\mu, \sigma^2/n).$$

### Homework

1. Obtain the Cramer-Rao lower bound for the sample mean of a random sample from a normal distribution with known variance.
2. Obtain the Cramer-Rao lower bound for the sample proportion which is the estimator of population proportion of a Bernoulli distribution.
3. (Properties of Least Squares). Consider the regression model:

   $$\begin{aligned} y_i &= \beta x_i + \mu_i, \\ \mathbb{E}[u_i \,|\, x_i] &= 0, \\ \mathbb{E}[u_i^2 \,|\, x_i] &= \sigma^2. \end{aligned}$$

   Suppose that we have an iid sample $(y_i, x_i)_{i=1}^n$. Show that the least squares estimator is consistent and derive its asymptotic distribution.

# 5 Confidence Intervals and Hypothesis Testing

## 5.1 Confidence Intervals

**Definition (Confidence Interval):** Let $(X_1, X_2, \ldots, X_n)$ be a random sample from a density $f_X(x;\theta)$. Let $T_1 = t_1(X_1, X_2, \ldots, X_n)$ and $T_2 = t_2(X_1, X_2, \ldots, X_n)$ be two statistics satisfying $T_1 \leq T_2$ for which

$$P[T_1 \leq g(\theta) \leq T_2] = 1 - \alpha, \tag{2}$$

where $\alpha$ does not depend on $\theta$. Then, the random interval $(T_1, T_2)$ is called a $100(1-\alpha)$ percent confidence interval for $g(\theta)$.

- $(1-\alpha)$ is called the confidence coefficient;
- $T_1$ and $T_2$ are called the lower and upper confidence limits for $g(\theta)$;
- A value for $(t_1, t_2)$ of the random interval is called the value of the confidence interval.
- The interpretation is that if $(t_1, t_2)$ are constructed for repeated samples, $100(1-\alpha)$ *% of the intervals contain the true value* $g(\theta)$.
- If we are interested in a vector of parameters, then we talk about *confidence regions* instead of confidence intervals.

Confidence intervals are usually computed using the distribution of a random variable that is a function of the sample but does not depend on the unknown parameter vector, referred to as a *pivotal quantity* or *pivot.*

**Definition (Pivotal Quantity):** Let $(X_1, X_2, \ldots, X_n)$ be a random sample from the density $f_X(x:\theta)$. Let $Q = q(X_1, X_2, \ldots, X_n : \theta)$ be a random variable whose distribution does not depend on $\theta$. Then, $Q$ is a *pivotal quantity.*

Pivotal quantity method: If $Q = q(X_1, X_2, \ldots, X_n : \theta)$ is a pivotal quantity and has a probability density function, then for any given $\alpha$ in the interval $(0,1)$ there will exist $q_1$ and $q_2$ such that:

$$P[q_1 \leq Q \leq q_2] = 1 - \alpha$$

Then, using $q_1$ and $q_2$, we find suitable intervals for functions like $g(\theta)$.

If $\hat{\theta}$ satisfies a central limit theorem, we can use its asymptotic distribution to derive confidence intervals for $\hat{\theta}$. To do so we must have a consistent estimator of $\sigma_\theta^2$ (unless it is known), in which case we can use the asymptotic distribution,

$$\hat{\theta} \sim N(\theta, \frac{\hat{\sigma_\theta^2}}{n}),$$

so that $Q = \frac{\hat{\theta}-\theta}{\hat{\sigma_\theta}/\sqrt{n}} \sim N(0,1)$ is our pivotal quantity.

Since we have used an asymptotic approximation of the distribution of $\hat{\theta}$, we obtain an asymptotically valued confidence interval,

$$\lim_{n\to\infty} P[q_1 \leq Q \leq q_2] = 1 - \alpha.$$

*Remark.*

1. We generally use the sampling distributions of the estimators of the parameters of the distribution.
2. Sample distributions may not be tractable when sample sizes are too small. For small samples, usually we need to rely on the normal distribution.
3. When $n$ is large, we can use large sample theory and the central limit theorems to find asymptotic distributions of random variables which are then used in constructing confidence intervals.
4. The question is whether a *pivotal quantity exists.* If we are sampling from a continuous distribution function then a pivotal quantity exists.

**Examples**

1. Construct a 95% confidence interval for $\mu$ if $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$ and $\sigma^2$ is known.
2. Construct a 95% confidence interval for $\mu$ if $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.
3. Construct a 95% confidence interval for $\mu$ if $X_i \sim iid(\mu, \sigma^2)$ for $i = 1, \ldots, n$.
4. Construct a 95% confidence interval for $\sigma^2$ if $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

## 5.2 Testing Hypotheses on Parameters

So far we focused on the *estimation* of parameters as part of drawing inference from the sample. Now we turn to the problem of testing hypotheses on parameters. There are two types of hypotheses we may consider:

1. Hypotheses on the distributional form (e.g. is the data normally distributed?).
2. Hypotheses concerning the parameters.

In general, we start with a maintained hypothesis regarding the parameters and then based on the sample information, we either reject the maintained hypothesis (or not) and make some conclusions. Whenever we conduct hypothesis tests, we identify two types of hypotheses regarding the parameters:

- *Null Hypothesis* - $H_0$: maintained hypothesis. This hypothesis will be maintained unless there is convincing evidence against the hypothesis from the sample information.

- *Alternative Hypothesis* - $H_1$: alternative. This hypothesis will be the outcome when the evidence supports rejection of the null hypothesis in favour of the alternative hypothesis.

A test requires a clear statement of the null and the alternative hypotheses. A sample of size $n$ is represented by the vector $x = (X_1, X_2, \ldots, X_n)$ and it belongs to the *n-dimensional Euclidean space*, $\mathbb{R}^n$.

> **Definition (Hypothesis Test):** A *test* of the hypothesis $H_0$ mathematically means determining a subset of $R$ of $\mathbb{R}_n$ such that if an observed sample represented by $x = (X_1, X_2, \ldots, X_n)$ is the region $R$ then we reject the null hypothesis. If $x \in \bar{R}$ then we cannot reject $H_0$.

> **Definition (Critical Region):** The *critical region* of the test is the set $R$ in the tests procedure. If an observed sample belongs to the critical region then the null hypothesis is rejected.

> **Definition (Test Statistic):** The *test statistic* is a real valued function of the sample $x$, denoted by $T(x)$. We reject $H_0$ if $T(x) \in R$ where $R$ is now a subset of the real line, as $T(\cdot)$ represents a real-valued function.

> **Definition (Simple vs Composite Hypotheses):** A hypothesis is called *simple* if it specifies the values of all the parameters of a probability distribution. Otherwise, it is called *composite*.

We consider the problem of testing a *simple null hypothesis* versus a *simple alternative hypothesis*. Note that the test procedure will be used against the observed sample data. Given that a random sample $x = (X_1, X_2, \ldots, X_n)$ consists of $n$ independently and identically distributed random variables, the test statistic is also a random variable. This means that any test applied to data collected through random sampling can result in errors. We consider two types of errors – *Type I and Type II* errors.

## 5.3 Type I and Type II Errors

> **Definition (Type I Error):** A *Type I error* is the error of rejecting the null hypothesis when it is true.

> **Definition (Type II Error):** A *Type II error* is the error of accepting the null when it is false (i.e., when the alternative is true).

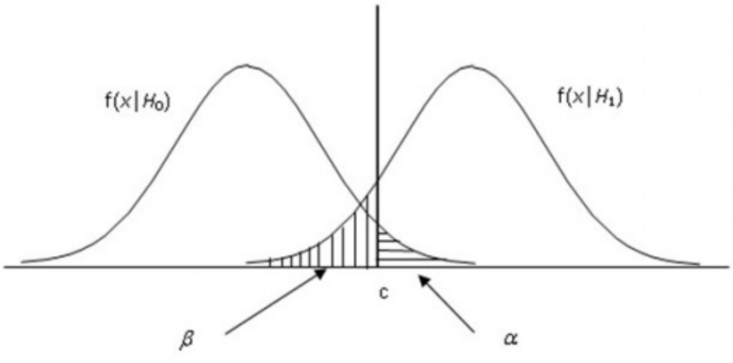
Associated with any given test procedure we have probabilities of both of these types We use the following notation:

$$\begin{aligned} P[\text{Rejecting } H_0 \text{ when it is true}] &= P[\text{Type I Error}] &&= \alpha \\ P[\text{Accepting } H_0 \text{ when it is false}] &= P[\text{Type II Error}] &&= \beta \end{aligned}$$

| | $\boldsymbol{H_0}$ **True** | $\boldsymbol{H_0}$ **False** |
|---|---|---|
| **Reject** $\boldsymbol{H_0}$ | Type I error ($\alpha$) | correct |
| **Accept** $\boldsymbol{H_0}$ | correct | Type II error ($\beta$) |

*Remark.* We can evaluate the probabilities of Type I and Type II errors by looking at the density function under the null and alternative hypotheses.

We consider: $f_X(x \mid H_0)$ and $f(x \mid H_1)$ and a test procedure: Reject $H_0$ if $x$ belongs to the rejection region: $R = \{x \mid x > c\}$ where $c$ is a real number.



> **Definition (Power):** The *power* of a test is the probability of rejecting a false null hypothesis under a fixed alternative. It is equal to 1 minus the probability of a Type II error $(1 - \beta)$.

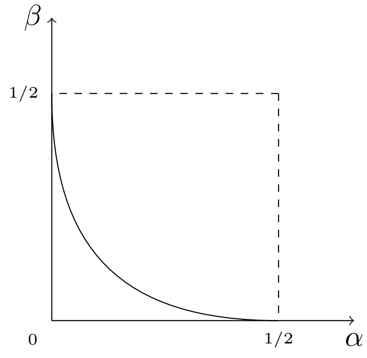
## 5.4 Power of Tests and Admissible Tests

> **Definition (Size vs Level):** $R$ is the most powerful test of *size* $\alpha$ if $\alpha(R) = \alpha$ and for any test of level $\alpha, \beta(R) \leq \beta(R_1)$ (note that it may not be unique). $R$ is the most powerful test of *level* $\alpha$ if $\alpha(R) \leq \alpha$ and for any test of level $\alpha$, $\beta(R) \leq \beta(R_1)$.

**Note:** It is impossible to minimise both $\alpha$ and $\beta$ at the same time.

**Result 5.1.** *The set of admissible characteristics plotted on the $(\alpha, \beta)$ plane is a continuous, monotonically decreasing, convex function which starts at a point within*

$[0,1]$ *on the* $\beta$ *axis and ends at a point within* $[0,1]$ *on the* $\alpha$ *axis.*



**Example:** Consider a test with a single observation with density

$$f_X(x) = \begin{cases} 1-\theta+x, & \text{if } \theta-1 \le x \le \theta \\ 1+\theta-x, & \text{if } \theta \le x \le \theta+1 \end{cases}$$

and $H_0 : \theta,\ H_1 = \theta = 1$. Compute the probabilities of type I and II errors for a test with rejection region $R = \{x > t\}$ where $t \in [0,1]$.

### 5.5 Neyman-Pearson Lemma

This is a famous result which allows us to identify the best test of a given size. Here the result is stated without proof.

**Lemma (Neyman-Pearson Lemma):** In testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, the best critical region of size $\alpha$ is given by:

$$R = \left\{ x \,\middle|\, \frac{L(x \,|\, H_1)}{L(x \,|\, H_0)} > c \right\},$$

where $L$ is the likelihood function and $c$ (the *critical value*) is determined so as to satisfy

$$P(R \,|\, \theta_0) = \alpha,$$

provided that such $c$ exists. Here, as well as in the following analysis, $\theta$ may be a vector.

**Example:** Suppose that $X \sim Bin(n,p)$ with known $n$. Derive the rejection region based on one observation for a test of $H_0 : p = p_0$ and $H_1 : p = p_1$.

### 5.6 Power Function

**Definition (Power Function):** If the distribution of the sample $X$ depends on a vector of parameters $\boldsymbol{\theta}$, we define the *power function* of a test $s$ based on the critical region $R$ as:

$$Q_s(\boldsymbol{\theta}) = P(X \in R \,|\, \boldsymbol{\theta}) \tag{3}$$

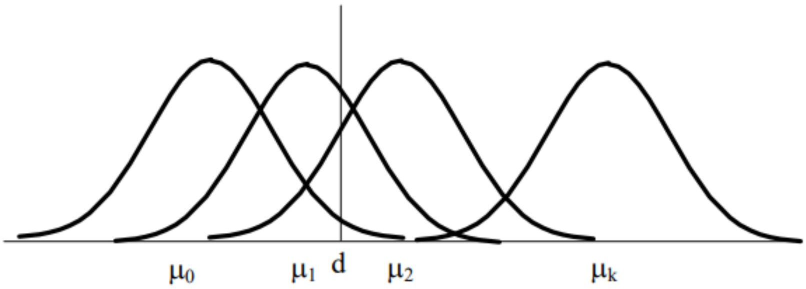
Therefore, the *power function* represents the probability of rejection of the null hypothesis when the parameter takes a value $\boldsymbol{\theta}$. For example, the size of the test $\alpha$ which is the probability of Type I error is:

$$\alpha = P[X \in R \,|\, H_0] = Q_s(\boldsymbol{\theta}_0)$$

Similarly,

$$Q_s(\boldsymbol{\theta}_1) = P[X \in R \,|\, H_1] = 1 - P[X \in \bar{R} \,|\, H_1] = 1 - \beta$$

**Example:** Suppose we want to test: $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. Consider the following test: Reject $H_0$ is $\bar{X} > d$. Then the power of the test can be seen from the image below:



As $\mu$ increases the area to the right of $d$ under $f$ increases. That is, the probability of correctly rejecting $H_0$ (i.e., the power of the test) increases.

### 5.7 Uniformly Most Powerful Test

**Definition (Uniformly Most Powerful (UMP) Test):** Let $Q_1(\theta)$ and $Q_2(\theta)$ be the power functions of two tests respectively. Then, we say that the first test is *uniformly more powerful* than the second test in testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$ if $Q_1(\theta_0) = Q_2(\theta_0) = \alpha$, and

$$Q_1(\theta) \geq Q_2(\theta) \; \forall \, \theta \in \Theta_1$$
$$Q_1(\theta) > Q_2(\theta) \text{ for at least one } \theta \in \Theta_1$$

A test with rejection region $R$ is the *uniformly most powerful* test of size (level) $\alpha$ if $P(R \,|\, \theta_0) = (\leq) \, \alpha$ and for any other test with rejection region $R_1$ such that $P(R_1 \,|\, \theta_0) = (\leq) \, \alpha$, we have

$$P(R \,|\, \theta) \geq P(R_1 \,|\, \theta) \text{ for any } \theta \in \Theta_1 \tag{4}$$

Note that the UMP may not always exist and it may not be unique.

### 5.8 Likelihood Ratio Test

The likelihood ratio test is a generalisation of the Neyman-Pearson test. It provides a UMP test when it exists. When a UMP test does not exist, the likelihood ratio test still has good asymptotic properties. In many cases, the likelihood ratio rest statistic has asymptotic distribution which is Chi square.

**Definition (Likelihood Ratio Test):** Let $L(x \,|\, \theta)$ be the likelihood function and let the null and alternative hypotheses be: $H_0 : \theta = \theta_0$ and $H_1 : \theta \in \Theta_1$ where $\Theta_1$ is a subset of the parameter space $\Theta$. The *likelihood ratio test* of $H_0$ against $H_1$ is defined by the critical region:

$$\Lambda = \frac{L(\theta_0)}{\sup\limits_{\theta_0 \cup \Theta_1} L(\theta)} < c \text{ where } c \text{ satisfies } P(\Lambda < c | \, H_0) = \alpha$$

> for a specified value of $\alpha$. It is easy to see that $\Lambda$ is in the range $(0, 1)$, so $c$ should be too.

**Example:** Consider a random sample from a normal distribution with unknown mean $\mu$ and known variance $\sigma^2$. Conduct a likelihood ratio test of $H_0 : \mu = \mu_0,\ H_1 : \mu > \mu_0$.

The likelihood ratio test can be generalised for cases where the null and alternative hypotheses are both composite. Suppose we wish to test: $H_0 : \theta \in \Theta_0,\ H_1 : \theta \in \Theta_1$ where $\Theta_0$ and $\Theta_1$ are both subsets of $\Theta$. Then, the *likelihood ratio test* of $H_0$ against $H_1$ is defined by the critical region:

$$\Lambda = \frac{\sup\limits_{\Theta_0} L(\theta_0)}{\sup\limits_{\Theta_0 \cup \Theta_1} L(\theta)} < c, \text{ where } c \text{ satisfies } P(\Lambda < c \,|\, H_0) = \alpha.$$

*Remark.* It may not always be possible to find $c$ required for the test as it involves the distribution of the statistics involved. In such cases, the following result is quite useful.

**Result 5.2.** *The likelihood ratio test-statistic, $\Lambda$, expressed as the ratio $LR_R/LR_U$ where $R$ and $U$ represent "restricted" and "unrestricted", verifies*

$$-2 \ln \Lambda = -2 \left[\ln(LR_R) - \ln(LR_U)\right] \xrightarrow{d} \chi^2_p$$

*under the null, where $p$ is the degrees of freedom usually representing the number of exact restrictions imposed by $H_0$. An alternative to the likelihood ratio test is the Wald Test. This can be used for testing a null hypothesis against two-sided alternative hypotheses.*

## 5.9 Wald Test

> **Definition (Wald Test):** Let $\boldsymbol{\theta}$ be a vector of parameters. Suppose we wish to test a set of linear or non-linear restrictions which are independent. Suppose we have
>
> $$H_0 : h(\boldsymbol{\theta}) = q \quad \text{versus} \quad H_1 : h(\boldsymbol{\theta}) \neq q \text{ where } q \text{ is a vector of known constants.}$$
>
> Let $\hat{\boldsymbol{\theta}}_n$ be an estimator of $\boldsymbol{\theta}$ obtained using a sample of size $n$ without imposing the restrictions of the null hypothesis. Then the *Wald test* simply checks if the value of the function $h(\boldsymbol{\theta})$ evaluated $\hat{\boldsymbol{\theta}}_n$ is different from $q$ in the null hypothesis. This is done using the Wald statistic.
>
> $$W = \left[h(\hat{\boldsymbol{\theta}}_n) - q\right]' \left(\text{var}\left[h(\hat{\boldsymbol{\theta}}_n)\right]\right)^{-1} \left[h(\hat{\boldsymbol{\theta}}_n) - q\right] \xrightarrow{d} \chi^2_M,$$
>
> where $\text{var}[h(\hat{\boldsymbol{\theta}}_n)]$ can be obtained using the *Delta method* and the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ and M is the number of restrictions (i.e. the size of $q$).

**Example:** Suppose that we have an estimator of a 2 dimensional vector $\boldsymbol{\theta}$ with asymptotic distribution: $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \Sigma)$. Let the realised value of $\hat{\boldsymbol{\theta}}$ in a random sample be

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \text{ and a consistent estimator } \hat{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Perform a Wald Test for

$$H_0 : \boldsymbol{\theta} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ H_1 : \boldsymbol{\theta} \neq \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$