

Middle East Technical University
Department of Economics
Econ 206:
Statistics for Economists-II

Güneş A. Aşık

December 29, 2025

1 Overview of Basic Concepts in Statistics

Basic probability theory

A *random process* (also called a **stochastic process**) is a process that can be repeated many times with different outcomes. It is a collection of random variables that evolve over-time, or some other index such as space. *The sample space* is the set of all possible outcomes of a random process. Random processes can be discrete or continuous.

A *discrete random variable* takes values from a *countable set* (finite or countably infinite). Formally, a random variable X is discrete if there exists a countable set:

$$\mathcal{X} = \{x_1, x_2, x_3, \dots\}$$

such that

$$\Pr(X = x_i) \geq 0, \quad \sum_i \Pr(X = x_i) = 1.$$

Examples:

- Number of heads in 5 coin tosses.
- Number of students present in a class.

- Roll of a fair die ($X \in \{1, 2, 3, 4, 5, 6\}$).

A *continuous random variable* can take values from an *uncountable set*, typically an interval of real numbers. Formally, X is continuous if there exists a probability density function (pdf) $f(x)$ such that

$$\Pr(a \leq X \leq b) = \int_a^b f(x) dx, \quad \text{with } f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x) dx = 1.$$

Examples:

- Housing prices.
- Time (in seconds) until a bus arrives.
- Daily return of a stock (often modeled as continuous).

Discrete random variables: probabilities are assigned to *individual values* whereas continuous random variables: probabilities are assigned to *intervals of values* via a probability density function.

The *expected value* or *mean* of a random variable X is defined as follows:

$$E(X) = \sum_x xp(x), \quad \text{if } X \text{ is discrete,}$$

and

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx, \quad \text{if } X \text{ is continuous,}$$

provided that the sum or the integral exists. Remember that the probability density of a continuous random variable requires that $\int_{-\infty}^{\infty} f(x) dx = 1$. So expectation is weighted average of all possible outcomes.

Expectation of a Constant: Consider a constant random variable, that is, a random variable with all its mass at a constant k . This is a discrete random variable with pmf $p(k) = 1$. By definition, we have

$$E(k) = k \cdot p(k) = k.$$

Example 1:

Let the random variable X of the discrete type have the probability mass function (pmf-a function that gives the probability that a discrete random variable is exactly equal to some value) given by the table:

x	1	2	3	4
$p(x)$	$\frac{4}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{2}{10}$

Here $p(x) = 0$ if x is not equal to one of the first four positive integers. We have

$$E(X) = (1)\left(\frac{4}{10}\right) + (2)\left(\frac{1}{10}\right) + (3)\left(\frac{3}{10}\right) + (4)\left(\frac{2}{10}\right).$$

Simplifying,

$$E(X) = \frac{23}{10} = 2.3.$$

Example 2:

Let the continuous random variable X have the pdf

$$f(x) = \begin{cases} 4x^3, & 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Then

$$E(X) = \int_0^1 x \cdot (4x^3) dx = \int_0^1 4x^4 dx = \left. \frac{4x^5}{5} \right|_0^1 = \frac{4}{5}.$$

Expectation of a Function of a Random Variable: Let X be a random variable and let $Y = g(X)$ for some function g .

1. **Continuous case:** Suppose X is continuous with pdf $f_X(x)$. If

$$\int_{-\infty}^{\infty} |g(x)| f_X(x) dx < \infty,$$

then the expectation of Y exists and is given by

$$E(Y) = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \quad (1.8.2)$$

2. **Discrete case:** Suppose X is discrete with pmf $p_X(x)$, and let the support of X be denoted by S_X . If

$$\sum_{x \in S_X} |g(x)| p_X(x) < \infty,$$

then the expectation of Y exists and is given by

$$E(Y) = \sum_{x \in S_X} g(x) p_X(x).$$

Example-3

Suppose X is the outcome of rolling a fair six-sided die, and let

$$g(X) = 2X^2 + 1.$$

Since each face is equally likely,

$$E(X) = \sum_{x=1}^6 x \cdot p(x) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5.$$

The expected value of $g(X)$ is

$$E(g(X)) = \sum_{x=1}^6 (2x^2 + 1) \cdot \frac{1}{6}.$$

Expanding:

$$E(g(X)) = (2 \cdot 1^2 + 1) \frac{1}{6} + (2 \cdot 2^2 + 1) \frac{1}{6} + \cdots + (2 \cdot 6^2 + 1) \frac{1}{6}.$$

Simplifying,

$$E(g(X)) = \frac{94}{3}.$$

Note that if a and b are constants, then

$$E(aX + b) = aE(X) + b.$$

Moments

The mean of a distribution is denoted by μ . In the case of a discrete random variable, the r th moment about the mean is (for $r = 0, 1, 2, \dots$)

$$\mu_r = E[(X - \mu)^r] = \sum_x (x - \mu)^r f(x).$$

The second central moment, μ_2 , is called the **variance** of the distribution and is denoted by σ^2 or $\text{Var}(X)$, where σ is the standard deviation:

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

This expression can be rewritten as

$$E[(X - \mu)^2] = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2 = E(X^2) - (E(X))^2.$$

If $Y = a + bX$, where a and b are constants, then the variance of Y is

$$\sigma_Y^2 = \text{Var}(a + bX) = b^2 \sigma_X^2,$$

so that the standard deviation of Y is

$$\sigma_Y = |b| \sigma_X.$$

Multivariate Distributions: Two random variables X and Y defined on the same probability space, the joint distribution for X and Y defines the probability of events in terms of both X and Y . In the case of only two random variables, this is called a *bivariate distribution*, but the concept generalizes to any number of random variables, giving a *multivariate distribution*.

If X and Y are discrete random variables, the joint probability distribution of X and Y is given by

$$f(x, y) = P(X = x \cap Y = y).$$

Product Moments: The r th and s th product moments of the random variables about the means (for $r, s = 0, 1, 2, \dots$) is

$$\mu_{r,s} = E[(X - \mu_X)^r (Y - \mu_Y)^s] = \sum_x \sum_y (x - \mu_X)^r (y - \mu_Y)^s f(x, y).$$

The product moment $\mu_{1,1}$ is called the covariance of X and Y , and it is denoted by $\sigma_{X,Y}$ or $\text{Cov}(X, Y)$. Thus,

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

Statistical Independence: If X and Y are independent, then

$$E(XY) = E(X)E(Y) \quad \text{and} \quad \text{Cov}(X, Y) = 0.$$

The correlation coefficient between X and Y is a standardized measure of their linear dependence, defined as

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{X,Y}}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}.$$

The correlation coefficient satisfies $-1 \leq \rho_{X,Y} \leq 1$. A value of $\rho_{X,Y} = 1$ indicates perfect positive linear dependence, $\rho_{X,Y} = -1$ indicates perfect negative linear dependence, and $\rho_{X,Y} = 0$ indicates no linear correlation between X and Y .

Moments of Linear Combination of Random Variables: If X and Y are random variables, then

$$\begin{aligned} E(X + Y) &= \mu_X + \mu_Y, & E(X - Y) &= \mu_X - \mu_Y, \\ \text{Var}(X + Y) &= \sigma_X^2 + \sigma_Y^2 + 2\text{Cov}(X, Y), \\ \text{Var}(X - Y) &= \sigma_X^2 + \sigma_Y^2 - 2\text{Cov}(X, Y). \end{aligned}$$

In the more general case, if $X_1, X_2, \dots, X_n$ are random variables, $a_1, a_2, \dots, a_n$ are constants, and

$$Y = \sum_{i=1}^n a_i X_i,$$

then

$$\begin{aligned} E(Y) &= \sum_{i=1}^n a_i E(X_i), \\ \text{Var}(Y) &= \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i < j} a_i a_j \text{Cov}(X_i, X_j). \end{aligned}$$

If $X_1, X_2, \dots, X_n$ are independent, then the covariance terms vanish, so the variance reduces to

$$\text{Var}(Y) = \sum_{i=1}^n a_i^2 \text{Var}(X_i).$$

Numerical Illustration: Two Independent Dice

Let X_1 and X_2 be independent outcomes of fair six-sided dice, so that for each i ,

$$P(X_i = x) = \frac{1}{6}, \quad x \in \{1, 2, 3, 4, 5, 6\}.$$

We have

$$E(X_i) = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5,$$

and

$$E(X_i^2) = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = \frac{91}{6},$$

so the variance of each die is

$$\text{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{35}{12}.$$

Define the linear combination

$$Y = 2X_1 - X_2.$$

Using linearity of expectation,

$$E(Y) = 2E(X_1) - E(X_2) = 2 \cdot 3.5 - 3.5 = 3.5.$$

Since X_1 and X_2 are independent, the covariance terms vanish, and

$$\text{Var}(Y) = 2^2 \text{Var}(X_1) + (-1)^2 \text{Var}(X_2) = 4 \cdot \frac{35}{12} + 1 \cdot \frac{35}{12} = 5 \cdot \frac{35}{12} = \frac{175}{12}.$$

Marginal Probability

Let X and Y be two discrete random variables with joint probability distribution $f_{X,Y}(x, y) = P(X = x, Y = y)$.

The **marginal probability distribution** of X is obtained by summing over all possible values of Y :

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

Similarly, the marginal probability distribution of Y is

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

For continuous random variables with joint density $f_{X,Y}(x, y)$, the marginal densities are obtained by integration:

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

Conditional Probability

The **conditional probability** of X given $Y = y$ is defined as

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad f_Y(y) > 0.$$

For continuous random variables, the conditional density of X given $Y = y$ is

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}, \quad f_Y(y) > 0.$$

Example-4: Joint and Marginal Distributions

Problem: From a group of 3 Republicans (R), 2 Democrats (D), and 1 Independent (I), a committee of 2 people is randomly selected. Let X be the number of Republicans and Y be the number of Democrats in the committee. Find the joint distribution of X and Y , and then find the marginal distribution of X .

Step 1: List all possible committees

The total number of ways to select 2 people from 6 is

$$\binom{6}{2} = 15.$$

Probability that 1 Republican and 1 Democrat is selected out of 6 is

$$\frac{\binom{3}{1}\binom{2}{1}\binom{1}{0}}{\binom{6}{2}} = \frac{6}{15}$$

We can list all possible combinations and categorize them by (X, Y) :

- $(X = 2, Y = 0)$: both members are Republicans. Number of ways: $\binom{3}{2} = 3$.
- $(X = 1, Y = 1)$: one Republican and one Democrat. Number of ways: $3 \cdot 2 = 6$.
- $(X = 1, Y = 0)$: one Republican and the Independent. Number of ways: $3 \cdot 1 = 3$.
- $(X = 0, Y = 2)$: both Democrats. Number of ways: $\binom{2}{2} = 1$.
- $(X = 0, Y = 1)$: one Democrat and the Independent. Number of ways: $2 \cdot 1 = 2$.
- $(X = 0, Y = 0)$: not possible (committee must have at least one person).

Step 2: Compute probabilities

Divide the counts by total number of committees (15):

$$P(X = 2, Y = 0) = \frac{3}{15} = 0.2$$

$$P(X = 1, Y = 1) = \frac{6}{15} = 0.4$$

$$P(X = 1, Y = 0) = \frac{3}{15} = 0.2$$

$$P(X = 0, Y = 2) = \frac{1}{15} \approx 0.0667$$

$$P(X = 0, Y = 1) = \frac{2}{15} \approx 0.1333$$

Step 3: Joint Probability Table

	$Y = 0$	$Y = 1$	$Y = 2$
$X = 0$	0	$2/15$	$1/15$
$X = 1$	$3/15$	$6/15$	0
$X = 2$	$3/15$	0	0

Step 4: Marginal Distribution of X

Sum across the rows:

$$P(X = 0) = P(0, 0) + P(0, 1) + P(0, 2) = 0 + \frac{2}{15} + \frac{1}{15} = \frac{3}{15} = 0.2$$

$$P(X = 1) = P(1, 0) + P(1, 1) + P(1, 2) = \frac{3}{15} + \frac{6}{15} + 0 = \frac{9}{15} = 0.6$$

$$P(X = 2) = P(2, 0) + P(2, 1) + P(2, 2) = \frac{3}{15} + 0 + 0 = \frac{3}{15} = 0.2$$

Answer: The marginal distribution of X is

X	$P(X)$
0	0.2
1	0.6
2	0.2

Conditional Expectation

The **conditional expectation** of X given $Y = y$ is defined as

$$E[X | Y = y] = \sum_x x P(X = x | Y = y), \quad (\text{discrete case})$$

$$E[X | Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x | y) dx, \quad (\text{continuous case}).$$

Conditional expectation is itself a random variable, often denoted as $E[X | Y]$, which assigns to each value of Y the expected value of X given that Y .

Law of Iterated Expectations

Let X and Y be random variables. Then

$$E[X] = E(E[X | Y]),$$

where $X | Y$ is the conditional probability of X given Y .

If Y takes on the outcomes $A_1, A_2, \dots, A_n$, the law can be written as

$$E[X] = \sum_{i=1}^n E[X | A_i] P(A_i).$$

This means that the expected value of X can be calculated from the probability distribution of $X | Y$ and Y .

Example-5: Law of Iterated Expectations

Suppose we want to compute the expected GPA of a student chosen at random from a class. Let X denote the GPA of a student, and let Y denote the gender of the student, where

$$Y = \begin{cases} \text{Boy,} & \text{if the student is male,} \\ \text{Girl,} & \text{if the student is female.} \end{cases}$$

Assume the following information is known:

$$E[X \mid Y = \text{Boy}] = 2.8, \quad E[X \mid Y = \text{Girl}] = 3.2.$$

Also suppose that in the class, 40% of the students are boys and 60% are girls. That is,

$$P(Y = \text{Boy}) = 0.4, \quad P(Y = \text{Girl}) = 0.6.$$

By the **Law of Iterated Expectations**,

$$E[X] = E(E[X \mid Y]).$$

Substituting the values,

$$E[X] = E[X \mid Y = \text{Boy}] \cdot P(Y = \text{Boy}) + E[X \mid Y = \text{Girl}] \cdot P(Y = \text{Girl}).$$

$$E[X] = (2.8)(0.4) + (3.2)(0.6) = 1.12 + 1.92 = 3.04.$$

Conclusion: The expected GPA of a randomly chosen student is

$$E[X] = 3.04.$$

2 Sampling Distributions and Central Limit Theorem

In a typical statistical problem, we have a random variable X of interest, but its pdf $f(x)$ or pmf $p(x)$ is not known. Our ignorance about $f(x)$ or $p(x)$ can roughly be classified in one of two ways:

1. $f(x)$ or $p(x)$ is completely unknown.
2. The form of $f(x)$ or $p(x)$ is known down to a parameter θ , where θ may be a vector.

In this process, our information about the unknown distribution of X or the unknown parameters of the distribution of X comes from a sample on X . The sample observations have the same distribution as X , and we denote them as the random variables $X_1, X_2, \dots, X_n$, where n denotes the sample size. When the sample is actually drawn, we use lower case letters $x_1, x_2, \dots, x_n$ as the values or realizations of the sample. Often we assume that the sample observations $X_1, X_2, \dots, X_n$ are also mutually independent, in which case we call the sample a random sample, which we now formally define:

Definition: If the random variables $X_1, X_2, \dots, X_n$ are independent and identically distributed (iid), then these random variables constitute a random sample of size n from the common distribution.

Definition: Independent and Identically Distributed (i.i.d.) Variables Let $\{X_i\}_{i=1}^n$ be a sequence of random variables defined on the same sample space. We say that $X_1, X_2, \dots, X_n$ are **independent and identically distributed (i.i.d.)** if:

1. **Independence:**

$$P(X_{i_1} \in A_1, X_{i_2} \in A_2, \dots, X_{i_k} \in A_k) = \prod_{j=1}^k P(X_{i_j} \in A_j)$$

for all measurable sets $A_1, \dots, A_k$ and any finite subset $\{i_1, \dots, i_k\}$.

2. **Identical Distribution:**

$$P(X_i \leq x) = F_X(x) \quad \text{for all } i,$$

where $F_X(x)$ is the common cumulative distribution function.

Intuition:

- The outcome of one draw provides no information about another.

- Each X_i is drawn from the same population distribution.

Example:

- Let X_i denote the height (in cm) of the i th person in a random sample.
- Random sampling $\Rightarrow$ i.i.d.:

$$X_1, X_2, \dots, X_n \sim F_X(x), \quad \text{i.i.d.}$$

Example: Toss a coin 10 times and write down the results into variables $A_1, \dots, A_{10}$.

- **Independent:** Each outcome A_i will not affect the other outcome A_j (for $i \neq j$ from 1 to 10), which means the variables $A_1, \dots, A_{10}$ are independent of each other.
- **Identically distributed:** Regardless of whether the coin is fair (with a probability of 1/2 for heads) or biased, as long as the same coin is used for each flip, the probability of getting heads remains consistent across all flips.

The i.i.d. assumption is the cornerstone of:

- The Law of Large Numbers (for convergence in probability)
- The Central Limit Theorem (for asymptotic normality)
- The consistency and efficiency of OLS and other estimators in econometrics.

Often, functions of the sample are used to summarize the information in a sample. These are called statistics, which we define as:

Definition Let $X_1, X_2, \dots, X_n$ denote a sample on a random variable X . Let $T = T(X_1, X_2, \dots, X_n)$ be a function of the sample. Then T is called a statistic.

Before moving to the Central Limit Theorem, let's remember the definitions of convergence in probability and convergence in distribution. Convergence in statistics formalizes what we suggest by a sequence of random variables $\{X_n\}$ getting “close” to another random variable X , as $n \rightarrow \infty$.

Definition: Convergence in Probability Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. We say that X_n converges in probability to X if, for all $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P[|X_n - X| \geq \epsilon] = 0,$$

or equivalently,

$$\lim_{n \rightarrow \infty} P[|X_n - X| < \epsilon] = 1.$$

If so, we write

$$X_n \xrightarrow{P} X.$$

If $X_n \xrightarrow{P} X$, we often say that the mass of the difference $X_n - X$ is converging to 0. In statistics, often the limiting random variable X is a constant; i.e., X is a degenerate random variable with all its mass at some constant a . In this case, we write $X_n \xrightarrow{P} a$. Also, for a sequence of real numbers $\{a_n\}$, $a_n \rightarrow a$ is equivalent to $a_n \xrightarrow{P} a$.

Theorem 1 (Chebyshev's Theorem). *For any random variable X with a finite mean μ and finite variance σ^2 ,*

$$\forall k > 1, \quad P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

or, equivalently,

$$\forall k > 1, \quad P(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

or, alternatively, by setting $k = \frac{\epsilon}{\sigma}$,

$$\forall \epsilon > 0, \quad P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

When σ^2 is known, one of k and ϵ can be arbitrarily chosen.

Chebyshev's theorem tells you that for any kind of data distribution (salary, height, test scores, etc.), the majority of the data points must lie within a few "standard deviations" of the average.

Theorem 2 ((Weak) Law of Large Numbers). *Let $\{X_n\}$ be a sequence of i.i.d. random variables with common mean μ and variance $\sigma^2 < \infty$. Let $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then:*

$$\forall \epsilon > 0, \quad \lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| \leq \epsilon) = 1$$

$$\bar{X}_n \xrightarrow{P} \mu.$$

Proof. The mean and variance of $\bar{X}_n$ are μ and σ^2/n , respectively. Hence, by Chebyshev's Theorem, we have for every $\varepsilon = \frac{k\sigma}{\sqrt{n}} > 0$,

$$P[|\bar{X}_n - \mu| \geq \varepsilon] = P\left[|\bar{X}_n - \mu| \geq \left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) \left(\frac{\sigma}{\sqrt{n}}\right)\right] \leq \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0.$$

□

This theorem says that all the mass of the distribution of $\bar{X}_n$ is converging to μ , as $n \rightarrow \infty$. In a sense, for n large, $\bar{X}_n$ is close to μ .

Theorem 3. Suppose $X_n \xrightarrow{P} X$ and a is a constant. Then

$$aX_n \xrightarrow{P} aX.$$

Theorem 4 (Continuous Mapping Theorem). Suppose $X_n \xrightarrow{P} a$ and the real function g is continuous at a . Then

$$g(X_n) \xrightarrow{P} g(a).$$

This theorem tells us that continuity preserves convergence in probability, and it gives us many useful results. For instance, if $X_n \xrightarrow{P} a$, then

$$\begin{aligned} X_n^2 &\xrightarrow{P} a^2 \\ 1/X_n &\xrightarrow{P} 1/a, \quad \text{provided } a \neq 0 \\ \sqrt{X_n} &\xrightarrow{P} \sqrt{a}, \quad \text{provided } a \geq 0. \end{aligned}$$

This theorem implies that if $\bar{X}_n \xrightarrow{P} \mu$, then $\bar{X}_n^2 \xrightarrow{P} \mu^2$. We will use this result in proving that the sample variance, S^2 is a consistent estimator for σ^2 .

So far, we introduced the concept of convergence in probability. With this concept, we can formally say, for instance, that a statistic converges to a parameter and, furthermore, in many situations we can show this without having to obtain the distribution function of the statistic. But how close is the statistic to the estimator? We will now try to answer this question.

Convergence in Distribution Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be, respectively, the cdfs of X_n and X . Let $C(F_X)$ denote the set of all points where F_X is continuous. We say that X_n converges in distribution to X if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x), \quad \text{for all } x \in C(F_X).$$

We denote this convergence by

$$X_n \xrightarrow{D} X.$$

This material on convergence in probability and in distribution comes under what statisticians and probabilists refer to as asymptotic theory. Often, we say that the distribution of X is the asymptotic distribution or the limiting distribution of the sequence $\{X_n\}$.

When X has a standard normal distribution, instead of writing $X_n \xrightarrow{D} X$, we will write $X_n \xrightarrow{D} N(0, 1)$. Here we need to be careful because the right-hand member of this last expression is a distribution and not a random variable. But as a short-cut, we will make use of this convention. When we say X_n has a limiting standard normal distribution, we mean that $X_n \xrightarrow{D} X$, where X has a standard normal distribution, or equivalently $X_n \xrightarrow{D} N(0, 1)$.

Relationship Between Modes of Convergence

Recall:

- **Convergence in probability:**

$$X_n \xrightarrow{p} X \quad \text{if} \quad \forall \epsilon > 0, \quad \lim_{n \rightarrow \infty} P(|X_n - X| > \epsilon) = 0$$

- **Convergence in distribution:**

$$X_n \xrightarrow{d} X \quad \text{if} \quad \lim_{n \rightarrow \infty} F_n(x) = F(x) \quad \text{for all } x \text{ where } F \text{ is continuous}$$

where F_n and F are the cumulative distribution functions of X_n and X respectively.

Theorem 5. *Convergence in probability implies convergence in distribution, but not conversely.*

1. **Convergence in probability $\Rightarrow$ Convergence in distribution:**

$$X_n \xrightarrow{p} X \quad \Rightarrow \quad X_n \xrightarrow{d} X$$

2. **Convergence in distribution $\Rightarrow$ Convergence in probability** only when the limit is constant:

$$X_n \xrightarrow{d} c \quad \Rightarrow \quad X_n \xrightarrow{p} c$$

where c is a constant.

Example in Class

Intuition: Convergence in probability is stronger: requires the actual random variables to get close to each other. Convergence in distribution is weaker: only requires the distribution functions to converge. The only case where they are equivalent is when converging to a constant.

Theorem 6 (Central Limit Theorem). *Let $X_1, \dots, X_n$ be i.i.d. random variables satisfying $\mathbb{E}(X_j) = \mu$, $\text{var}(X_j) = \sigma^2 < \infty$ and let $\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$. Then*

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

Another way to state the CLT is as follows: let Z be a standard normal random variable. For any $z \in \mathbb{R}$, we have

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z).$$

Informally, CLT states that for sufficiently large n , the random variable

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal distributed regardless of the distribution of X_i and exactly standard normal distributed if X_i are normally distributed. This theorem asserts the remarkable and important fact that if $X_1, X_2, \dots, X_n$ denote the observations of a random sample of size n from any distribution having finite variance $\sigma^2 > 0$ (and hence finite mean μ), then the random variable

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

converges in distribution to a random variable having a standard normal distribution. Thus, whenever the conditions of the theorem are satisfied, for large n the random variable $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ has an approximate normal distribution with mean zero and variance 1. It is then possible to use this approximate normal distribution to compute approximate probabilities concerning $\bar{X}$. This is a super powerful theorem that lays the foundations of hypothesis testing. CLT suggests that even if the underlying distribution from which the data is observed is not normally distributed, the distribution of $\bar{X}_n$, i.e. the sample means will converge to a standard normal distribution.

Example in class

Example in class

Example in Class

CLT Simulation in R

The R code that produces the following histograms will be posted in the METU class page.

Simulation Example-1: We have a population that is distributed exponentially with $\mu = 2$ and $\sigma^2 = 4$. As an exercise, we draw samples with size $n=2$, $n=5$, $n=10$, $n=30$, $n=100$ and $n=1000$, repeating the exercise 1000 times for each sample size.

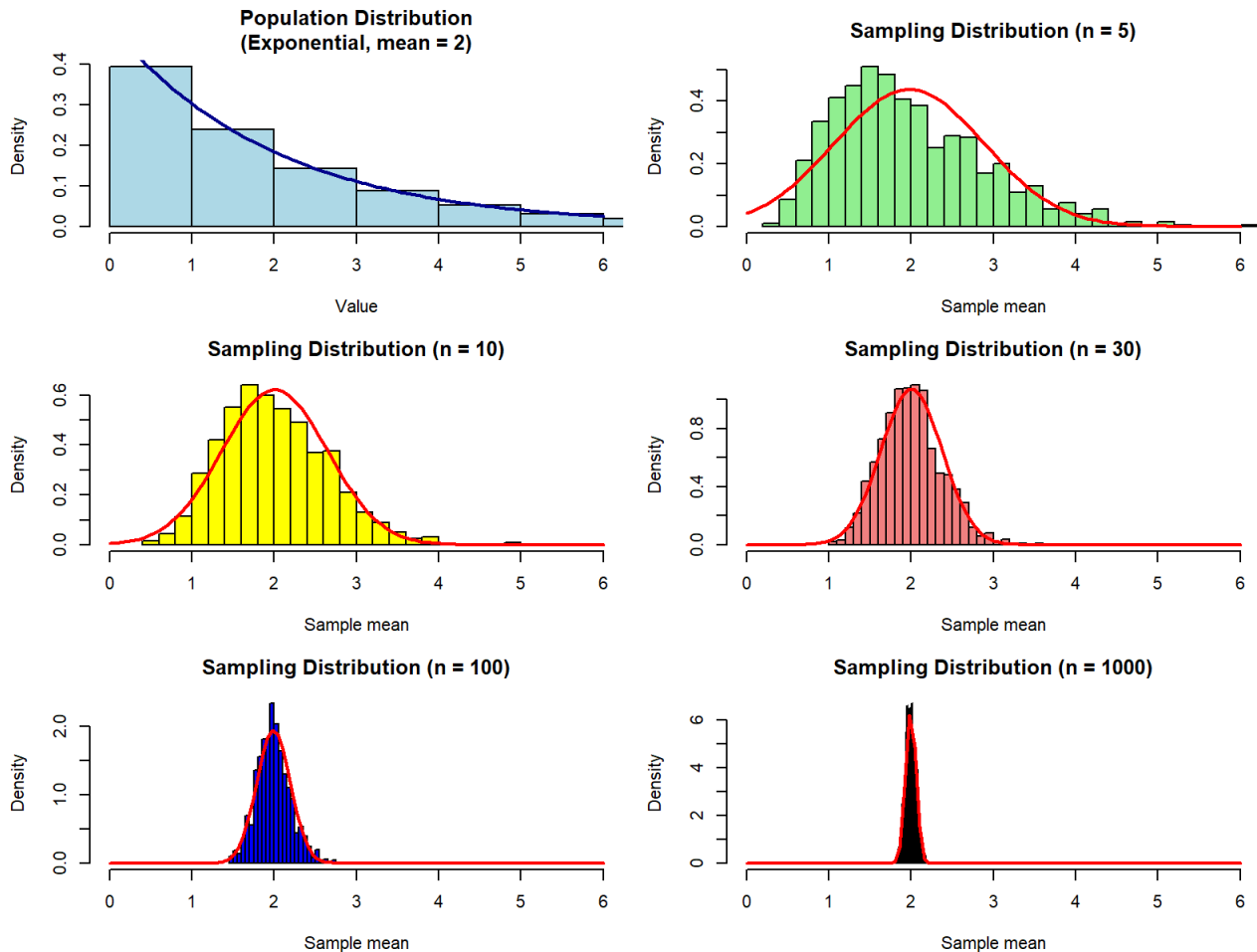


Figure 1: Central Limit Theorem: Sample means follow $\text{Normal}(\mu, \sigma^2/n)$ distribution when the population follows exponential distribution

Simulation Example-2: We have a population that is distributed uniformly between $[0,10]$. In that case $\mu = 5$ and $\sigma^2 = 8.333$. As before, we draw samples with size $n=2$, $n=5$, $n=10$, $n=30$, $n=100$ and $n=1000$, repeating the exercise 1000 times for each sample size.¹

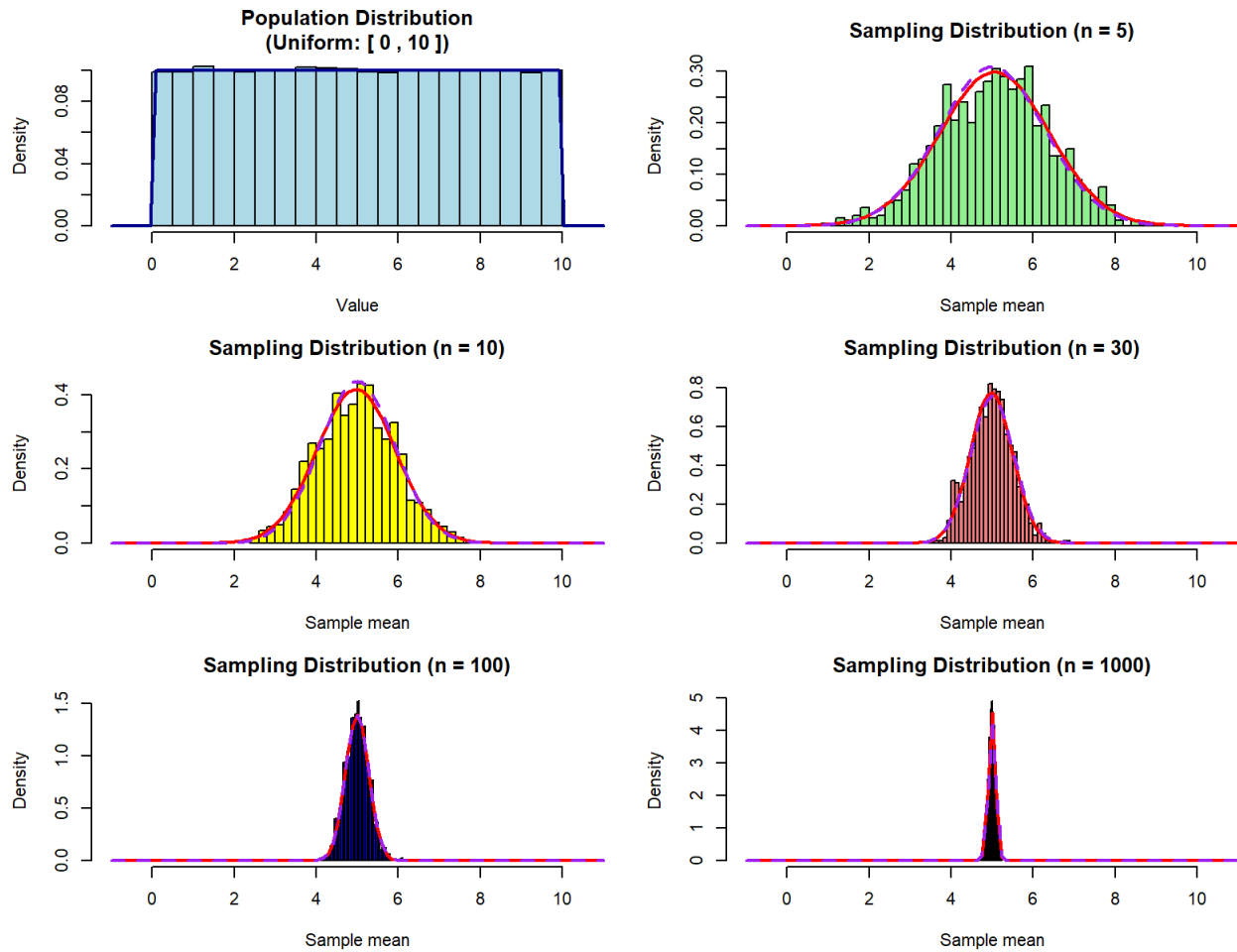


Figure 2: Central Limit Theorem: Sample means follow $\text{Normal}(\mu, \sigma^2/n)$ distribution when the population follows uniform distribution

The Central Limit Theorem is fundamental in econometrics as it provides the theoretical justification for statistical inference with economic data. Since most economic variables (income, prices, consumption) are not normally distributed in their raw form, CLT assures us that sample means will approach normality as sample sizes increase, enabling valid hypothesis testing, confidence intervals, and regression analysis. This allows econometricians to make reliable inferences about population parameters using standard normal and t-distributions, even when the underlying data generating processes are unknown or non-normal.

¹The population distribution looks a bit off because of binning in R.

3 Point Estimators

As we have highlighted many times in our class, most of the times our aim in economics is to estimate an unknown population parameter. All we know/have is a sequence of values which are coming out of the population. Formally, $\{x_i\}_{i=1}^N$ is the population and $\{x_i\}_{i=1}^n$ is our sample. Naturally the second one is a subset of the first.

Consider the situation where we have a random variable X whose pdf (or pmf) is written as $f(x; \theta)$ for an unknown parameter $\theta \in \Omega$. For example, the distribution of X is normal with unknown mean μ and variance σ^2 . Then $\theta = (\mu, \sigma^2)$ and $\Omega = \{\theta = (\mu, \sigma^2) : -\infty < \mu < \infty, \sigma > 0\}$.

Our information consists of a random sample $X_1, X_2, \dots, X_n$ on X ; i.e., $X_1, X_2, \dots, X_n$ are independent and identically distributed (iid) random variables with the common pdf $f(x; \theta)$, $\theta \in \Omega$. We say that T is a statistic if T is a function of the sample; i.e., $T = T(X_1, X_2, \dots, X_n)$. Here, we want to consider T as a point estimator of θ .

For example, if μ is the unknown mean of X , then we may use as our point estimator the sample mean $\bar{X} = n^{-1} \sum_{i=1}^n X_i$.

When the sample is drawn let $x_1, x_2, \dots, x_n$ denote the observed values of $X_1, X_2, \dots, X_n$. We call these values the realized values of the sample and call the realized statistic $t = t(x_1, x_2, \dots, x_n)$ a point estimate of θ .

Key distinctions to keep in mind:

- **Estimator:** $\hat{\theta}(\cdot)$ An estimator is a function, a formula, or a rule to estimate an unknown population parameter. For example $\bar{X} = n^{-1} \sum_{i=1}^n X_i$
- **Estimate:** The specific numerical value obtained by applying the estimator to sample data. In other words, it is the value that you get when you apply the estimator formula to your data. We use an *estimator* to obtain an *estimate*.
- **Estimation:** The overall process of using sample data to infer population parameters.

Since an estimator is a rule or formula to predict an unknown population parameter, we want our estimators to give us the right prediction on average. This means, we do not want our estimators to *overestimate* or *underestimate* the unknown parameter in a systematic way. This brings us to the statistical properties of **unbiasedness** and **consistency**.

Definition 1 (Unbiasedness). *The point estimator T for θ is unbiased if $E(T) = \theta$.*

Or, we can express unbiasedness as $E(\hat{\theta}) = \theta$. Generally (especially in econometrics), instead of using the notation T , we use the notation $\hat{\theta}$ as the estimator for θ . Both are fine, but T is a more general expression for denoting the statistic. And why do we call $\hat{\theta}$ point estimator? Unbiasedness is an individual property. $\hat{\theta}$ being an unbiased estimator means that when we repeat the estimation several times (using a different random sample of n observations each time), the average of the $\hat{\theta}$ s that we estimate are equal to the true θ .

Now, consider the following intuitive example. Suppose we don't know the average height of students in our class and want to estimate it. In this example, the population of interest is all the students enrolled in the class. Let's suppose the population mean is 1.70 cm in our class but since some students don't show up in classes, we do not know the mean for sure. Suppose that we randomly select 20 students to measure their height and calculate $\bar{X}$. In the first week, our sample gives us an estimate of 1.68 cm. In the second week, we get 1.73 cm. We repeat this sampling exercise until the end of term and record all the $\bar{X}$ s we collected. If this sample mean is an unbiased estimator, we would expect that the average of the sample means we collected throughout the term to be very close to 1.70 cm which is the population mean. Notice here that it does not have to be exactly 1.70. It is sufficient that this rule or formula ensures that theoretically $E(\bar{X}) = 1.70$. Or at least, we require that the mean of the $\bar{X}$ s gets closer and closer to 1.70 as we collect more and more samples. In contrast, consider another estimator; instead of sampling 20 students randomly without conditioning on gender, we collect samples only among boys. Since in the population, men tend to be taller than women, our sample mean would systematically overestimate the true height of students in our class. In other words, this estimator would be biased and the amount of bias would not disappear as we increase the sample size.

Notice that so far, we have been discussing a single prediction value for the unknown population parameter as opposed to estimating an interval, i.e. range of values for the true population θ . This is why we call this a point estimator. Later we will discuss predicting a range of values when we learn about confidence intervals.

Let's show now why the sample mean $\bar{X}$ and the sample variance S^2 are unbiased estimators of μ and σ^2 respectively.

Unbiasedness of the Sample Mean We want to show that $E(\bar{X}) = \mu$.

$$\begin{aligned}
 E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\
 &= \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) && \text{(Linearity of Expectation)} \\
 &= \frac{1}{n} \sum_{i=1}^n E(X_i) \\
 &= \frac{1}{n} \sum_{i=1}^n \mu && \text{(Since } E(X_i) = \mu \text{ for all } i) \\
 &= \frac{1}{n} (n\mu) \\
 &= \mu.
 \end{aligned}$$

Therefore, the sample mean $\bar{X}$ is an **unbiased estimator** of the population mean μ .

Unbiasedness of the Sample Variance (In Class)

Definition 2 (Efficiency). $\hat{\theta}_1$ and $\hat{\theta}_2$ being two estimators of θ , if

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$

then $\hat{\theta}_1$ is a more efficient estimator than $\hat{\theta}_2$.

Important note: Efficiency is a comparative property. We might have several alternative estimators for an unknown population parameter among the alternative unbiased estimators, we always want to choose the estimator with minimum variance. Because it gives us more precise (narrower) confidence intervals for our estimates (we will start learning about confidence intervals very soon).

Example 3.1.

Problem: Choose the best estimator for β from:

$$\begin{aligned}\hat{\beta}_1 &= \max\{X_1, X_2, \dots, X_n\} \\ \hat{\beta}_2 &= \frac{n+1}{n} \max\{X_1, X_2, \dots, X_n\} \\ \hat{\beta}_3 &= 2\bar{X}\end{aligned}$$

Given Properties:

Estimator	Expected Value	Variance
$\hat{\beta}_1$	$\frac{n}{n+1}\beta$	$\frac{n\beta^2}{(n+2)(n+1)^2}$
$\hat{\beta}_2$	β	$\frac{\beta^2}{n(n+2)}$
$\hat{\beta}_3$	β	$\frac{\beta^2}{3n}$

Both have similar variance behavior, but:

- $\hat{\beta}_1$ is biased downward ($E[\hat{\beta}_1] < \beta$)
- $\hat{\beta}_2$ is unbiased ($E[\hat{\beta}_2] = \beta$)
- Comparing coefficients: $\frac{n}{(n+2)(n+1)^2} < \frac{1}{n(n+2)}$ for all $n \geq 1$

So $\hat{\beta}_2$ has strictly smaller variance than $\hat{\beta}_1$.

- $\hat{\beta}_3$ has the largest variance
- $\hat{\beta}_1$ is biased and has larger variance than $\hat{\beta}_2$
- $\hat{\beta}_2$ is unbiased and has the smallest variance among the three

We should choose $\hat{\beta}_2$ because it is unbiased and has the smallest variance among all three estimators.

Definition 3 (Consistency). *Let X be a random variable with cdf $F(x, \theta)$, $\theta \in \Omega$. Let $X_1, \dots, X_n$ be a sample from the distribution of X and let T_n denote a statistic. We say T_n is a consistent estimator of θ if*

$$T_n \xrightarrow{P} \theta.$$

If $X_1, \dots, X_n$ is a random sample from a distribution with finite mean μ and variance σ^2 , then by the Weak Law of Large Numbers, the sample mean, $\bar{X}_n$, is a consistent estimator of μ .

Example 3.2.

Let $X_1, \dots, X_n$ denote a random sample from a distribution with mean μ and variance σ^2 . We now show that it is a consistent estimator of σ^2 . Recall that by the Law of Large Numbers, $\bar{X}_n \xrightarrow{P} \mu$. To show that the sample variance converges in probability to σ^2 , assume further that $E[X_1^4] < \infty$, so that $\text{Var}(S^2) < \infty$.

Using the preceding results, we can show the following:

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \\ &= \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) \\ &\xrightarrow{P} 1 \cdot [E(X_1^2) - \mu^2] = \sigma^2. \end{aligned}$$

Hence the sample variance is a consistent estimator of σ^2 . From the discussion above, we have immediately that $S_n \xrightarrow{P} \sigma$; that is, the sample standard deviation is a consistent estimator of the population standard deviation. Consistency is a very important property for an estimator to have. It is a poor estimator that does not approach its target as the sample size gets large. Note that the same cannot be said for the property of unbiasedness.

For example, instead of using the sample variance to estimate σ^2 , suppose we use $V = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Then V is consistent for σ^2 , but it is biased, because $E(V) = \frac{(n-1)\sigma^2}{n}$. Thus the bias of V is $-\sigma^2/n$, which vanishes as $n \rightarrow \infty$.

Example 3.3. (In Class)

Example 3.4. (In Class)

Example 3.5. (In Class)

A recap: the big picture

So far, we have learned about some desired properties of point estimators. We will keep in mind that we want to have unbiased estimators with minimum variance so that we have more precise estimates or predictions of the unknown population parameters. But when we can't have an unbiased estimator, we at least want an estimator that is consistent (i.e. converges in probability to the population parameter). This is why I recommend you try your best to understand the basic theorems, such as the Weak Law of Large Numbers, Chebyshev's Theorem, and the Central Limit Theorem which help us find unbiased or consistent estimators. You should also make sure that you are comfortable with taking expectations, and probability limits.

Now that we studied the properties of point estimators, we will now learn three techniques to yield point estimators of distributional parameters of populations or data generating processes. These are:

1. Method of Moment (MM) Technique
2. Maximum Likelihood (ML) Technique
3. Least Squares (LS) Technique

3.1 Moment Generating Functions (MGFs)

Definition

For a random variable X , the **moment generating function (MGF)** is defined as:

$$M_X(t) = E[e^{tX}],$$

provided that this expectation exists for t in some neighborhood around 0. The MGF can be used to obtain the moments of a random variable. By differentiating $M_X(t)$ and evaluating at $t = 0$, we get:

$$M_X^{(n)}(0) = E[X^n].$$

In particular:

$$M_X'(0) = E[X], \quad M_X''(0) = E[X^2], \quad \text{and so on.}$$

Where does this come from? We can differentiate inside the expectation operator for all t in an open interval around 0. The derivative of the exponential function is $\frac{d}{dt}e^{tX} = X e^{tX}$.

First Moment ($n = 1$):

$$M'_X(t) = \frac{d}{dt} E[e^{tX}] = E \left[\frac{d}{dt} e^{tX} \right] = E [X e^{tX}] .$$

$$\text{Evaluating at } t = 0: \quad M'_X(0) = E [X e^0] = E[X].$$

Second Moment ($n = 2$):

$$M''_X(t) = \frac{d}{dt} E [X e^{tX}] = E \left[\frac{d}{dt} (X e^{tX}) \right] = E [X^2 e^{tX}] .$$

$$\text{Evaluating at } t = 0: \quad M''_X(0) = E [X^2 e^0] = E[X^2].$$

Notice that since $\sigma^2 = E[X - E(X)]^2 = E[X^2] - [E(X)]^2$, we can write the theoretical moments as:

$$M'_X(0) = E(X) = \mu, \quad [E(X)]^2 = \mu^2$$

$$M''_X(0) = E [X^2] = \mu^2 + \sigma^2$$

General n -th Moment:

$$M_X^{(n)}(t) = E [X^n e^{tX}] .$$

$$\text{Evaluating at } t = 0: \quad M_X^{(n)}(0) = E [X^n] .$$

- For continuous X : $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$
- For discrete X : $M_X(t) = \sum_x e^{tx} P(X = x)$

The Intuition of MGFs, Why are MGFs Useful?

1. **To Uniquely Identify a Distribution.** If two random variables have the same MGF (and it exists in a neighborhood around 0), then they have the same distribution:

$$M_X(t) = M_Y(t) \implies X \stackrel{d}{=} Y.$$

Hence, the MGF completely characterizes the distribution.

2. **To Compute Moments Easily.** MGFs make it straightforward to find moments without directly computing integrals. For example, if $X \sim N(\mu, \sigma^2)$, then:

$$M_X(t) = \exp \left(\mu t + \frac{1}{2} \sigma^2 t^2 \right) ,$$

and

$$E[X] = M'_X(0) = \mu, \quad E[X^2] = M''_X(0) = \mu^2 + \sigma^2.$$

3. **To Analyze Sums of Random Variables.** If X and Y are independent, then:

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

This property makes MGFs extremely useful for studying the distribution of sums of independent random variables. For example:

- The sum of independent normal random variables is normal.
 - The sum of independent exponential random variables follows a gamma distribution.
 - The binomial distribution can be derived as the sum of Bernoulli random variables.
4. **To Prove the Central Limit Theorem (CLT).** In the proof of the CLT, MGFs (or characteristic functions) are used to show that the distribution of standardized sample means converges to a normal distribution. Specifically, the MGF of the standardized sum converges to:

$$M_{Z_n}(t) \rightarrow e^{t^2/2},$$

which is the MGF of a standard normal random variable $N(0, 1)$.

To find the limiting distribution function of a sequence of random variables $\{X_n\}$

$$F_{X_n}(x) \rightarrow F_X(x) \quad \text{at all continuity points of } F_X$$

obviously requires that we know $F_{X_n}(x)$ for each positive integer n . However, it is often difficult to obtain $F_{X_n}(x)$ in closed form.

Fortunately, if it exists, the moment generating function (mgf) that corresponds to the cumulative distribution function $F_{X_n}(x)$ often provides a convenient method of determining the limiting distribution.

Theorem 7 (Limiting Distributions via MGFs). *Let $\{X_n\}$ be a sequence of random variables with mgf $M_{X_n}(t)$ that exists for $-h < t < h$ for all n . Let X be a random variable with mgf $M_X(t)$, which exists for $|t| \leq h_1 \leq h$. If*

$$\lim_{n \rightarrow \infty} M_{X_n}(t) = M_X(t) \quad \text{for } |t| \leq h_1,$$

then

$$X_n \xrightarrow{D} X.$$

Example 3.6

Let $X_1, X_2, \dots, X_n$ be a random sample from a population with probability density function:

$$f(x; \theta) = \begin{cases} \theta e^{-\theta x}, & x > 0 \\ 0, & \text{elsewhere} \end{cases}$$

Find the method of moments estimator of θ , say $\hat{\theta}_1$.

Solution:

The moment generating function is:

$$\begin{aligned}
 M_X(t) &= E(e^{tX}) = \int_0^\infty e^{tx} \cdot \theta e^{-\theta x} dx \\
 &= \theta \int_0^\infty e^{-(\theta-t)x} dx \\
 &= \theta \left[\frac{-1}{\theta-t} e^{-(\theta-t)x} \right]_0^\infty \\
 &= \theta \left(0 - \frac{-1}{\theta-t} \right) = \frac{\theta}{\theta-t}, \quad \text{for } t < \theta
 \end{aligned}$$

Obtain the first moment from the MGF:

$$\begin{aligned}
 E(X) &= M'_X(0) = \frac{d}{dt} \left(\frac{\theta}{\theta-t} \right) \Big|_{t=0} \\
 &= \theta \cdot \frac{1}{(\theta-t)^2} \Big|_{t=0} = \frac{\theta}{\theta^2} = \frac{1}{\theta}
 \end{aligned}$$

So, $\mu_1 = \frac{1}{\theta}$.

Find the first sample moment as:

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

Step 4: Equate moments and solve for θ .

$$\frac{1}{\theta} = \bar{X} \quad \Rightarrow \quad \theta = \frac{1}{\bar{X}}$$

Therefore, the method of moments estimator is:

$$\hat{\theta}_1 = \frac{1}{\bar{X}}$$

Notice the hat on top of $\hat{\theta}$, because it is our guess or estimate of the true θ , we put a hat on top of it to distinguish the estimator from the true population parameter.

Example 3.7

Example 3.8

Example 3.9.

Example 3.10

Show that if Z is a variable with standard normal distribution, then the moment generating function is $M_Z(t) = e^{t^2/2}$.

Solution:

The standard normal distribution has probability density function:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

The moment generating function (MGF) is defined as:

$$M_Z(t) = E[e^{tZ}] = \int_{-\infty}^{\infty} e^{tz} f(z) dz$$

Substituting the PDF:

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tz} \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Combine the exponents:

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz - z^2/2} dz$$

Complete the square in the exponent:

Consider the exponent:

$$tz - \frac{z^2}{2} = -\frac{1}{2}(z^2 - 2tz)$$

Complete the square:

$$z^2 - 2tz = (z - t)^2 - t^2$$

Therefore:

$$tz - \frac{z^2}{2} = -\frac{1}{2}[(z - t)^2 - t^2] = -\frac{1}{2}(z - t)^2 + \frac{t^2}{2}$$

Substitute back into the integral:

$$M_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2 + \frac{t^2}{2}} dz$$

Factor out the constant term:

$$M_Z(t) = e^{t^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2} dz$$

Recognize the normal distribution integral:

The integral $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z-t)^2} dz$ is the total probability of a normal distribution with mean t and variance 1, which equals 1.

Therefore:

$$M_Z(t) = e^{t^2/2} \cdot 1 = e^{t^2/2}$$

$$\boxed{M_Z(t) = e^{t^2/2}}$$

Verification: We can verify this result by checking that $M_Z(0) = 1$ (which is true for any MGF) and by computing moments:

$$M'_Z(t) = te^{t^2/2} \Rightarrow E[Z] = M'_Z(0) = 0$$

$$M''_Z(t) = (t^2 + 1)e^{t^2/2} \Rightarrow E[Z^2] = M''_Z(0) = 1$$

These match the mean and variance of the standard normal distribution.

3.2 Maximum Likelihood Method

The rationale of the method of maximum likelihood is based on selecting the values of the parameters that maximize the likelihood (the joint probability function or joint density function) of the *observed sample*. Notice the use of the expression, "*observed sample*": we have a sample at hand, and we try to choose the parameter values that make the observed sample most probable under the assumed statistical model. Consider the following example. Suppose that we are confronted with a box that contains three balls. We know that each of the balls may be red or white, but we do not know the total number of either color. However, we are allowed to randomly sample two of the balls without replacement. If our random sample yields two red balls, what would be a good estimate of the total number of red balls in the box? Obviously, the number of red balls in the box must be two or three (if there were zero or one red ball in the box, it would be impossible to obtain two red balls when sampling without replacement).

If there are two red balls and one white ball in the box, the probability of randomly selecting two red balls is

$$\frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}.$$

On the other hand, if there are three red balls in the box, the probability of randomly selecting two red balls is

$$\frac{\binom{3}{2}}{\binom{3}{2}} = 1.$$

It should seem reasonable to choose three as the estimate of the number of red balls in the box because this estimate maximizes the probability of obtaining the observed sample. Of course, it is possible for the box to contain only two red balls, but the observed outcome gives more credence to there being three red balls in the box.

More formally, suppose that the likelihood function depends on k parameters $\theta_1, \theta_2, \dots, \theta_k$. Method of Maximum Likelihood chooses as estimates those values of the parameters that maximize the likelihood

$$L(y_1, y_2, \dots, y_n \mid \theta_1, \theta_2, \dots, \theta_k).$$

Now let's go back to our math courses, MAT119 and MAT120. Do you recall how you would find the maximum of a function? We simply take the first derivative with respect to the variable we want to maximize, and then set the first derivative equal to zero to find the value where the slope is zero (example, utility maximization w.r.t. a consumption good). The idea is the same here. We will define a likelihood function and then find an estimator which maximizes the probability.

In our problem, the information in the sample and the parameter θ are involved in the joint distribution of the random sample; i.e., $\prod_{i=1}^n f(x_i; \theta)$. We want to view this as a function of θ , so we write it as

$$L(\theta) = L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i; \theta).$$

This is called the **likelihood function** of the random sample. As an estimate of θ , a measure of the center of $L(\theta)$ seems appropriate. An often-used estimate is the value of θ that provides a maximum

of $L(\theta)$. If it is unique, this is called the **maximum likelihood estimator (MLE)**, and we denote it as $\hat{\theta}$; i.e.,

$$\hat{\theta} = \underset{\theta}{\operatorname{Argmax}} L(\theta).$$

In practice, it is often much easier to work with the log of the likelihood, that is, the function $l(\theta) = \log L(\theta)$. Because the log is a strictly increasing function, the value that maximizes $l(\theta)$ is the same as the value that maximizes $L(\theta)$. Furthermore, for most of the models discussed in this course, the pdf (or pmf) is a differentiable function of θ , and frequently $\hat{\theta}$ solves the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

Example 3.12

Suppose the common pdf of the random sample $X_1, X_2, \dots, X_n$ is the $\Gamma(1, \theta)$ density $f(x) = \theta^{-1} \exp\{-x/\theta\}$ with support $0 < x < \infty$. This gamma distribution is often called the exponential distribution. The log of the likelihood function is given by

$$l(\theta) = \log \prod_{i=1}^n \frac{1}{\theta} e^{-x_i/\theta} = -n \log \theta - \theta^{-1} \sum_{i=1}^n x_i.$$

The first partial of the log-likelihood with respect to θ is

$$\frac{\partial l(\theta)}{\partial \theta} = -n\theta^{-1} + \theta^{-2} \sum_{i=1}^n x_i.$$

Setting this partial to 0 and solving for θ , we obtain the solution $\bar{x}$. There is only one critical value and, furthermore, the second partial of the log-likelihood evaluated at $\bar{x}$ is strictly negative, verifying that it provides a maximum. Hence, for this example, the statistic $\hat{\theta} = \bar{X}$ is the mle of θ . Because $E(X) = \theta$, we have that $E(\bar{X}) = \theta$ and, hence, $\hat{\theta}$ is an unbiased estimator of θ .

Example 3.13 (In Class)

Example 3.14 (In Class)

Example 3.15 (In Class)

3.3 Appendix

If $X \sim N(\mu, \sigma^2)$, then its moment-generating function (MGF) is defined as:

$$M_X(t) = E(e^{tX})$$

To find this expected value, we start with the probability density function (PDF) of X :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right), \quad \text{for } x \in \mathbb{R}.$$

The MGF is then given by the integral:

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

Combining the exponents inside the integral:

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

Our goal is to simplify the exponent into a perfect square in x . Let's expand and complete the square.

$$\begin{aligned} tx - \frac{1}{2\sigma^2}(x - \mu)^2 &= tx - \frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2) \\ &= -\frac{1}{2\sigma^2}x^2 + \left(t + \frac{\mu}{\sigma^2}\right)x - \frac{\mu^2}{2\sigma^2}. \end{aligned}$$

Now, we complete the square for the terms involving x . Consider the general quadratic form $ax^2 + bx + c$. Here,

$$a = -\frac{1}{2\sigma^2}, \quad b = t + \frac{\mu}{\sigma^2}, \quad c = -\frac{\mu^2}{2\sigma^2}.$$

Completing the square:

$$ax^2 + bx + c = a\left(x^2 + \frac{b}{a}x\right) + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c.$$

Let's compute the necessary components:

$$\begin{aligned} \frac{b}{2a} &= \frac{t + \frac{\mu}{\sigma^2}}{2 \cdot \left(-\frac{1}{2\sigma^2}\right)} = \frac{t + \frac{\mu}{\sigma^2}}{-1/\sigma^2} = -\sigma^2 \left(t + \frac{\mu}{\sigma^2}\right) = -\mu - \sigma^2 t. \\ -\frac{b^2}{4a} + c &= -\frac{\left(t + \frac{\mu}{\sigma^2}\right)^2}{4 \cdot \left(-\frac{1}{2\sigma^2}\right)} - \frac{\mu^2}{2\sigma^2} = \frac{\sigma^2}{2} \left(t + \frac{\mu}{\sigma^2}\right)^2 - \frac{\mu^2}{2\sigma^2}. \end{aligned}$$

Now, expand the squared term:

$$\left(t + \frac{\mu}{\sigma^2}\right)^2 = t^2 + \frac{2\mu t}{\sigma^2} + \frac{\mu^2}{\sigma^4}.$$

Substituting back:

$$-\frac{b^2}{4a} + c = \frac{\sigma^2}{2} \left(t^2 + \frac{2\mu t}{\sigma^2} + \frac{\mu^2}{\sigma^4}\right) - \frac{\mu^2}{2\sigma^2} = \frac{\sigma^2 t^2}{2} + \mu t + \frac{\mu^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} = \mu t + \frac{1}{2}\sigma^2 t^2.$$

Thus, the original exponent becomes:

$$tx - \frac{1}{2\sigma^2}(x - \mu)^2 = -\frac{1}{2\sigma^2}(x - (\mu + \sigma^2 t))^2 + \left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Substituting this back into the integral for $M_X(t)$:

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

We can factor out the term that does not depend on x :

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

The expression inside the square brackets is the integral of the PDF of a normal distribution with mean $\mu + \sigma^2 t$ and variance σ^2 . By definition, this integral equals 1.

Therefore, we are left with the final result:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

3.4 Sufficient Statistics

Recall the definition of a statistic: a statistic is a function $T = r(X_1, X_2, \dots, X_n)$ of the random sample $X_1, X_2, \dots, X_n$. Examples are:

$$\begin{aligned}\bar{X}_n &= \frac{1}{n} \sum_{i=1}^n X_i, \quad (\text{the sample mean}) \\ s^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad (\text{the sample variance}) \\ T_1 &= \max\{X_1, X_2, \dots, X_n\} \\ T_2 &= 10\end{aligned}$$

Although the last one seems a bit strange and different than the others (as it completely ignores the random sample), it is still a statistic. We say a statistic T is an estimator of a population parameter θ if T is usually close to θ . The sample mean is an estimator for the population mean; the sample variance is an estimator for the population variation.

Obviously, there are lots of functions of $X_1, X_2, \dots, X_n$ and so lots of statistics. When we look for a good estimator, do we really need to consider all of them, or is there a much smaller set of statistics we could consider? Another way to ask the question is if there are a few key functions of the random sample which will by themselves contain all the information the sample does. For example, suppose we know the sample mean and the sample variance. Does the random sample contain any more information about the population than this?

Let $f(x|\theta)$ with $\theta \in \Theta$ be some parametric family. Let $X = (X_1, \dots, X_n)$ be a random sample from distribution $f(x|\theta)$. Suppose we would like to learn parameter value θ from our sample. The concept of sufficient statistic allows us to separate information contained in X into two parts. One part contains all the valuable information as long as we are concerned with parameter θ , while the other part contains pure noise in the sense that this part has no valuable information. Thus, we can ignore the latter part.

Definition 4. *Statistic $T(X)$ is **sufficient** for θ if the conditional distribution of X given $T(X)$ does not depend on θ .*

Let $T(X)$ be a sufficient statistic. Consider the pair $(X, T(X))$. Obviously, $(X, T(X))$ contains the same information about θ as X alone, since $T(X)$ is a function of X . But if we know $T(X)$, then X itself has no value for us since its conditional distribution given $T(X)$ is independent of θ . Thus, by observing X (in addition to $T(X)$), we cannot say whether one particular value of parameter θ is more likely than another. Therefore, once we know $T(X)$, we can discard X completely.

It is difficult to use the definition to check if a statistic is sufficient or to find a sufficient statistic. Luckily, there is a theorem that makes it easy to find sufficient statistics.

Theorem 8 (Factorization theorem). *Let $X_1, X_2, \dots, X_n$ be a random sample with joint density $f(x_1, x_2, \dots, x_n|\theta)$. A statistic $T = r(X_1, X_2, \dots, X_n)$ is sufficient if and only if the joint density can be factored as follows:*

$$f(x_1, x_2, \dots, x_n|\theta) = u(x_1, x_2, \dots, x_n) v(r(x_1, x_2, \dots, x_n), \theta) \quad (1)$$

where u and v are non-negative functions. The function u can depend on the full random sample $x_1, \dots, x_n$, but not on the unknown parameter θ . The function v can depend on θ , but can depend on the random sample only through the value of $r(x_1, \dots, x_n)$.

It is easy to see that if $f(t)$ is a one to one function and T is a sufficient statistic, then $f(T)$ is a sufficient statistic. In particular we can multiply a sufficient statistic by a nonzero constant and get another sufficient statistic.

We now apply the theorem to some examples.

Example (normal population, unknown mean, known variance) We consider a normal population for which the mean μ is unknown, but the variance σ^2 is known. The joint density is

$$\begin{aligned} f(x_1, \dots, x_n | \mu) &= (2\pi)^{-n/2} \sigma^{-n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2} \right) \end{aligned}$$

Since σ^2 is known, we can let

$$u(x_1, \dots, x_n) = (2\pi)^{-n/2} \sigma^{-n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 \right)$$

and

$$v(r(x_1, x_2, \dots, x_n), \mu) = \exp \left(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} r(x_1, x_2, \dots, x_n) \right)$$

where

$$r(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i$$

By the factorization theorem this shows that $\sum_{i=1}^n X_i$ is a sufficient statistic. It follows that the sample mean $\bar{X}_n$ is also a sufficient statistic.

Intuition for Sufficient Statistics

A **sufficient statistic** is a summary of the data that contains **all the information** needed to estimate a parameter. Once you know the sufficient statistic, the actual raw data provides no additional information about the parameter.

A Simple Example: The Coin Flip Story

Imagine you flip a coin 100 times to determine if it's fair. You could:

- **Option 1:** Record the entire sequence (H,T,H,H,T,...) - 100 pieces of information

- **Option 2:** Just record "43 heads" - 1 piece of information

For estimating the probability of heads, *Option 2 contains all the information* you need! The exact sequence doesn't matter - only the total number of heads. A sufficient statistic **compresses** the data without losing any information about the parameter. It's the **optimal summary** for estimation purposes.

Example 3.17: Find two sufficient Statistics for $\text{Uniform}[\theta_1, \theta_2]$.

Let $X_1, X_2, \dots, X_n \sim \text{Uniform}[\theta_1, \theta_2]$ be i.i.d. with $\theta_1 < \theta_2$. The joint density is:

$$f(x_1, \dots, x_n \mid \theta_1, \theta_2) = \prod_{i=1}^n \frac{1}{\theta_2 - \theta_1} \cdot I_{[\theta_1, \theta_2]}(x_i)$$

where $I_A(x)$ is the indicator function that equals 1 if $x \in A$ and 0 otherwise.

Write the joint density clearly:

$$f(x_1, \dots, x_n \mid \theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} \cdot I\{\theta_1 \leq x_1, \dots, x_n \leq \theta_2\}$$

Express the indicator function more usefully:

The condition $\theta_1 \leq x_1, \dots, x_n \leq \theta_2$ is equivalent to:

$$\theta_1 \leq \min\{x_1, \dots, x_n\} \quad \text{and} \quad \max\{x_1, \dots, x_n\} \leq \theta_2$$

So we can write:

$$f(x_1, \dots, x_n \mid \theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} \cdot I\{\theta_1 \leq X_{(1)}\} \cdot I\{X_{(n)} \leq \theta_2\}$$

where $X_{(1)} = \min\{X_1, \dots, X_n\}$ and $X_{(n)} = \max\{X_1, \dots, X_n\}$.

Apply the Factorization Theorem:

We can factor the joint density as:

$$f(x_1, \dots, x_n \mid \theta_1, \theta_2) = \underbrace{1}_{u(x_1, \dots, x_n)} \cdot \underbrace{\frac{1}{(\theta_2 - \theta_1)^n} I\{\theta_1 \leq X_{(1)}\} I\{X_{(n)} \leq \theta_2\}}_{v(T_1, T_2, \theta_1, \theta_2)}$$

By the Factorization Theorem, the following statistics are **sufficient**:

$$\boxed{T_1 = X_{(1)} = \min\{X_1, \dots, X_n\}}$$

$$T_2 = X_{(n)} = \max\{X_1, \dots, X_n\}$$

Verification: Why These Are Sufficient?

- $T_1 = \min X_i$ tells us the **smallest possible value** for θ_1
- $T_2 = \max X_i$ tells us the **largest possible value** for θ_2
- Once we know these two values, the individual data points provide no additional information about θ_1 and θ_2
- The conditional distribution of the data given (T_1, T_2) is uniform over all permutations of values between T_1 and T_2 , which doesn't depend on the parameters

The intuition is the following: if you want to estimate the interval $[\theta_1, \theta_2]$, you only need to know:

- The **smallest observation** (to estimate the left endpoint θ_1)
- The **largest observation** (to estimate the right endpoint θ_2)

The exact positions of the other observations between these extremes don't matter for estimating the parameters!

3.5 Rao-Blackwell Theorem and Minimum Variance Unbiased Estimator

Sufficient statistics play an important role in finding good estimators for parameters. If $\hat{\theta}$ is an unbiased estimator for θ and if U is a statistic that is sufficient for θ , then there is a function of U that is also an unbiased estimator for θ and has no larger variance than $\hat{\theta}$. If we seek unbiased estimators with small variances, we can restrict our search to estimators that are functions of sufficient statistics. The theoretical basis for the preceding remarks is provided in the following result, known as the Rao-Blackwell theorem.

Theorem 9 (Rao-Blackwell Theorem). *Let $\hat{\theta}$ be an unbiased estimator for θ such that $V(\hat{\theta}) < \infty$. If U is a sufficient statistic for θ , define $\hat{\theta}^* = E(\hat{\theta} | U)$. Then, for all θ ,*

$$E(\hat{\theta}^*) = \theta \quad \text{and} \quad V(\hat{\theta}^*) \leq V(\hat{\theta}).$$

Proof. Because U is sufficient for θ , the conditional distribution of any statistic (including $\hat{\theta}$), given U , does not depend on θ . Thus, $\hat{\theta}^* = E(\hat{\theta} | U)$ is not a function of θ and is therefore a statistic. Because $\hat{\theta}$ is an unbiased estimator for θ , Law of Iterated Expectations implies that

$$E(\hat{\theta}^*) = E[E(\hat{\theta} | U)] = E(\hat{\theta}) = \theta.$$

Thus, $\hat{\theta}^*$ is an unbiased estimator for θ .

$$V(\hat{\theta}) = V[E(\hat{\theta} | U)] + E[V(\hat{\theta} | U)] = V(\hat{\theta}^*) + E[V(\hat{\theta} | U)].$$

Because $V(\hat{\theta} | U = u) \geq 0$ for all u , it follows that $E[V(\hat{\theta} | U)] \geq 0$ and therefore that $V(\hat{\theta}) \geq V(\hat{\theta}^*)$, as claimed. $\square$

The Rao-Blackwell Theorem provides a method to improve any unbiased estimator. The core idea is that if you waste information, your estimator will be noisy. Use all available information to smooth out the noise.

Start with any unbiased estimator $\delta(X)$. Find a sufficient statistic $T(X)$ for the parameter θ . Then create a new estimator by taking the conditional expectation: $\delta^*(X) = E[\delta(X)|T(X)]$.

This new estimator is guaranteed to be unbiased and have variance less than or equal to the original estimator: $\text{Var}(\delta^*) \leq \text{Var}(\delta)$. The theorem smooths out unnecessary variability by using all the information contained in the sufficient statistic.

Example 3.18: Bernoulli Distribution Sufficient Statistic and MVUE

Let $Y_1, Y_2, \dots, Y_n$ denote a random sample from a distribution where $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$, with p unknown (such random variables are often called Bernoulli variables). Use the factorization criterion to find a sufficient statistic that best summarizes the data. Give an MVUE for p .

Notice that the preceding probability function can be written as

$$P(Y_i = y_i) = p^{y_i}(1 - p)^{1 - y_i}, \quad y_i = 0, 1.$$

Thus, the likelihood $L(p)$ is

$$\begin{aligned} L(y_1, y_2, \dots, y_n | p) &= p(y_1, y_2, \dots, y_n | p) \\ &= p^{y_1}(1 - p)^{1 - y_1} \times p^{y_2}(1 - p)^{1 - y_2} \times \dots \times p^{y_n}(1 - p)^{1 - y_n} \\ &= p^{\sum y_i} (1 - p)^{n - \sum y_i} \\ &= \underbrace{p^{\sum y_i} (1 - p)^{n - \sum y_i}}_{g(\sum y_i, p)} \times \underbrace{1}_{h(y_1, y_2, \dots, y_n)}. \end{aligned}$$

According to the factorization criterion, $U = \sum_{i=1}^n Y_i$ is sufficient for p . This statistic best summarizes the information about the parameter p .

Notice that $E(U) = np$, or equivalently, $E(U/n) = p$. Thus, $U/n = \bar{Y}$ is an unbiased estimator for p . Because this estimator is a function of the sufficient statistic $\sum_{i=1}^n Y_i$, the estimator $\hat{p} = \bar{Y}$ is the MVUE for p .

Example 3.19

Suppose that $Y_1, Y_2, \dots, Y_n$ denote a random sample from the Weibull density function, given by

$$f(y | \theta) = \begin{cases} \left(\frac{2y}{\theta}\right) e^{-y^2/\theta}, & y > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Find an MVUE for θ .

Solution: We begin by using the factorization criterion to find the sufficient statistic that best summarizes the information about θ .

$$\begin{aligned} L(y_1, y_2, \dots, y_n \mid \theta) &= f(y_1, y_2, \dots, y_n \mid \theta) \\ &= \left(\frac{2}{\theta}\right)^n (y_1 \times y_2 \times \dots \times y_n) \exp\left(-\frac{1}{\theta} \sum_{i=1}^n y_i^2\right) \\ &= \underbrace{\left(\frac{2}{\theta}\right)^n \exp\left(-\frac{1}{\theta} \sum_{i=1}^n y_i^2\right)}_{u(\sum y_i^2, \theta)} \times \underbrace{(y_1 \times y_2 \times \dots \times y_n)}_{r(y_1, y_2, \dots, y_n)}. \end{aligned}$$

Thus, $U = \sum_{i=1}^n Y_i^2$ is the minimal sufficient statistic for θ .

We now must find a function of this statistic that is unbiased for θ . Letting $W = Y_i^2$, we have

$$\begin{aligned} f_W(w) &= f(\sqrt{w}) \left| \frac{d(\sqrt{w})}{dw} \right| \\ &= \left(\frac{2}{\theta}\right) (\sqrt{w} e^{-w/\theta}) \left(\frac{1}{2\sqrt{w}}\right) \\ &= \left(\frac{1}{\theta}\right) e^{-w/\theta}, \quad w > 0. \end{aligned}$$

That is, Y_i^2 has an exponential distribution with parameter θ . Because

$$E(Y_i^2) = E(W) = \theta \quad \text{and} \quad E\left(\sum_{i=1}^n Y_i^2\right) = n\theta,$$

it follows that

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i^2$$

is an unbiased estimator of θ that is a function of the sufficient statistic $\sum_{i=1}^n Y_i^2$. Therefore, $\hat{\theta}$ is an MVUE of the Weibull parameter θ .

4 Confidence Intervals

So far we learned what a point estimator is, what some desired properties of point estimators are (unbiasedness, efficiency, i.e. minimum variance, and consistency), and how we can derive such estimators. We will now learn about confidence intervals for an unknown population parameter.

One way to measure the goodness of any point estimation procedure is in terms of the distances between the estimates that it generates and the target parameter. This quantity, which varies randomly in repeated sampling, is called the error of estimation. Naturally we would like the error of estimation to be as small as possible.

The **estimation error** ε measures how far our estimate $\hat{\theta}$ is from the true parameter θ we're trying to measure:

$$\varepsilon = |\hat{\theta} - \theta|$$

Since $\hat{\theta}$ is calculated from data (and is therefore random), the error ε is also random. We can't know the exact error for any particular estimate, but we can make **probability statements** about how large it might be.

Imagine the sampling distribution of $\hat{\theta}$ centered around the true population parameter θ (since it's unbiased). If we choose a margin b around θ , the probability that our estimate falls within this margin is:

$$P(|\hat{\theta} - \theta| < b) = P(\theta - b < \hat{\theta} < \theta + b)$$

This probability identifies the fraction of times, in repeated sampling, that the estimator $\hat{\theta}$ falls within b units of θ , the target parameter. When b is practically small, a high probability $P(\varepsilon < b)$ provides a measure of the goodness of a single estimate.

Suppose that we want to find the value of b so that $P(\varepsilon < b) = 0.90$. If we know the probability density function of $\hat{\theta}$, then we seek a value b such that:

$$\int_{\theta-b}^{\theta+b} f(\hat{\theta}) d\hat{\theta} = 0.90$$

where $f(\hat{\theta})$ is the probability density function of our estimator.

Even without knowing the exact distribution, if $\hat{\theta}$ is unbiased we can use its **standard error**² $\sigma_{\hat{\theta}}$ to create error bounds. Letting $b = k\sigma_{\hat{\theta}}$, Chebyshev's inequality tells us:

$$P(\varepsilon < k\sigma_{\hat{\theta}}) \geq 1 - \frac{1}{k^2}$$

For example, with $k = 2$:

$$P(\varepsilon < 2\sigma_{\hat{\theta}}) \geq 0.75$$

This means there's at least a 75% chance our estimation error will be less than two standard errors. The point is that $b = 2\sigma_{\hat{\theta}}$ is a good approximate bound on the error of estimation in most

²The standard deviation of an estimator is called the standard error.

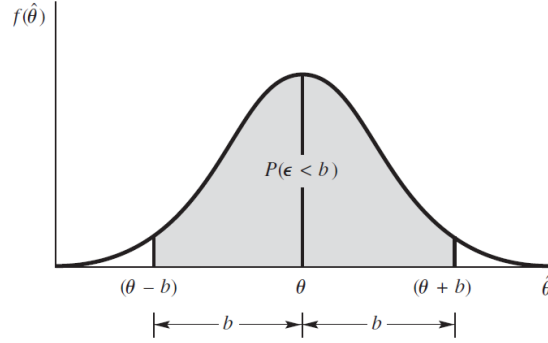


Figure 3: Sampling distribution of a point estimator $\hat{\theta}$

practical situations. The bounds for probabilities provided by Tchebysheff's theorem are usually very conservative; the actual probabilities usually exceed the Tchebysheff bounds by a considerable amount. The probability that a random variable Y falls within two standard deviations of its mean, $P(\mu - 2\sigma < Y < \mu + 2\sigma)$, varies across different distributions. For the Normal distribution, this probability is approximately 0.9544. For the Uniform distribution, the probability is exactly 1.0000, meaning all values necessarily fall within this range. For the Exponential distribution, the probability is approximately 0.9502.

In an ideal scenario, a confidence interval should satisfy two key conditions: it must include the true parameter θ , and it should be as precise (narrow) as possible. However, because the interval's endpoints are derived from sample data, they fluctuate randomly across different samples. This means the position and width of the interval are both random, making it impossible to guarantee that any specific computed interval contains the fixed parameter θ . Therefore, the aim is to develop an interval estimation method that consistently produces narrow intervals with a high likelihood of capturing θ .

Such intervals are referred to as **confidence intervals**. Their endpoints are known as the **upper and lower confidence limits**. The probability that a randomly constructed confidence interval contains θ is called the **confidence coefficient**. In practical terms, this coefficient indicates how frequently, over many repeated samples, the generated intervals would successfully include the target parameter θ .

Let's be a bit more formal; consider θ , an unknown parameter. Suppose that on the basis of sample information, random variables A and B are found such that $P(A < \theta < B) = 1 - \alpha$, where α is any number between 0 and 1. If the specific sample values of A and B are a and b , then the interval from a to b is called a $100(1 - \alpha)\%$ **confidence interval** of θ . The quantity $100(1 - \alpha)\%$ is called the **confidence level** of the interval. If the population is repeatedly sampled by a very large number of times, the true value of the parameter θ will be covered by $100(1 - \alpha)\%$ of intervals calculated this way.

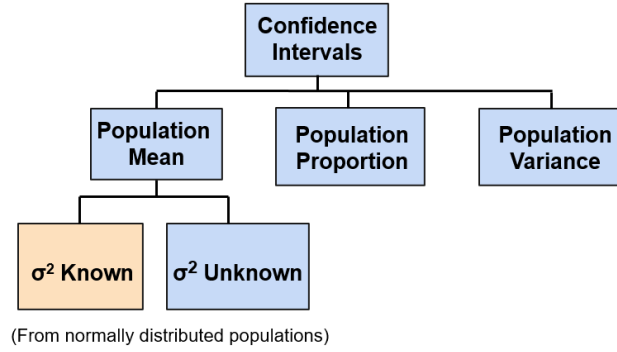


Figure 4: Confidence Intervals

Confidence Interval Estimation Based on the Normal Distribution: One Population Intervals

Let $x_1, x_2, \dots, x_n$ be a random sample of n observations from a normally distributed population with unknown mean μ and known variance σ^2 . Suppose that we want a $100(1 - \alpha)\%$ confidence interval for the population mean. We know from Econ205 that

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

has a standard normal distribution and $z_{\alpha/2}$ is the value from the standard normal distribution such that the upper tail probability is $\alpha/2$. We use the following algebraic manipulation:

$$\begin{aligned}
 1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
 &= P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \\
 &= P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{x} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \\
 &= P\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)
 \end{aligned}$$

For a 95% confidence level it follows that

$$P\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

We know that the probability that a standard normal random variable falls between the numbers -1.96 and 1.96 is 0.95 .

Confidence Interval Estimation for the Mean of a Population That Is Normally Distributed: Population Variance Known

Consider a random sample of n observations from a normal distribution with mean μ and variance σ^2 . If the sample mean is $\bar{x}$, then a $100(1 - \alpha)\%$ confidence interval for the population mean with known variance is given by

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or, equivalently,

$$\bar{x} \pm ME$$

where ME , the margin of error (also called the sampling error), is given by

$$ME = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The width, w , is equal to twice the margin of error:

$$w = 2(ME)$$

The upper confidence limit (UCL) is given by

$$UCL = \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

The lower confidence limit (LCL) is given by

$$LCL = \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

We need to interpret confidence intervals accurately. If random samples of n observations are drawn repeatedly and independently from the population and $100(1 - \alpha)\%$ confidence intervals are calculated by Equation $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$, then over a very large number of repeated trials, $100(1 - \alpha)\%$ of these intervals will contain the true value of the population mean.

This is a crucial point that most students tend to overlook. When we estimate a confidence interval, our specific interval *either includes the true parameter value or not* (simply because confidence intervals are estimated based on the random $\hat{\theta}$). But the idea is that if we repeat the exercise, say a 100 times, we can be sure that a certain percentage of such intervals will include the true parameter, say

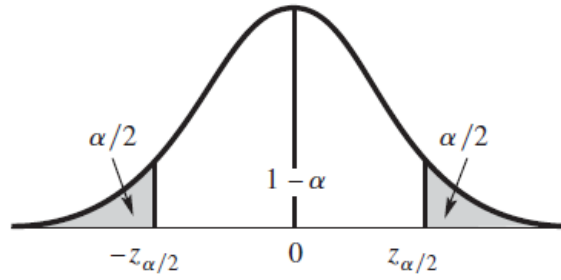


Figure 5: $100(1 - \alpha)\%$ Confidence Interval

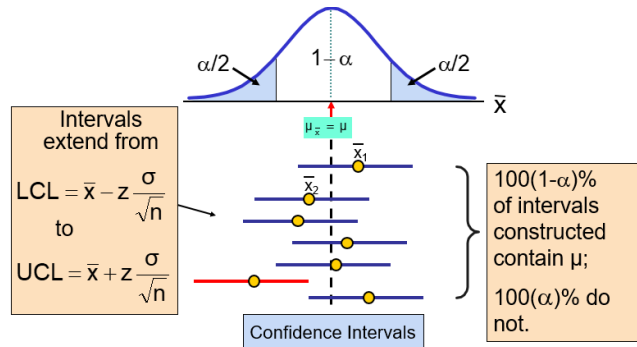


Figure 6: Intervals and Level of Confidence

95 percent. Let's take a look at Figure 7. Assume that we draw samples from a normally distributed population and calculate $\bar{x}$ for each sample. If we set $1 - \alpha$ to 0.95, this means that when we draw 100 samples and estimate 100 confidence intervals, 95 of these intervals will include the true μ . In Figure 7, all the blue intervals include the point estimates within the $(1 - \alpha)$ region, whereas the point estimate falls outside $(1 - \alpha)$ region.

For selected confidence levels, the following Table lists corresponding values of $z_{\alpha/2}$, sometimes called the reliability factor. Consequently, a 90% confidence interval is estimated as:

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

Table 1: Confidence Level and Corresponding Critical Values

Confidence Level	$z_{\alpha/2}$ Value
90%	1.645
95%	1.96
98%	2.33
99%	2.58
99.9%	3.27

Example 4.1

A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Find a 95% confidence interval for the true mean resistance of the population.

Solution

The following information is given: Sample mean: $\bar{x} = 2.20$ ohms, Population standard deviation: $\sigma = 0.35$ ohms, Sample size: $n = 11$. For a 95% confidence interval, the critical value is $z_{\alpha/2} = 1.96$.

First, calculate the standard error:

$$\frac{\sigma}{\sqrt{n}} = \frac{0.35}{\sqrt{11}} \approx \frac{0.35}{3.3166} \approx 0.1055$$

The margin of error is calculated as:

$$ME = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{0.35}{\sqrt{11}}$$

$$ME = 1.96 \cdot 0.1055 \approx 0.2068$$

The 95% confidence interval is:

$$\bar{x} \pm ME = 2.20 \pm 0.2068$$

So the confidence interval is:

$$(2.20 - 0.2068, 2.20 + 0.2068) = (1.9932, 2.4068)$$

The 95% confidence interval for the true mean resistance of the population is approximately **(1.993, 2.407) ohms**.

We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms. Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean.

Example 4.2

The shopping times of $n = 64$ randomly selected customers at a local supermarket were recorded. The average of the 64 shopping times were recorded as 33 minutes and variance is known to be 256 minutes. Estimate μ , the true average shopping time per customer, with a confidence coefficient of $1 - \alpha = 0.90$.

Solution

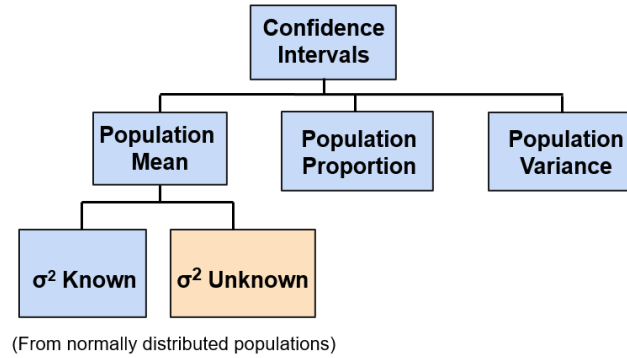


Figure 7: Confidence Intervals

Given: sample size $n = 64$, sample mean $\bar{x} = 33$ minutes, variance $\sigma^2 = 256$ minutes², and confidence level $1 - \alpha = 0.90$. The sample standard deviation is $\sigma = \sqrt{256} = 16$ minutes.

For a 90% confidence interval, the critical value is $z_{\alpha/2} = z_{0.05} = 1.645$.

The margin of error is calculated as:

$$ME = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.645 \cdot \frac{16}{\sqrt{64}} = 1.645 \cdot \frac{16}{8} = 1.645 \cdot 2 = 3.29$$

The 90% confidence interval is:

$$\bar{x} \pm ME = 33 \pm 3.29$$

$$\text{Lower limit} = 33 - 3.29 = 29.71, \quad \text{Upper limit} = 33 + 3.29 = 36.29$$

The 90% confidence interval for the true average shopping time per customer is **(29.71, 36.29) minutes**.

4.1 Confidence Interval Estimation for the Mean of a Population That Is Normally Distributed: Population Variance UnKnown

In most cases when the true population mean is unknown, the population variance is also unknown. In such cases, whenever we want to construct a confidence interval, we use the sample variance, s^2 in lieu of the unknown population variance. This introduces additional uncertainty into our inference, especially if the sample size is small. When we know σ , the quantity $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows exactly $N(0, 1)$ by the Central Limit Theorem. However, when we replace σ with its estimate s , we get a different distribution. Therefore when the population variance is unknown, we can no longer rely on critical z values in small samples and we rely on the t -distribution. The t -distribution accounts for uncertainty in three key ways:

- **Heavier tails:** The t -distribution has thicker tails than the normal distribution, reflecting the increased variability that comes from estimating σ^2 from the sample.

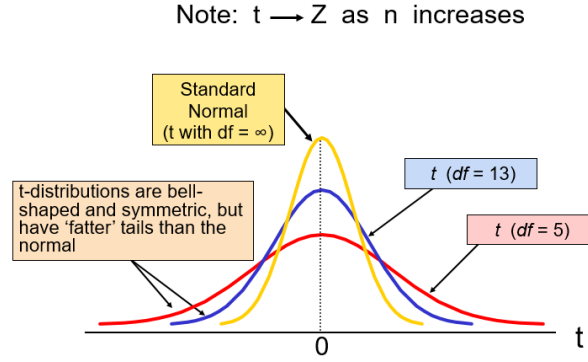


Figure 8: Confidence Intervals

- **Sample size dependence:** As the sample size n increases, s^2 becomes a better estimate of σ^2 , and the t-distribution converges to the normal distribution.
- **Uncertainty penalty:** Using critical t-values instead of z-values automatically widens our confidence intervals to account for the uncertainty in our variance estimate, giving us more conservative (and more accurate) intervals.

If we collect a random sample of n observations with mean $\bar{x}$ and standard deviation s from a normally distributed population with mean μ , then the variable:

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

follows the Student's t distribution with $(n - 1)$ degrees of freedom.

We therefore make the following adjustment for confidence interval estimation:

Consider $\{x_i\}_{i=1}^n$ a random sample of n observations from a normal population $N(\mu, \sigma^2)$. If the sample mean is $\bar{x}$, then a $(1 - \alpha)100\%$ confidence interval for μ with unknown σ^2 is:

$$\bar{x} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

where,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Example 4.3

A random sample of $n = 25$ with $\bar{x} = 50$ and $s = 8$. Form a 95% confidence interval for μ .

Solution

The confidence interval for μ when σ is unknown is given by:

$$\bar{x} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

We have the following values: sample size $n = 25$, sample mean $\bar{x} = 50$, sample standard deviation $s = 8$, confidence level 95% so $\alpha = 0.05$, degrees of freedom $df = n - 1 = 24$, and critical t-value $t_{24, 0.975} = 2.064$ (from t-table).

Calculating the margin of error:

$$E = t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 2.064 \times \frac{8}{\sqrt{25}} = 2.064 \times \frac{8}{5} = 2.064 \times 1.6 = 3.3024$$

The 95% confidence interval is:

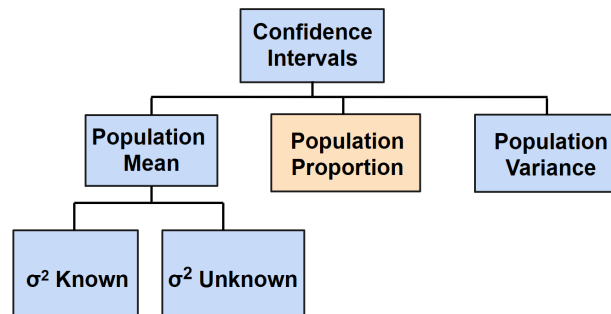
$$50 \pm 3.3024 = (46.6976, 53.3024)$$

Therefore, the 95% confidence interval for μ is:

$$(46.70, 53.30)$$

Example 4.4: In class

4.2 Confidence Interval Estimation for Population Proportion



Consider $\{x_i\}_{i=1}^n$, a random sample of n observations from a Bernoulli(p) population. Notice that each x_i is either 1 (success) or 0 (failure). $\bar{x}$ in this case is the observed proportion of successes, denoted as $\hat{p}$. Then, if $n\hat{p}(1 - \hat{p}) \geq 5$, a $(1 - \alpha)100\%$ confidence interval for p is:

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

The distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation:

$$\sigma_p = \sqrt{\frac{P(1-P)}{n}}$$

When we have a sample at hand, we estimate it using the formula: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Example 4.5

A random sample of $n = 100$ people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

Solution

The sample proportion is:

$$\hat{p} = \frac{25}{100} = 0.25$$

Check the normal approximation condition:

$$n\hat{p}(1-\hat{p}) = 100 \times 0.25 \times 0.75 = 18.75 > 5$$

The condition is satisfied, so we can use the normal approximation.

For a 95% confidence interval:

$$\alpha = 0.05, \quad z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$$

The standard error is:

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.25 \times 0.75}{100}} = \sqrt{\frac{0.1875}{100}} = \sqrt{0.001875} = 0.0433$$

The margin of error is:

$$E = z_{1-\frac{\alpha}{2}} \cdot \sigma_{\hat{p}} = 1.96 \times 0.0433 = 0.0849$$

The 95% confidence interval is:

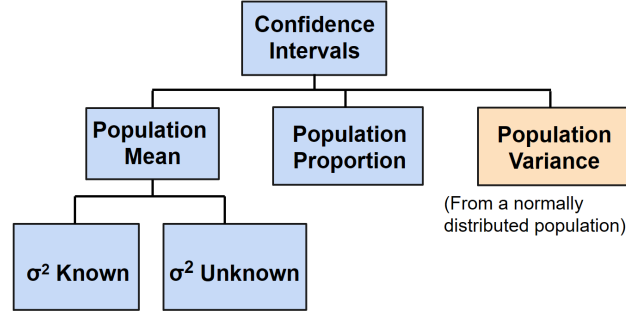
$$\hat{p} \pm E = 0.25 \pm 0.0849 = (0.1651, 0.3349)$$

Therefore, the 95% confidence interval for the true proportion of left-handers is:

$$\boxed{(0.165, 0.335)}$$

Example 4.6: In class

4.3 Confidence Interval Estimation for Population Variance



The chi-square (χ^2) distribution describes the sum of squared independent standard normal random variables. This distribution is fundamental for making inferences about population variance.

When sampling from a normal distribution $N(\mu, \sigma^2)$, the variable:

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$$

follows a chi-square distribution with $n - 1$ degrees of freedom, where:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

is the sample variance.

The notation $\chi_{n-1,\alpha}^2$ represents the critical value satisfying:

$$P(\chi_{n-1}^2 > \chi_{n-1,\alpha}^2) = \alpha$$

Using this relationship, we can construct a $(1 - \alpha)100\%$ confidence interval for σ^2 as:

$$\left[\frac{(n-1)s^2}{\chi_{n-1,1-\frac{\alpha}{2}}^2}, \frac{(n-1)s^2}{\chi_{n-1,\frac{\alpha}{2}}^2} \right]$$

This interval is asymmetric, reflecting the skewed nature of the chi-square distribution and providing proper coverage for the unknown population variance.

Example 4.7

A process engineer is concerned with the variation of temperature in an industrial furnace. She collects a random sample of temperatures as:

975, 1075, 1050, 900, 1000, 950, 1025, 1050, 975 °C

Calculate a 95% confidence interval for the (population) variance of temperatures in this furnace.

Solution

Given: $n = 9$ temperatures

First, calculate the sample mean:

$$\bar{x} = \frac{975 + 1075 + 1050 + 900 + 1000 + 950 + 1025 + 1050 + 975}{9} = \frac{9000}{9} = 1000$$

Now calculate the sample variance:

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n - 1} \\ &= \frac{(975 - 1000)^2 + (1075 - 1000)^2 + (1050 - 1000)^2 + (900 - 1000)^2 + (1000 - 1000)^2}{8} \\ &\quad + \frac{(950 - 1000)^2 + (1025 - 1000)^2 + (1050 - 1000)^2 + (975 - 1000)^2}{8} \\ &= \frac{625 + 5625 + 2500 + 10000 + 0 + 2500 + 625 + 2500 + 625}{8} \\ &= \frac{25000}{8} = 3125 \end{aligned}$$

For a 95% confidence interval for variance with $n = 9$:

$$\begin{aligned} \alpha &= 0.05 \\ df &= n - 1 = 8 \\ \chi_{8,0.975}^2 &= 2.180 \quad (\text{from chi-square table}) \\ \chi_{8,0.025}^2 &= 17.535 \quad (\text{from chi-square table}) \end{aligned}$$

The 95% confidence interval for σ^2 is:

$$\left[\frac{(n-1)s^2}{\chi_{n-1,1-\frac{\alpha}{2}}^2}, \frac{(n-1)s^2}{\chi_{n-1,\frac{\alpha}{2}}^2} \right] = \left[\frac{8 \times 3125}{17.535}, \frac{8 \times 3125}{2.180} \right]$$

Calculate the bounds:

$$\begin{aligned} \text{Lower bound} &= \frac{25000}{17.535} = 1425.8 \\ \text{Upper bound} &= \frac{25000}{2.180} = 11467.9 \end{aligned}$$

Therefore, the 95% confidence interval for the population variance is:

$$\boxed{(1425.8, 11467.9)}$$

4.4 Confidence Interval Estimation: Finite Populations

If the sample size is more than 5% of the population size (and sampling is without replacement) then a finite population correction factor must be used when calculating the standard error. Suppose sampling is without replacement and the sample size is large relative to the population size. Assuming the population size is large enough to apply the central limit theorem, we can apply the finite population correction factor when estimating the population variance.

Finite population correction factor: $\frac{N-n}{N-1}$

Estimating the Population Mean

Let a simple random sample of size n be taken from a population of N members with mean μ . The sample mean is an **unbiased estimator** of the population mean μ . The point estimate is:

$$\hat{\mu} = \bar{x}$$

Finite Populations: Mean

If the sample size is more than 5% of the population size, an unbiased estimator for the variance of the sample mean is:

$$\text{Var}(\bar{x}) = \frac{s^2}{n} \left(\frac{N-n}{N-1} \right)$$

So the $100(1 - \alpha)\%$ confidence interval for the population mean is:

$$\bar{x} \pm t_{n-1, \alpha/2} \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N-1} \right)}$$

Estimating the Population Total

Consider a simple random sample of size n from a population of size N . The quantity to be estimated is the population total $N\mu$. An unbiased estimation procedure for the population total $N\mu$ yields the point estimate $N\bar{x}$.

An unbiased estimator of the **variance** of the population total is:

$$\text{Var}(N\bar{x}) = N^2 \left(\frac{s^2}{n} \right) \left(\frac{N-n}{N-1} \right)$$

A $100(1 - \alpha)\%$ confidence interval for the population total is:

$$N\bar{x} \pm t_{n-1, \alpha/2} \sqrt{N^2 \left(\frac{s^2}{n} \right) \left(\frac{N-n}{N-1} \right)}$$

Confidence Interval for Population Total: Example

A firm has a population of 1000 accounts and wishes to estimate the value of the total population balance. A sample of 80 accounts is selected with average balance of \$87.60 and standard deviation of \$22.30. Find the 95% confidence interval estimate of the total balance.

Solution: The 95% confidence interval for the population total balance is \$82,837.53 to \$92,362.47.

$N = 1000$, $n = 80$, $\bar{x} = 87.6$, $s = 22.3$

$$\begin{aligned} N\hat{\sigma}_{\bar{x}}^2 &= N^2 \left(\frac{s^2}{n} \right) \left(\frac{N-n}{N-1} \right) \\ &= (1000)^2 \left(\frac{(22.3)^2}{80} \right) \left(\frac{920}{999} \right) \\ N\bar{X} \pm t_{0.025,79} \cdot N\hat{\sigma}_{\bar{x}} &= (1000)(87.6) \pm (1.9905)(2392.6) \end{aligned}$$

$$82,837.53 < N\mu < 92,362.47$$

Estimating the Population Proportion: Finite Population

Let the true population proportion be P . Let $\hat{p}$ be the sample proportion from n observations from a simple random sample. The sample proportion, $\hat{p}$, is an unbiased estimator of the population proportion, P .

Confidence Intervals for Population Proportion: Finite Population

If the sample size is more than 5% of the population size, an unbiased estimator for the variance of the population proportion is:

$$\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)$$

So the $100(1-\alpha)\%$ confidence interval for the population proportion is:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \left(\frac{N-n}{N-1} \right)}$$

4.5 Sample-Size Determination: Finite Populations

Finite Populations - For the Mean

How can we find the required sample size for a level of margin or error that we are willing to tolerate? Remember the margin of error formula:

$$ME = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Now using this formula, we can work our way to find the required sample size n_0 using the prior formula:

$$n_0 = \frac{z_{\alpha/2}^2 \sigma^2}{ME^2}$$

To determine the required sample size for the mean, you must decide on the desired level of confidence $(1 - \alpha)$, which determines the critical z value, the acceptable margin of error (sampling error), ME, and you must know the population standard deviation, σ .

Example 4.8

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

Solution

For a 90% confidence level:

$$\alpha = 0.10, \quad z_{\alpha/2} = z_{0.05} = 1.645$$

The required sample size formula is:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{ME} \right)^2$$

Substitute the values:

$$n = \left(\frac{1.645 \times 45}{5} \right)^2 = \left(\frac{74.025}{5} \right)^2 = (14.805)^2 = 219.188$$

Since sample size must be an integer, we round up:

$$n = 220$$

Therefore, the required sample size is:

$$\boxed{220}$$

The logic is similar for finding the required sample size for population proportions. The ME formula for population proportion is:

$$ME = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Using this formula, one can find the required sample size. However, the sample and population proportions, $\hat{p}$ and P , are generally not known (since no sample has been taken yet). Since the value $P(1 - P) = 0.25$ generates the largest possible margin of error, taking a conservative value such as $P = 0.5$ guarantees that the resulting sample size will meet the desired level of confidence. Therefore, to determine the required sample size for the proportion, you must decide on the desired level of confidence $(1 - \alpha)$, which then gives us the critical $z_{\alpha/2}$ value; the acceptable sampling error (margin of error), ME ; and estimate $P(1 - P) = 0.25$.

Example 4.9: In class

4.6 Confidence interval estimation: Two populations

In statistical inference, we often encounter situations where we need to compare parameters from two different populations. While one-population confidence intervals provide estimates for a single population parameter, two-population confidence intervals allow us to make inferences about the *difference* between two population parameters, such as means or proportions. These intervals are essential for comparing treatment effects, analyzing group differences, and testing hypotheses about population disparities. Whether comparing the average incomes of two regions, the effectiveness of two medical treatments, or the success rates of two marketing strategies, two-population confidence intervals provide a rigorous framework for quantifying the uncertainty in our estimates of the differences between population parameters. The construction of these intervals depends on whether the samples are independent or paired, and whether the population variances are known, unknown but assumed equal, or unknown and assumed unequal.

Difference between Two Normal Population Means, Case 1: Dependent (Matched) Samples

Let $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be two matched samples. We can then create:

$$\{d_i\}_{i=1}^n \quad \text{where} \quad d_i = x_i - y_i, \quad \forall i$$

Then, a $(1 - \alpha)100\%$ confidence interval for $\mu_d = \mu_x - \mu_y$ is:

$$\bar{d} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$$

where

$$s_d = \sqrt{s_d^2} = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}}$$

and

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n}$$

Example 4.10

Six people sign up for a weight loss program. You collect the following data:

Table 2: Before and After Measurements with Differences

Person	Before (x)	After (y)	Difference, d_i
1	136	125	11
2	205	195	10
3	157	150	7
4	138	140	-2
5	175	165	10
6	166	160	6
Total: 42			

Using the information given, $\bar{d} = 7$ and $s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}} = 4.82$. For a 95% confidence level, the appropriate t value with d.f.=5 is 2.571. Using the formula: $\bar{d} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$, we find:

$$-1.94 < \mu_d < 12.06$$

Confidence Interval Estimation of the Difference Between Two Normal Population Means, Case-2 Independent Samples

Our aim is to form a confidence interval for the difference between two population means, $\mu_x - \mu_y$. Independence here means that sample selected from one population has no effect on the sample selected from the other population. As before, there are two possible cases:

1. When σ_x^2 and σ_y^2 are known: $\Rightarrow$ use $z_{\alpha/2}$ in confidence interval estimation.
2. When σ_x^2 and σ_y^2 are unknown: $\Rightarrow$ use $t_{n-1, \alpha/2}$ in confidence interval estimation.
 - (a) When σ_x^2 and σ_y^2 are assumed equal.
 - (b) When σ_x^2 and σ_y^2 are assumed unequal.

Case 1: When the σ_x^2 and σ_y^2 are known: the random variable:

$$z = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

has a standard normal distribution and A $(1 - \alpha)100\%$ confidence interval for $\mu_x - \mu_y$ is given by:

$$\bar{x} - \bar{y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Example 4.11

A researcher wants to compare the mean wages of workers in Ankara and Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variance of wages in Ankara and Istanbul are known to be 640000 and 810000, respectively. Calculate/estimate a 95% confidence interval for the difference of (population) means of wages in Ankara and Istanbul.

Solution

$$\begin{array}{lll} n_A = 49, & \bar{x}_A = 6000, & \sigma_A^2 = 640000 \\ n_I = 81, & \bar{x}_I = 7000, & \sigma_I^2 = 810000 \\ \text{Confidence level} = 95\% & \Rightarrow z_{0.025} = 1.96 & \end{array}$$

The point estimate for the difference in means is:

$$\bar{x}_A - \bar{x}_I = 6000 - 7000 = -1000$$

The standard error for the difference is:

$$SE = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_I^2}{n_I}} = \sqrt{\frac{640000}{49} + \frac{810000}{81}} = \sqrt{13061.2245 + 10000} = \sqrt{23061.2245} = 151.86$$

The margin of error is:

$$ME = z_{0.025} \times SE = 1.96 \times 151.86 = 297.65$$

The 95% confidence interval is:

$$(\bar{x}_A - \bar{x}_I) \pm ME = -1000 \pm 297.65 = (-1297.65, -702.35)$$

Therefore, the 95% confidence interval for the difference in mean wages (Ankara - Istanbul) is:

$$\boxed{(-1297.65, -702.35)}$$

Case 2a: When the σ_x^2 and σ_y^2 are unknown, but assumed equal, then a $(1 - \alpha)100\%$ confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm t_{n_x+n_y-2, 1-\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

In our formulation:

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

and

$$s_x^2 = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2}{n_x - 1}, \quad s_y^2 = \frac{\sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_y - 1}$$

Case 2b: When the σ_x^2 and σ_y^2 are unknown, and assumed unequal, then a $(1 - \alpha)100\%$ confidence interval for $\mu_x - \mu_y$ is:

$$\bar{x} - \bar{y} \pm t_{v, 1 - \frac{\alpha}{2}} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

In our formulation:

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

Notice that, if $n_x = n_y = n$:

$$v = \left(1 + \frac{2}{\frac{s_x^2}{s_y^2} + \frac{s_y^2}{s_x^2}}\right) (n - 1)$$

Example 4.12

A researcher wants to compare the mean wages of workers in Ankara and Istanbul. She has the following data and information: Mean wage rate of 49 workers from Ankara is 6000. Mean wage rate of 81 workers from Istanbul is 7000. Population variances of wages in Ankara and Istanbul are unknown and they are assumed to be unequal. Sample variance of wages in Ankara and Istanbul are calculated as 490000 and 640000, respectively. Calculate/estimate a 95% confidence interval for the difference of (population) means of wages in Ankara and Istanbul.

Solution

$$\begin{array}{lll} n_A = 49, & \bar{x}_A = 6000, & s_A^2 = 490000 \\ n_I = 81, & \bar{x}_I = 7000, & s_I^2 = 640000 \end{array}$$

Confidence level = 95%

The point estimate for the difference in means is:

$$\bar{x}_A - \bar{x}_I = 6000 - 7000 = -1000$$

$$v = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_I^2}{n_I}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A-1} + \frac{\left(\frac{s_I^2}{n_I}\right)^2}{n_I-1}} = \frac{\left(\frac{490000}{49} + \frac{640000}{81}\right)^2}{\frac{\left(\frac{490000}{49}\right)^2}{48} + \frac{\left(\frac{640000}{81}\right)^2}{80}}$$

Therefore:

$$v = \frac{320454200}{2863577.1} = 111.91 \approx 112$$

Using $t_{112,0.975} \approx 1.981$ (approximately the same as $z_{0.975} = 1.96$ for large degrees of freedom):

The standard error is:

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_I^2}{n_I}} = \sqrt{10000 + 7901.2346} = \sqrt{17901.2346} = 133.78$$

The margin of error is:

$$ME = t_{v,0.975} \times SE = 1.981 \times 133.78 = 265.02$$

The 95% confidence interval is:

$$(\bar{x}_A - \bar{x}_I) \pm ME = -1000 \pm 265.02 = (-1265.02, -734.98)$$

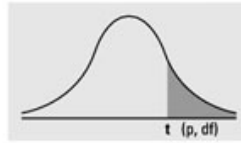
Therefore, the 95% confidence interval for the difference in mean wages (Ankara - Istanbul) is:

$$\boxed{(-1265.02, -734.98)}$$

Appendix

Figure 9: Student's T-Distribution Table

Numbers in each row of the table are values on a t -distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p) .



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

Figure 10: Z Table

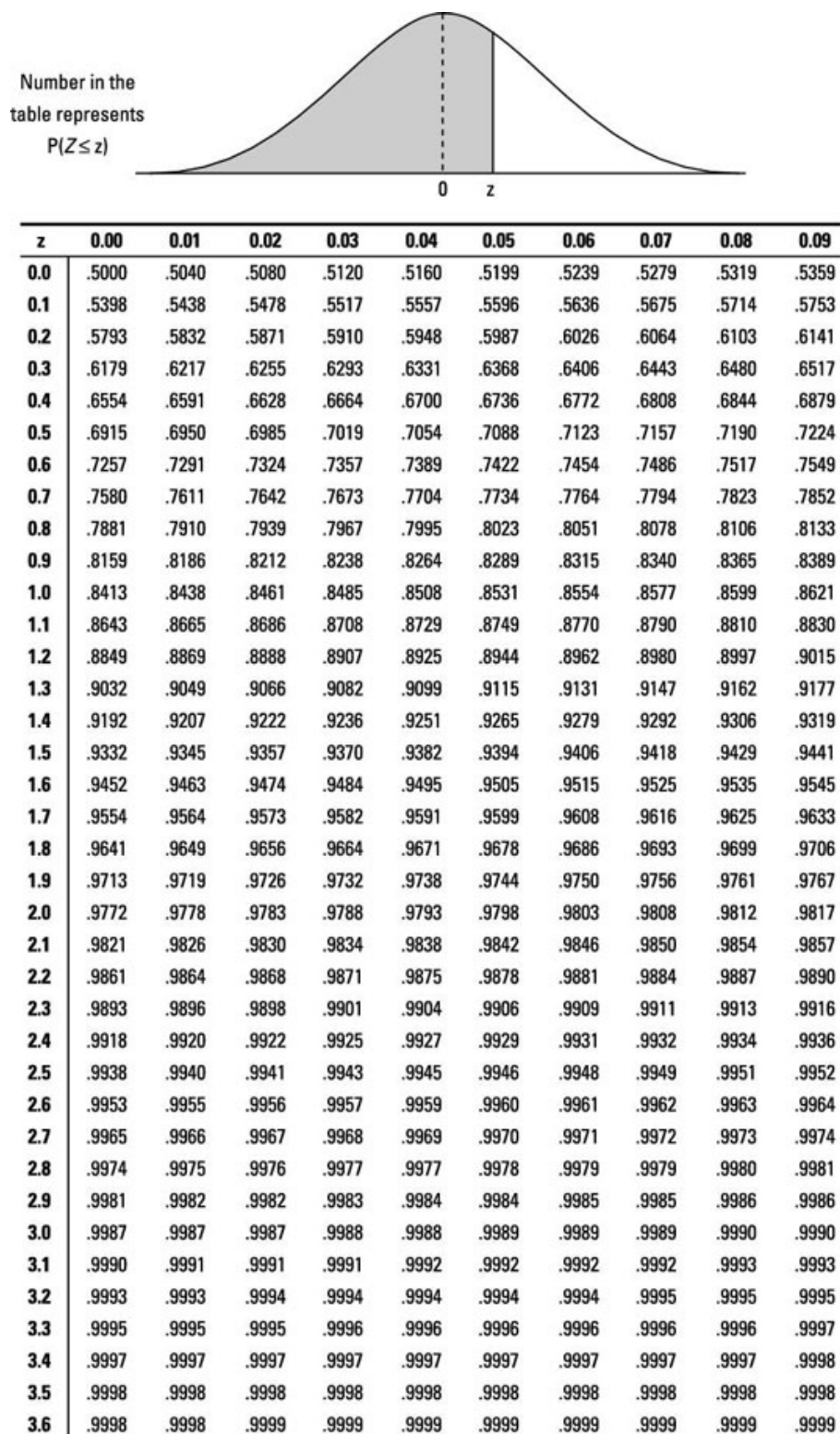


Figure 11: Chi-Square Distribution Table (Right-tail probabilities)

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-----	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

5 Hypothesis Testing

The objective of statistics is to make **inferences** about unknown population parameters based on sample data. These inferences come in two forms:

- **Estimation** of parameters
- **Hypothesis testing** about parameter values

Hypothesis testing is the foundation of scientific progress. Scientists observe events or nature, collect sample data, formulate a theory (state a hypothesis about parameters) and then test theory against observation. So far, we learned about estimation of parameters and constructing a confidence interval which contains an unknown population parameter with a specified probability $(1 - \alpha) \times 100\%$.

In contrast, hypothesis testing evaluates whether a proposed or hypothesized value is plausible for that unknown parameter. This process formally examines the validity of a claim about the parameter. A statistical hypothesis is a precise statement regarding the numerical value of a population parameter.

There are four essential components of a statistical test:

1. **Null hypothesis (H_0):** The hypothesis to be tested (often representing status quo or no effect)
2. **Alternative hypothesis (H_a):** The research hypothesis we seek to support
3. **Test statistic:** A function of sample measurements on which the decision is based
4. **Rejection region (RR):** The set of test statistic values for which we reject H_0

Decision Rule

- If test statistic $\in$ RR $\Rightarrow$ Reject H_0 ,
- If test statistic $\notin$ RR $\Rightarrow$ Fail to reject H_0

In the exact words of my dear wise friend, Eray Yücel Hoca, there are two deadly sins with respect to Hypothesis Testing

1. **Never Say “Accept H_0 ”** When H_a is not sufficiently supported by the data, use one of these phrases: i) *Fail to reject H_0* *Insufficient evidence to reject H_0*
2. **Never Include H_a in your concluding statement** Avoid creative wording when making technical statistical statements.

Confidence intervals and hypothesis tests are two sides of the same coin.

Consider testing:

$$H_0 : \mu = 2 \quad \text{vs} \quad H_1 : \mu \neq 2$$

If $\mu = 2$ falls within a 95% confidence interval, then we **cannot reject** $H_0 : \mu = 2$ at $\alpha = 0.05$. Otherwise, we reject H_0 .

This illustrates an important principle: any value of μ contained in a 95% confidence interval (if it were our null hypothesis) *cannot* be rejected at the 5% significance level based on that sample.

This is precisely why we say “do not reject H_0 ” rather than “accept H_0 ”—because multiple plausible parameter values may be consistent with the data.

Types of Errors

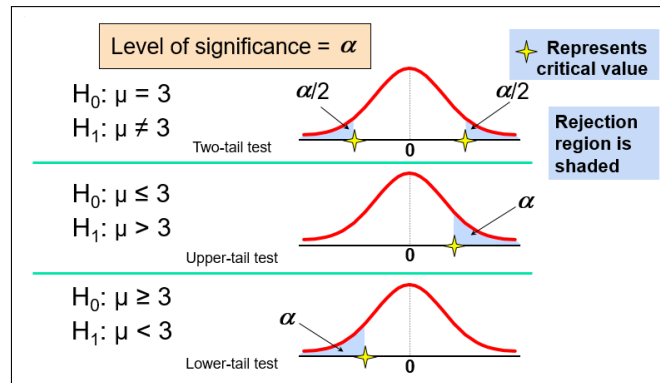
	Accept H_0	Reject H_0
H_0 True	Correct decision	Type I error
H_a True	Type II error	Correct decision

- **Type I error:** Rejecting H_0 when it is true
 - Probability denoted by α (level of significance)
- **Type II error:** Accepting H_0 when H_a is true
 - Probability denoted by β

The significance level, α defines the unlikely values of the sample statistic if the null hypothesis is true. It defines the rejection region of the sampling distribution. Typical values for α are 0.01, 0.05, or 0.10 (Recall that the typical confidence levels, i.e. $(1 - \alpha)100\%$ are 99, 95 and 90 percent in confidence interval estimation). Level of significance, α is selected by the researcher at the beginning of the hypothesis testing and provides the critical value(s) of the test (z or t or chi-square).

Hypothesis testing can be two-sided or one sided.

Figure 12: Level of Significance and the Rejection Region



Source: Newborne, Carlston and Thorne. Statistics for Business and Economics, 2013 Pearson Education

5.1 Hypothesis testing for the mean of a normal population: known population variance

Consider the one-sided hypothesis test:

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

with $\{x_i\}_{i=1}^n$, a random sample of n observations from a normal population $N(\mu, \sigma^2)$ with known σ^2 .

Given a statistical significance level α : H_0 is rejected if

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

and we fail to reject H_0 otherwise. The alternative rule is to reject H_0 if:

$$\bar{x} > \mu_0 + z_\alpha \sigma / \sqrt{n}$$

There is a p-value approach to hypothesis testing as well. The **p-value** is the probability of obtaining a test statistic as extreme as, or more extreme than, the observed sample value, *given that H_0 is true*. Alternatively, it is called the **observed level of significance**. It represents the **smallest value of α** for which H_0 can be rejected based on the observed data. What we need to do is to convert sample result (such as, $\bar{x}$) to test statistic (e.g., z statistic) and then obtain the p-value. For an upper tail test:

$$\begin{aligned} p\text{-value} &= P \left(z > \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \mid H_0 \text{ is true} \right) \\ &= P \left(z > \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0 \right) \end{aligned}$$

The decision rule is as follows:

If p-value $< \alpha$, reject H_0

If p-value $\geq \alpha$, do not reject H_0 .

Example 5.1.

A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over USD 52 per month. The company wishes to test this claim. The manager selects a sample with size 64 and calculates $\bar{x} = 53.1$. Test this hypothesis assuming $\sigma = 10$ is known and $\alpha = 0.10$.

Solution:

Our hypothesis test is constructed as follows:

$$H_0 : \mu \leq 52 \quad (\text{the average is not over \$52 per month})$$

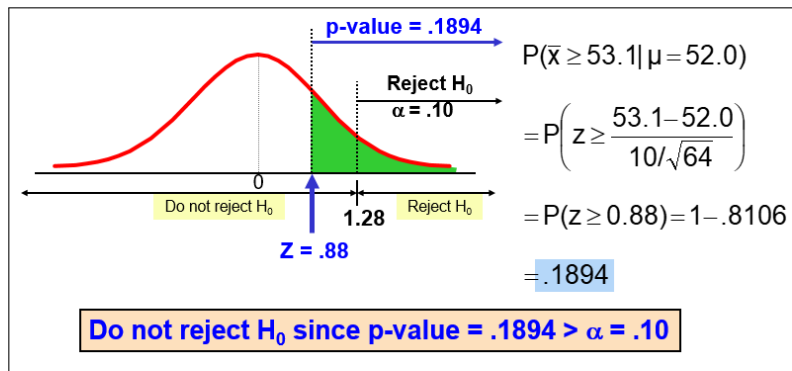
$$H_1 : \mu > 52 \quad (\text{the average is greater than \$52 per month})$$

(i.e., sufficient evidence exists to support the manager's claim)

Since $p\text{-value} = 0.1894 > 0.10 = \alpha$, we **fail to reject** H_0 .

There is insufficient statistical evidence at the 10% significance level to support the manager's claim that average monthly cell phone bills exceed \$52. The observed sample mean of \$53.10 is not significantly greater than \$52 given the sample size and population variability.

Figure 14: P-Value Approach



Confidence Interval Perspective A 90% confidence interval for μ is:

$$\bar{x} \pm z_{0.95} \cdot \frac{\sigma}{\sqrt{n}} = 53.1 \pm 1.645 \cdot \frac{10}{\sqrt{64}} = 53.1 \pm 1.645 \cdot 1.25 = 53.1 \pm 2.05625$$

$$(51.04, 55.16) \quad \text{or approximately} \quad (51.04, 55.16)$$

Since $\mu_0 = 52$ falls within this 90% confidence interval, we cannot reject $H_0 : \mu = 52$ at the 10% significance level.

Example 5.2

Test the claim that the true mean number of TV sets in US homes is equal to 3. Assume that $\sigma = 0.8$, a sample of size $n = 100$ is taken, and $\bar{x}$ is calculated as 2.84. Use a significance level of $\alpha = 0.05$ to test this claim.

Solution

Since we are testing if the mean is *equal* to 3, this is a two-tailed test:

$$H_0 : \mu = 3 \quad (\text{the mean number of TV sets is equal to 3})$$

$$H_1 : \mu \neq 3 \quad (\text{the mean number of TV sets is not equal to 3})$$

The following information is given:

Hypothesized mean: $\mu_0 = 3$

Sample size: $n = 100$

Sample mean: $\bar{x} = 2.84$

Population standard deviation: $\sigma = 0.8$

Significance level: $\alpha = 0.05$

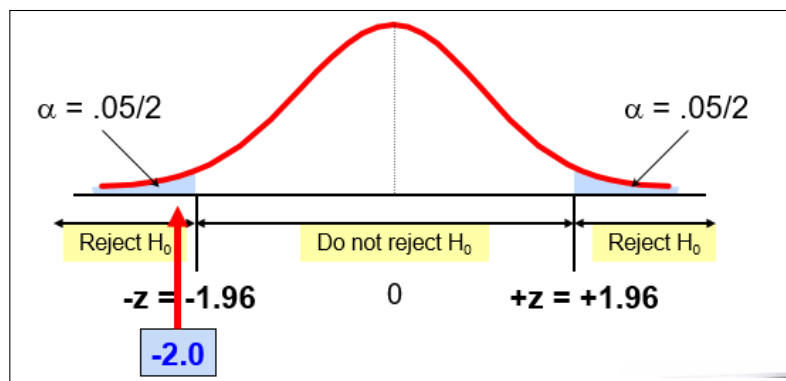
Critical values: $z_{\alpha/2} = z_{0.025} = -1.96$, $z_{1-\alpha/2} = z_{0.975} = 1.96$

Calculate the test statistics:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.84 - 3}{0.8/\sqrt{100}} = \frac{-0.16}{0.08} = -2.0$$

Since $z = -2.0 < -1.96$, we **reject** H_0 .

Figure 15: Two-tailed Test



P-value Approach

How likely is it to see a sample mean of 2.84 (or something further from the mean, in either direction) if the true mean is $\mu = 3.0$?

For a two-tailed test:

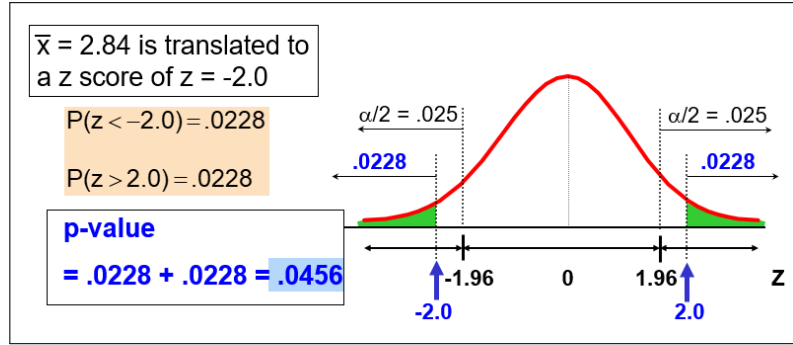
$$p\text{-value} = 2 \cdot P(Z < -2.0) = 2 \times 0.0228 = 0.0456$$

Since $p\text{-value} = 0.0456 < 0.05 = \alpha$, we **reject** H_0 .

Critical Sample Mean Boundaries The rejection region boundaries for $\bar{x}$ are:

$$\mu_0 \pm z_{0.975} \cdot \frac{\sigma}{\sqrt{n}} = 3 \pm 1.96 \times 0.08 = 3 \pm 0.1568$$

Figure 16: P-Value Approach



$$(2.8432, 3.1568)$$

Since $\bar{x} = 2.84 < 2.8432$, it falls in the rejection region.

95% Confidence Interval A 95% confidence interval for μ is:

$$\bar{x} \pm z_{0.975} \cdot \frac{\sigma}{\sqrt{n}} = 2.84 \pm 1.96 \times 0.08 = 2.84 \pm 0.1568$$

$$(2.6832, 2.9968)$$

1. The test statistic $z = -2.0$ is less than the critical value -1.96 .
2. The p -value of 0.0456 is less than $\alpha = 0.05$.
3. The 95% confidence interval $(2.6832, 2.9968)$ does **not** contain $\mu_0 = 3$.
4. The sample mean $\bar{x} = 2.84$ is below the lower critical boundary of 2.8432.

All four methods lead to the same conclusion. There is sufficient statistical evidence at the 5% significance level to reject the claim that the true mean number of TV sets in US homes is equal to 3. Based on the sample of 100 homes with mean 2.84, we conclude that the true mean is significantly different from 3.

5.2 Hypothesis testing for the mean of a normal population: Unknown population variance

Consider the one-sided hypothesis test:

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0$$

Let $\{x_i\}_{i=1}^n$ be a random sample of n observations from a normal population $N(\mu, \sigma^2)$, where σ^2 is unknown.

At the statistical significance level α :

- H_0 is rejected if

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{n-1, 1-\alpha}$$

- and we fail to reject H_0 otherwise.

where:

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean
- $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ is the sample standard deviation
- $t_{n-1, 1-\alpha}$ is the $(1 - \alpha)$ -quantile of the t -distribution with $n - 1$ degrees of freedom

Example 5.3

The average cost of a hotel room in Chicago is said to be \$168 per night. A random sample of 25 hotels resulted in $\bar{x} = \$172.50$ and $s = \$15.40$. Test this claim at the $\alpha = 0.05$ significance level.

Solution

We are testing if the average cost is equal to \$168, which is a two-tailed test:

$$H_0 : \mu = 168 \quad (\text{the average cost is \$168 per night})$$

$$H_a : \mu \neq 168 \quad (\text{the average cost is different from \$168 per night})$$

Given Information:

Hypothesized mean: $\mu_0 = 168$

Sample size: $n = 25$

Sample mean: $\bar{x} = 172.50$

Sample standard deviation: $s = 15.40$

Significance level: $\alpha = 0.05$

Degrees of freedom: $df = n - 1 = 24$

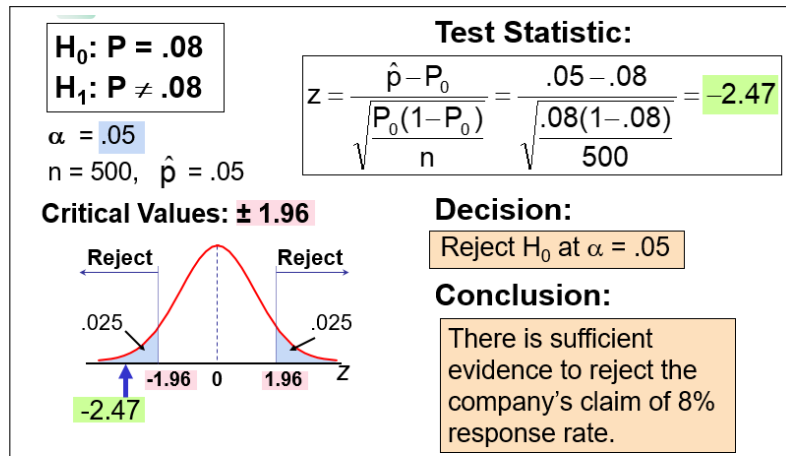
Since σ is unknown and $n < 30$, we use the t -test:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{172.50 - 168}{15.40/\sqrt{25}} = \frac{4.50}{15.40/5} = \frac{4.50}{3.08} = 1.461$$

For a two-tailed t -test with $df = 24$ and $\alpha = 0.05$:

$$t_{\alpha/2, df} = t_{0.025, 24} = -2.064 \quad \text{and} \quad t_{1-\alpha/2, df} = t_{0.975, 24} = 2.064$$

Figure 17: Hypothesis Testing for Population Proportion



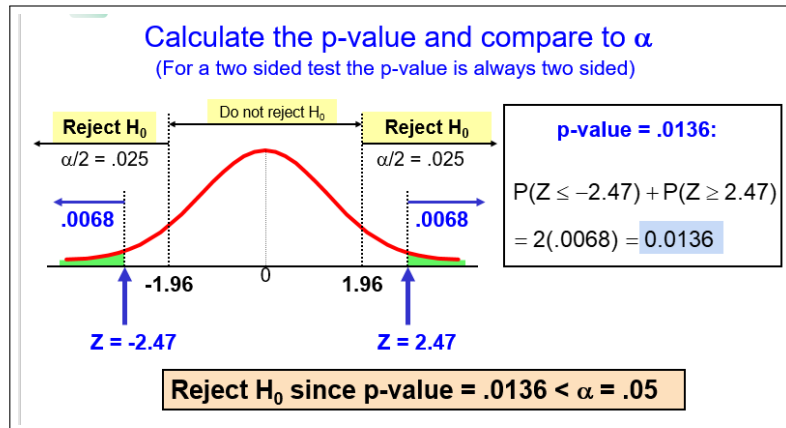
Since $|t| = 1.461 < 2.064$, we **fail to reject** H_0 .

P-value Approach For a two-tailed t -test with $df = 24$:

$$p\text{-value} = 2 \cdot P(T_{24} > 1.461) \approx 2 \times 0.0785 = 0.1570$$

Since $p\text{-value} = 0.1570 > 0.05 = \alpha$, we **fail to reject** H_0 .

Figure 18: P-Value for Hypothesis Testing for Population Proportion



95% Confidence Interval A 95% confidence interval for μ is:

$$\bar{x} \pm t_{0.975,24} \cdot \frac{s}{\sqrt{n}} = 172.50 \pm 2.064 \times \frac{15.40}{\sqrt{25}} = 172.50 \pm 2.064 \times 3.08$$

$$172.50 \pm 6.357 = (166.143, 178.857)$$

1. The test statistic $t = 1.461$ falls between the critical values -2.064 and 2.064 .

2. The p -value of approximately 0.1570 is greater than $\alpha = 0.05$.
3. The 95% confidence interval (166.143, 178.857) contains $\mu_0 = 168$.

All three approaches lead to the same conclusion.

5.3 Hypothesis testing for population proportion

Hypothesis testing for population proportion involves categorical variables. There are two possible outcomes; “Success” (a certain characteristic is present) “Failure” (the characteristic is not present). Fraction or proportion of the population in the “success” category is denoted by P . We assume the sample size is large.

The sample proportion in the success category is denoted by $\hat{p}$ where

$$\hat{p} = \frac{\text{Number of successes in sample}}{\text{Sample size}}$$

When $nP(1-P) > 5$, $\hat{p}$ can be approximated by a normal distribution with mean and standard deviation $\mu_{\hat{p}}$ and $\frac{\sqrt{P(1-P)}}{\sqrt{n}}$.

Consider the one-sided hypothesis test for a population proportion:

$$H_0 : p \leq p_0 \quad \text{vs} \quad H_1 : p > p_0$$

Let $\{x_i\}_{i=1}^n$ be a random sample of n observations from a Bernoulli(p) population, where each x_i takes values 0 or 1.

At the statistical significance level α :

- H_0 is rejected if

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-\alpha}$$

- and we fail to reject H_0 otherwise.

where:

- $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample proportion
- p_0 is the hypothesized population proportion under H_0
- z_α is the $(1 - \alpha)$ -quantile of the standard normal distribution
- The test statistic uses p_0 in the denominator because the test is conducted under the assumption that H_0 is true

5.4 Example 5.4.

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 people were surveyed with 25 responses. Test at the $\alpha = 0.05$ significance level.

Solution

We are testing if the response rate is equal to 8%, which is a two-tailed test:

$$H_0 : p = 0.08 \quad (\text{the response rate is 8\%})$$

$$H_1 : p \neq 0.08 \quad (\text{the response rate is different from 8\%})$$

Given Information:

Hypothesized proportion: $p_0 = 0.08$

Sample size: $n = 500$

Number of successes: $x = 25$

$$\text{Sample proportion: } \hat{p} = \frac{x}{n} = \frac{25}{500} = 0.05$$

Significance level: $\alpha = 0.05$

$$\text{Critical values: } z_{\alpha/2} = z_{0.025} = -1.96, \quad z_{1-\alpha/2} = z_{0.975} = 1.96$$

Calculate the test statistics: For testing a proportion with large n , we use the z -test:

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.05 - 0.08}{\sqrt{\frac{0.08 \times 0.92}{500}}} = \frac{-0.03}{\sqrt{\frac{0.0736}{500}}} \\ &= \frac{-0.03}{\sqrt{0.0001472}} = \frac{-0.03}{0.01213} = -2.473 \end{aligned}$$

Check Conditions for Normal Approximation:

$$\bullet \quad np_0(1 - p_0) = 500 \times 0.08 \times 0.92 = 36.8 > 5$$

Decision Using Critical Value: Since $|z| = 2.473 > 1.96$, we **reject** H_0 .

P-value Approach For a two-tailed test:

$$p\text{-value} = 2 \cdot P(Z < -2.473) = 2 \times 0.0067 = 0.0134$$

Since $p\text{-value} = 0.0134 < 0.05 = \alpha$, we **reject** H_0 .

5.5 Power of a Test

The power of a hypothesis test is its ability to correctly reject a false null hypothesis—essentially, the probability of detecting an effect when that effect truly exists. Imagine you're searching for a faint

star in a dark sky using a telescope. A more powerful test is like using a telescope with a better lens and less distortion; it's more likely to reveal the star if it's actually there, rather than missing it and concluding the patch of sky is empty. In statistical terms, higher power means you're less likely to make a Type II error (failing to reject a false null), and it increases with larger effect sizes, larger sample sizes, or by allowing a higher significance level—all of which sharpen your capacity to identify a real signal amid the noise.

As you may recall, a Type I error occurs if the researcher rejects the null hypothesis in favor of the alternative hypothesis when, in fact, the null hypothesis is true. The probability of Type I error is denoted as α . A Type II error, on the other hand, occurs if the researcher fails to reject the null hypothesis when, in fact, the null hypothesis is false. The probability of Type II error is denoted as β . So,

$$\begin{aligned}P(\text{Reject } H_0 \mid H_0 \text{ true}) &= \alpha \\P(\text{Fail to reject } H_0 \mid H_0 \text{ true}) &= 1 - \alpha \\P(\text{Fail to reject } H_0 \mid H_0 \text{ false}) &= \beta \\ \text{Power} &= 1 - \beta = P(\text{Reject } H_0 \mid H_0 \text{ false})\end{aligned}$$

In sum, power is the ability of a test to reject a *false* null hypothesis.

Assume the population is normal and the population variance is known. Consider the test:

$$\begin{aligned}H_0 : \mu &= \mu_0 \\H_a : \mu &\neq \mu_0\end{aligned}$$

$$\text{Reject } H_0 \text{ if } z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

or,

$$\text{Reject } H_0 \text{ if } \bar{X} > \bar{X}_c = \mu_0 + z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

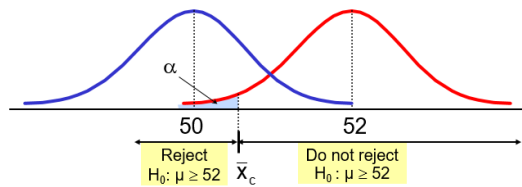
If the null hypothesis is false and the true mean is μ^* , then the probability of type II error is:

$$\beta = P(\bar{X} < \bar{X}_c \mid H_a \text{ true}) = P\left(z < \frac{\bar{X}_c - \mu^*}{\sigma/\sqrt{n}}\right)$$

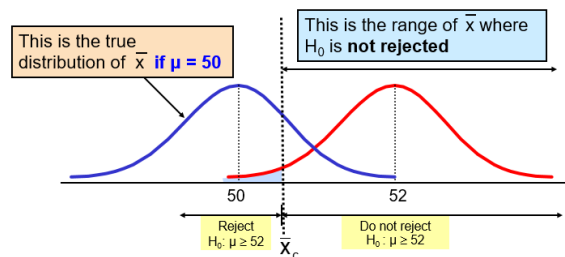
Let's go through an example, suppose we are testing the hypothesis that true population mean is higher than 52 at the 5 percent significance level. Assume that sample size is 64 and population standard deviation is known to be 6. Here are the visual steps for calculating the power of this test:

- Type II error is the probability of failing to reject a false H_0

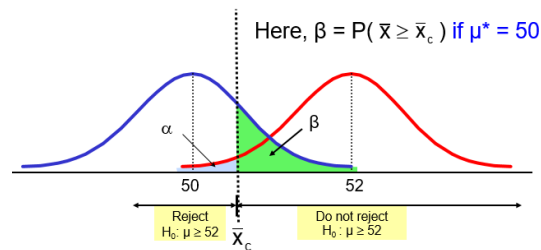
Suppose we fail to reject $H_0: \mu \geq 52$ when in fact the true mean is $\mu^* = 50$



- Suppose we do not reject $H_0: \mu \geq 52$ when in fact the true mean is $\mu^* = 50$



- Suppose we do not reject $H_0: \mu \geq 52$ when in fact the true mean is $\mu^* = 50$

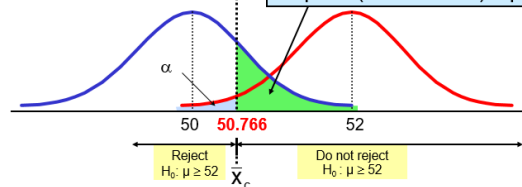


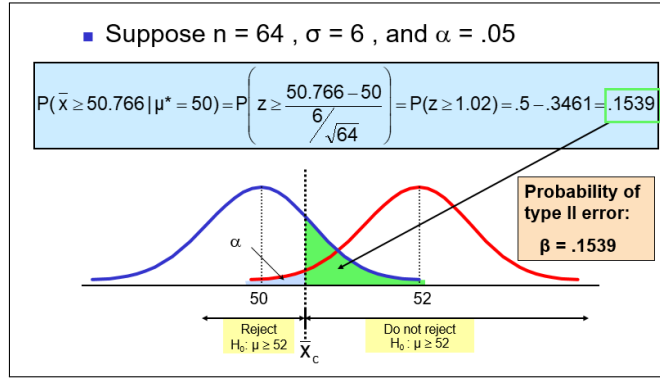
- Suppose $n = 64$, $\sigma = 6$, and $\alpha = .05$

$$\bar{x}_c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} = 52 - 1.645 \frac{6}{\sqrt{64}} = 50.766$$

(for $H_0: \mu \geq 52$)

So $\beta = P(\bar{x} \geq 50.766)$ if $\mu^* = 50$





5.6 Tests of the Variance of a Normal Distribution

Consider

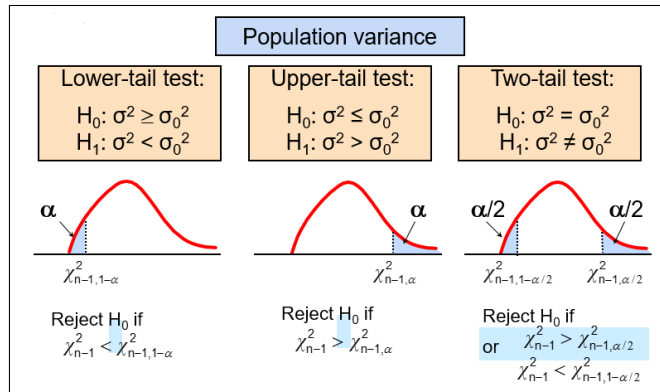
$$H_0: \sigma^2 \leq \sigma_0^2 \quad \text{vs} \quad H_1: \sigma^2 > \sigma_0^2$$

and $\{x_i\}_{i=1}^n$, a random sample of n observations from a normal population $N(\mu, \sigma^2)$.

At the statistical significance level α , H_0 is rejected if

$$\frac{(n-1)s^2}{\sigma_0^2} > \chi_{n-1, 1-\alpha}^2$$

and we fail to reject H_0 otherwise.



Example 5.6

A process engineer is concerned with the variation of temperature in an industrial furnace and wonders if it exceeds 1500. She collects a random sample of temperatures as:

975, 1075, 1050, 900, 1000, 950, 1025, 1050, 975 °C

Conduct and conclude the relevant hypothesis test at the significance level of 5%.

$$H_0: \sigma^2 \leq 1500 \quad \text{vs} \quad H_1: \sigma^2 > 1500$$

Solution:

The relevant distribution is χ^2 and the degrees of freedom is 8.

Since:

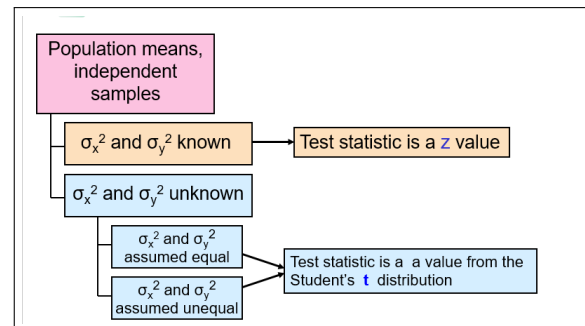
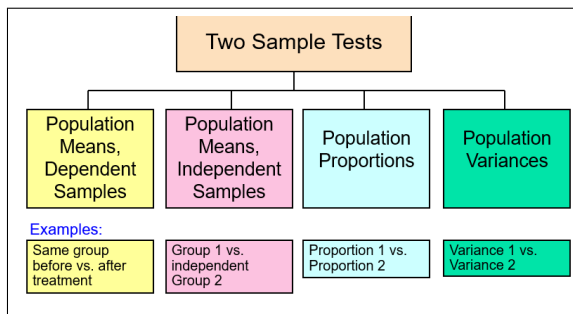
$$\frac{(9-1) \cdot 3125}{1500} = 16.667 > 15.507$$

we reject H_0 at $\alpha = 0.05$.

5.7 Hypothesis testing: Two populations

The logic of the hypothesis testing is very similar to the confidence interval estimation. And this applies to the case with two populations as well. Recall that earlier, we learned about different cases when we had two populations;

1. Dependent Samples
2. Independent Samples
 - (a) Variance known
 - (b) Variance unknown
 - i. variance of each population assumed equal
 - ii. variance of each population assumed unequal



Paired t-Test Formulas

Let $t = \frac{\bar{d}}{s_d/\sqrt{n}}$ be the test statistic.

- **Right-tailed test:** $H_0 : \mu_d \leq 0, H_1 : \mu_d > 0$
Reject H_0 if $t > t_{n-1, 1-\alpha}$
- **Left-tailed test:** $H_0 : \mu_d \geq 0, H_1 : \mu_d < 0$
Reject H_0 if $t < t_{n-1, \alpha}$

- **Two-tailed test:** $H_0 : \mu_d = 0$, $H_1 : \mu_d \neq 0$
Reject H_0 if $|t| > t_{n-1, 1-\alpha/2}$

Where:

- n = number of matched pairs
- $\bar{d}$ = mean of the differences
- s_d = standard deviation of the differences
- $t_{n-1, 1-\alpha}$ = critical value from t-distribution with $n - 1$ df

Example 5.7

Assume you send your salespeople to a "customer service" training workshop. Has the training made a difference in the number of complaints? You collect the following data:

Salesperson	Before (1)	After (2)	Difference, $d_i = (2) - (1)$
C.B.	6	4	-2
T.F.	20	6	-14
M.H.	3	2	-1
R.K.	0	0	0
M.O.	4	0	-4
Total			-21

Test this at $\alpha = 0.05$.

Solution

For a two-tailed test (testing if there's **any** difference):

$$H_0 : \mu_d = 0 \quad (\text{no difference in complaints})$$

$$H_1 : \mu_d \neq 0 \quad (\text{training changed complaint numbers})$$

The table shows $d_i = \text{After} - \text{Before}$, but for our test we want:

$$d'_i = \text{Before} - \text{After} = -d_i$$

So we'll work with:

$$d'_i = 2, 14, 1, 0, 4$$

Then $\sum d'_i = 21$.

Calculate Sample Statistics

$$\begin{aligned}n &= 5 \\ \bar{d} &= \frac{21}{5} = 4.2 \\ \sum (d'_i)^2 &= 2^2 + 14^2 + 1^2 + 0^2 + 4^2 = 4 + 196 + 1 + 0 + 16 = 217 \\ s_d^2 &= \frac{\sum (d'_i)^2 - \frac{(\sum d'_i)^2}{n}}{n-1} = \frac{217 - \frac{441}{5}}{4} = \frac{217 - 88.2}{4} = \frac{128.8}{4} = 32.2 \\ s_d &= \sqrt{32.2} \approx 5.674\end{aligned}$$

Test Statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{4.2}{5.674/\sqrt{5}} = \frac{4.2}{2.537} \approx 1.655$$

Degrees of freedom: $df = n - 1 = 4$.

Two-tailed test at $\alpha = 0.05$, so $\alpha/2 = 0.025$ in each tail, $df = 4$:

$$t_{4,0.975} \approx 2.776 \quad \text{and} \quad t_{4,0.025} \approx -2.776$$

(from t-table; since t-distribution is symmetric, $t_{4,0.025} = -t_{4,0.975}$)

Reject H_0 if $|t| > t_{critical}$ or if $t < -2.776$ or $t > 2.776$.

$$|t| = |1.655| = 1.655 < 2.776 \Rightarrow \text{Fail to reject } H_0.$$

5.8 Difference between two normal population means Case: Independent samples and Known population variances

Consider:

$$H_0 : \mu_x - \mu_y \leq 0 \quad \text{vs} \quad H_1 : \mu_x - \mu_y > 0$$

Let:

$$\begin{aligned}\{x_i\}_{i=1}^{n_x} &\subset \{x_i\}_{i=1}^{N_x} \sim N(\mu_x, \sigma_x^2) \\ \{y_i\}_{i=1}^{n_y} &\subset \{y_i\}_{i=1}^{N_y} \sim N(\mu_y, \sigma_y^2)\end{aligned}$$

where σ_x^2 and σ_y^2 are **known**.

Right-Tailed Test For:

$$H_0 : \mu_x - \mu_y \leq 0 \quad \text{vs} \quad H_1 : \mu_x - \mu_y > 0$$

At the statistical significance level α :

- H_0 is rejected if

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} > z_{1-\alpha}$$

- and we fail to reject H_0 otherwise.

Left-Tailed Test For:

$$H_0 : \mu_x - \mu_y \geq 0 \quad \text{vs} \quad H_1 : \mu_x - \mu_y < 0$$

At the statistical significance level α :

- H_0 is rejected if

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} < z_\alpha$$

- and we fail to reject H_0 otherwise.

Two-Tailed Test For:

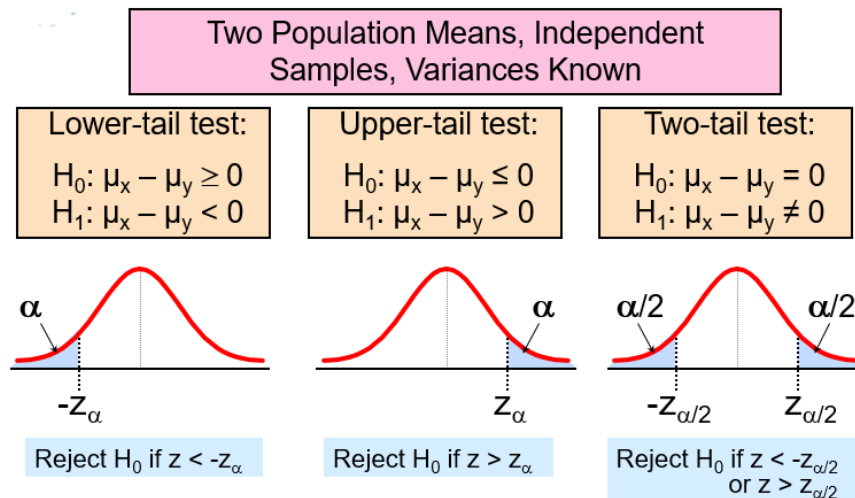
$$H_0 : \mu_x - \mu_y = 0 \quad \text{vs} \quad H_1 : \mu_x - \mu_y \neq 0$$

At the statistical significance level α :

- H_0 is rejected if

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} < z_{\alpha/2} \quad \text{or} \quad \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} > z_{1-\alpha/2}$$

- and we fail to reject H_0 otherwise.



Example 5.8

A researcher wonders if the mean wage of workers in Ankara falls short of that in Istanbul. She has the following data and information:

- Mean wage rate of 49 workers from Ankara is 6000

- Mean wage rate of 81 workers from Istanbul is 7000
- Population variance of wages in Ankara: 640000
- Population variance of wages in Istanbul: 810000

Conduct and conclude a hypothesis test at the significance level of 5% to assess if mean wage rate in Ankara is less than the mean wage rate in Istanbul.

Solution

Let:

- μ_x = mean wage in Ankara
- μ_y = mean wage in Istanbul
- $n_x = 49$, $\bar{x} = 6000$, $\sigma_x^2 = 640000$
- $n_y = 81$, $\bar{y} = 7000$, $\sigma_y^2 = 810000$

We test if Ankara wages are **less than** Istanbul wages:

$$H_0 : \mu_x - \mu_y \geq 0 \quad \text{vs} \quad H_1 : \mu_x - \mu_y < 0$$

Since population variances are known, we use the **z-test** (normal distribution).

4. Test statistic calculation

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = \frac{6000 - 7000}{\sqrt{\frac{640000}{49} + \frac{810000}{81}}}$$

Calculate denominator:

$$\begin{aligned} \frac{\sigma_x^2}{n_x} &= \frac{640000}{49} \approx 13061.2245 \\ \frac{\sigma_y^2}{n_y} &= \frac{810000}{81} = 10000 \\ \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} &= \sqrt{13061.2245 + 10000} = \sqrt{23061.2245} \approx 151.86 \end{aligned}$$

Thus:

$$z = \frac{-1000}{151.86} \approx -6.585$$

Left-tailed test at $\alpha = 0.05$:

$$z_{\alpha} = z_{0.05} = -1.645$$

Reject H_0 if $z < z_{\alpha}$.

Since:

$$-6.585 < -1.645$$

we reject H_0 at $\alpha = 0.05$.

6 Linear Regression Analysis

Recall that previously, we learned to find point estimators using the Method of Moment and Maximum Likelihood Techniques. We are now ready to learn about one of the most commonly used methods in economics, namely the Least Squares (LS) Estimator. This method relies on fitting a curve to the data points at hand and trying to find the best fit by minimizing the distance between the actual data points and the curve that we fit.

Most of the times, the aim of this exercise is to understand the relationship between two economic variables. We have seen this before when we learned how to estimate the covariance or the correlation coefficient between, say X and Y . The linear regression analysis also gives us this correlation, but this time the way we model the correlation implies that X might have a causal effect on Y . We call X the *independent* or *explanatory variable* (because it explains the variation in Y) and Y the *dependent variable*.

We model the relationship between dependent and independent variable(s) as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

However, a minimalist design is also possible:

$$y = \beta_0 + \beta_1 x$$

Even this may be a good enough model:

$$y = \beta_0$$

The process of inference begins with the specification of an economic model. Then a statistical model describes the sampling process that we visualize was used to produce the sample data. See the structure below:

Economic model:

$$y = \beta_0 + \beta_1 x$$

Statistical model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

The random error term (ϵ) serves three main purposes:

[label=0.] ϵ captures the combined effect of all other influences other than x . These other effects are assumed to be unobservable, otherwise they would be included in the model. ϵ captures any approximation error that arises because of the linear functional form. ϵ captures any element of random behavior present in each individual observation.

See the structures below:

Case 1: Unconditional model of mean

Economic model:

$$y = \beta_0$$

Statistical model:

$$y = \beta_0 + \epsilon$$

Case 2: Simple Linear model

Economic model:

$$y = \beta_0 + \beta_1 x$$

Statistical model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Case 3: Multiple Linear model

Economic model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Statistical model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

Example: Keynesian Consumption Model

Keynes postulated the marginal propensity to consume (MPC), suggesting that as the income increases, individuals increase their consumption, but not one to one. In other words; $0 < mpc < 1$

Keynesian Consumption Function:

• $Y = \beta_1 + \beta_2 X$ where $0 < \beta_2 < 1$

- Y =consumption expenditure, X =income and β_1 and β_2 are parameters of the model, respectively the intercept and slope coefficients.

This is a **single equation model** with one equation. If there are more than one equation, it is called a **multiple equation model**. In our example, consumption expenditure is the dependent variable and income is the independent/explanatory variable.

To allow for an inexact relationship between economic variables (why inexact?), the econometrician would modify the deterministic consumption as follows:

- $Y = \beta_1 + \beta_2 X + e$
- where e is known as the **disturbance or error term** which is a **random (stochastic)** variable with well defined probabilistic properties. The disturbance term includes everything that affects consumption but not taken into account explicitly.

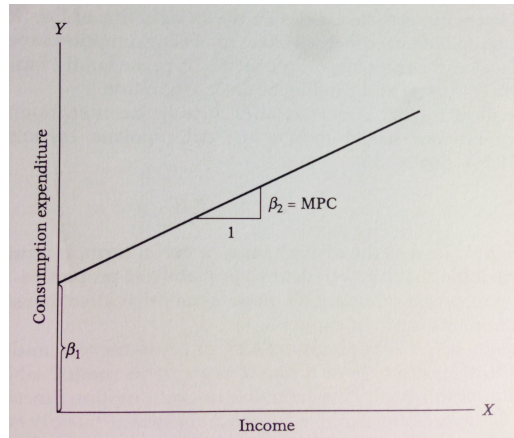
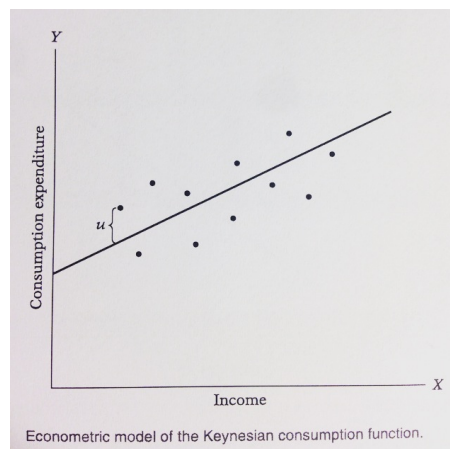


Figure 20: Keynesian Consumption Function



Obtaining Data:

Data for U.S. Economy between 1981-1996. Y variable is the aggregate consumption expenditure and X is the gross domestic product. Both measured in 1992 US dollars (they are in real terms).

DATA ON Y (PERSONAL CONSUMPTION EXPENDITURE)
AND X (GROSS DOMESTIC PRODUCT, 1982–1996), BOTH
IN 1992 BILLIONS OF DOLLARS

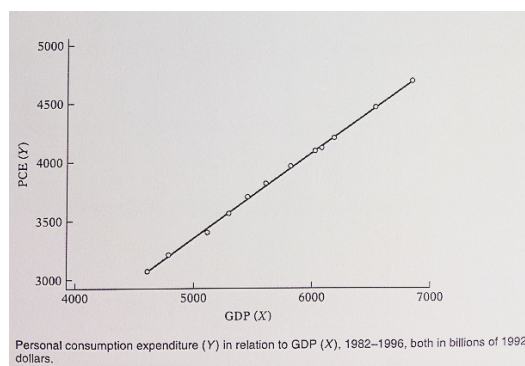
Year	Y	X
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062.0
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486.0	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4

Source: *Economic Report of the President*, 1998, Table B–2, p. 282.

Estimation of the Econometric Model:

Next task is to estimate the parameters of the consumption function. The following is the estimated consumption function:

- $\hat{Y} = -184.08 + 0.7064X_i$
- The hat on Y indicates that it is an estimate and β_1 and β_2 are estimated as -184.08 and 0.7064 accordingly.
- The estimated consumption function (i.e. regression line) is as follows:



What does this tell us?

- marginal propensity to consume was around 0.70 suggesting that for the sample period, a 1 dollar real increase in income led on average to an increase about 70 cents in real consumption.
- we say ‘**on average**’ because the relationship between consumption and income is inexact, not all the data points lie exactly on the regression line.

But how did we calculate $mpc = 0.706$? Here is how we do it:

Building and estimating a Simple Linear Regression model

Consider a variable y which we believe is explained by another variable x via a linear relationship like:

$$y = \beta_0 + \beta_1 x + e$$

In this expression,

- β_0 stands for the autonomous/unconditional component of y
- $\beta_1 x$ stands for the part of y attributable to x ; depending on the sign of β_1 , an increase in x may induce an increase or a decrease in y

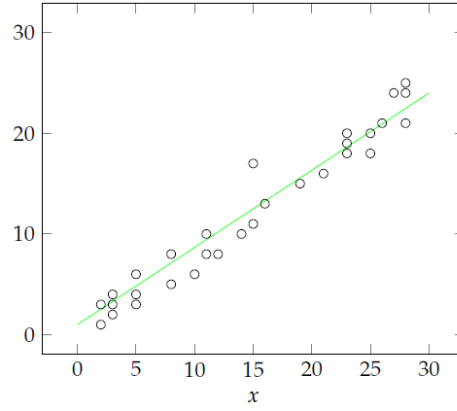


Figure 21: Simple Linear Regression Line. (Source: Karademir and Yücel)

- A case of $\beta_1 = 0$ corresponds to our unconditional model of mean

Estimating this model means estimating β_0 (the intercept) and β_1 (the slope) because these two parameters characterize the relationship between x and y . Hence, we have two unknown parameters that we would like to estimate. The Least squares method tries to find the "best estimates" of β_0 and β_1 such that the distance between our *predicted* and the *actual* y is minimized. Since we used the word "minimized", this can actually be considered as an optimization problem where we find the minimum of a function with respect to certain choice variables. In this model, our choice variables are $\hat{\beta}_0$ and $\hat{\beta}_1$, which are our prediction or *estimators* for the unknown β_0 and β_1 .

Let us define the predicted y , $\hat{y}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 y_i$$

The difference between y_i and $\hat{y}_i$ are the error terms:

$$\hat{e}_i = y_i - \hat{y}_i$$

Now we will choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the sum of the squared prediction errors, $\sum_{i=1}^n \hat{e}_i^2$ is going to be minimum.³ Let us go over how this can be done:

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i \\ \hat{e}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ S &= \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ \min_{\{\hat{\beta}_0, \hat{\beta}_1\}} &\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\end{aligned}$$

³Why do we minimize the **sum of squared errors** and not just the sum of errors? You will hopefully learn this in your econometrics course.

Minimizing a function requires taking partial derivatives and setting them equal to zero:

$$\frac{\partial S}{\partial \hat{\beta}_0} = \sum_{i=1}^n 2 \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) (-1) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = \sum_{i=1}^n 2 \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) (-x_i) = 0$$

This gives us the normal equations:

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0$$

As you see, we now have two equations and two unknowns. Expanding these:

$$\sum y_i - n\hat{\beta}_0 - \left(\sum x_i \right) \hat{\beta}_1 = 0$$

$$\sum x_i y_i - \left(\sum x_i \right) \hat{\beta}_0 - \left(\sum x_i^2 \right) \hat{\beta}_1 = 0$$

Rewriting as a system of equations:

$$n\hat{\beta}_0 + \left(\sum x_i \right) \hat{\beta}_1 = \sum y_i$$

$$\left(\sum x_i \right) \hat{\beta}_0 + \left(\sum x_i^2 \right) \hat{\beta}_1 = \sum x_i y_i$$

Solving for $\hat{\beta}_0$:

$$\left[\frac{n \sum x_i^2}{\sum x_i} \quad \hat{\beta}_0 \right] + \left(\sum x_i^2 \right) \hat{\beta}_1 = \frac{\sum x_i^2 \sum y_i}{\sum x_i}$$

$$\left(\sum x_i \right) \hat{\beta}_0 + \left(\sum x_i^2 \right) \hat{\beta}_1 = \sum x_i y_i$$

Subtracting the equations:

$$\left(\frac{n \sum x_i^2}{\sum x_i} - \sum x_i \right) \hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i}{\sum x_i} - \sum x_i y_i$$

$$\left(\frac{n \sum x_i^2 - (\sum x_i)^2}{\sum x_i} \right) \hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\sum x_i}$$

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Alternatively, from the first normal equation:

$$\begin{aligned}\sum y_i - n\hat{\beta}_0 - \left(\sum x_i\right) \hat{\beta}_1 &= 0 \\ n\hat{\beta}_0 &= \sum y_i - \left(\sum x_i\right) \hat{\beta}_1 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

$$\begin{aligned}\sum x_i y_i - \left(\sum x_i\right) \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) - \left(\sum x_i^2\right) \hat{\beta}_1 &= 0 \\ \sum x_i y_i - \bar{y} \sum x_i - \hat{\beta}_1 \bar{x} \sum x_i - \hat{\beta}_1 \sum x_i^2 &= 0 \\ \frac{\sum x_i y_i}{n} - \bar{y} \frac{\sum x_i}{n} - \hat{\beta}_1 \bar{x} \frac{\sum x_i}{n} - \hat{\beta}_1 \frac{\sum x_i^2}{n} &= 0 \\ \frac{\sum x_i y_i}{n} - \bar{x} \bar{y} - \hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \frac{\sum x_i^2}{n} &= 0 \\ \hat{\beta}_1 &= \frac{\frac{\sum x_i y_i}{n} - \bar{x} \bar{y}}{\bar{x}^2 + \frac{\sum x_i^2}{n}} = \frac{n \sum x_i y_i - n^2 \bar{x} \bar{y}}{n^2 \bar{x}^2 + n \sum x_i^2} \\ \hat{\beta}_1 &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 + (\sum x_i)^2}\end{aligned}$$

Now, reconsider that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and that $\sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$. Substituting the first one into the second:

$$\begin{aligned}\sum x_i y_i - \sum x_i \bar{y} + \hat{\beta}_1 \sum x_i \bar{x} - \hat{\beta}_1 \sum x_i^2 &= 0 \\ \hat{\beta}_1 \left(\sum x_i^2 - \sum x_i \bar{x}\right) &= \sum x_i (y_i - \bar{y}) \\ \hat{\beta}_1 \sum x_i (x_i - \bar{x}) &= \sum x_i (y_i - \bar{y}) \\ \hat{\beta}_1 &= \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}\end{aligned}$$

Now notice the following:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})y_i = \sum (y_i - \bar{y})x_i$$

as,

$$\begin{aligned}
\sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum (x_i - \bar{x})y_i - \bar{y} \underbrace{\sum (x_i - \bar{x})}_0 \\
&= \sum (y_i - \bar{y})x_i - \bar{x} \underbrace{\sum (y_i - \bar{y})}_0
\end{aligned}$$

and, as the sum of the deviations from the mean is zero, i.e.,

$$\begin{aligned}
\sum (x_i - \bar{x}) &= \sum x_i - \sum \bar{x} = \sum x_i - n\bar{x} = 0 \\
\sum (y_i - \bar{y}) &= \sum y_i - \sum \bar{y} = \sum y_i - n\bar{y} = 0
\end{aligned}$$

The same logic applies in:

$$\sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})x_i - \bar{x} \underbrace{\sum (x_i - \bar{x})}_0$$

At the end, the above-derived expression for $\hat{\beta}_1$, i.e.,

$$\hat{\beta}_1 = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})}$$

can be rewritten as:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

so, can be written as:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

To sum up, our **Least Squares estimators** (notice the use of word, "estimator" here) $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model parameters β_0 and β_1 are found to be:

$$\boxed{\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Why do we call $\hat{\beta}_0$ and $\hat{\beta}_1$ estimators? Because they are the formulas or rules to estimate the unknown β_0 and β_1 , and this is the definition of what an estimator is.

Now do not get scared by the amount of algebra that we endured to get to this point. We will almost never do these by hand and software programs like R or Stata will do it for us. But we still need to understand how we got here and how we derived these formulas.

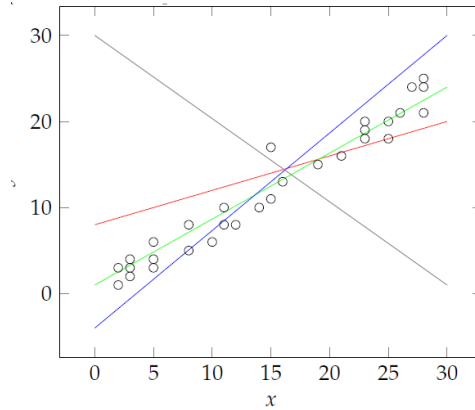


Figure 22: Simple Linear Regression Line. (Source: Karademir and Yücel)

Example 6.1

$X = x$	-3	-2	-1	0	1	2	3	4	5	6
$Y = y$	3	3	4	5	6	7	8	9	10	12

Assume that the following data is appropriate to a simple linear regression equation given by:

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \dots, n$$

Calculate the values of the OLS estimators of β_0 and β_1 .

Solution

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS estimators. From the data, we have $n = 10$.

Calculate the necessary sums

$$\begin{aligned}\sum x_i &= (-3) + (-2) + (-1) + 0 + 1 + 2 + 3 + 4 + 5 + 6 = 15 \\ \sum y_i &= 3 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 12 = 67 \\ \sum x_i^2 &= (-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 \\ &= 9 + 4 + 1 + 0 + 1 + 4 + 9 + 16 + 25 + 36 = 105 \\ \sum x_i y_i &= (-3)(3) + (-2)(3) + (-1)(4) + 0(5) + 1(6) + 2(7) + 3(8) + 4(9) + 5(10) + 6(12) \\ &= -9 - 6 - 4 + 0 + 6 + 14 + 24 + 36 + 50 + 72 = 183\end{aligned}$$

Calculate sample means

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{15}{10} = 1.5 \\ \bar{y} &= \frac{\sum y_i}{n} = \frac{67}{10} = 6.7\end{aligned}$$

Calculate the OLS estimator for β_1

The formula for $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

First, calculate the numerator:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y} = 183 - 10(1.5)(6.7) = 183 - 100.5 = 82.5$$

Then, calculate the denominator:

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n \bar{x}^2 = 105 - 10(1.5)^2 = 105 - 10(2.25) = 105 - 22.5 = 82.5$$

Therefore:

$$\hat{\beta}_1 = \frac{82.5}{82.5} = 1$$

Calculate the OLS estimator for β_0

The formula for $\hat{\beta}_0$ is:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Substituting the values:

$$\hat{\beta}_0 = 6.7 - 1(1.5) = 6.7 - 1.5 = 5.2$$

Thus, the estimated regression line is:

$$\hat{Y}_i = 5.2 + 1 \cdot x_i$$

Let us now calculate the **predicted values** and the **residuals**. From the previous solution, we have the estimated regression equation:

$$\hat{Y}_i = 5.2 + 1 \cdot x_i$$

The residuals are calculated as:

$$\hat{e}_i = y_i - \hat{y}_i$$

i	x_i	y_i	$\hat{y}_i = 5.2 + x_i$	$\hat{e}_i = y_i - \hat{y}_i$	$\hat{e}_i^2$
1	-3	3	$5.2 + (-3) = 2.2$	$3 - 2.2 = 0.8$	0.64
2	-2	3	$5.2 + (-2) = 3.2$	$3 - 3.2 = -0.2$	0.04
3	-1	4	$5.2 + (-1) = 4.2$	$4 - 4.2 = -0.2$	0.04
4	0	5	$5.2 + 0 = 5.2$	$5 - 5.2 = -0.2$	0.04
5	1	6	$5.2 + 1 = 6.2$	$6 - 6.2 = -0.2$	0.04
6	2	7	$5.2 + 2 = 7.2$	$7 - 7.2 = -0.2$	0.04
7	3	8	$5.2 + 3 = 8.2$	$8 - 8.2 = -0.2$	0.04
8	4	9	$5.2 + 4 = 9.2$	$9 - 9.2 = -0.2$	0.04
9	5	10	$5.2 + 5 = 10.2$	$10 - 10.2 = -0.2$	0.04
10	6	12	$5.2 + 6 = 11.2$	$12 - 11.2 = 0.8$	0.64
Total				$\sum \hat{e}_i = 0$	$\sum \hat{e}_i^2 = 1.6$

Observation	x_i	y_i	$\hat{y}_i$	$\hat{e}_i$
1	-3	3	2.2	0.8
2	-2	3	3.2	-0.2
3	-1	4	4.2	-0.2
4	0	5	5.2	-0.2
5	1	6	6.2	-0.2
6	2	7	7.2	-0.2
7	3	8	8.2	-0.2
8	4	9	9.2	-0.2
9	5	10	10.2	-0.2
10	6	12	11.2	0.8

Example 6.2

For a sample of 20 monthly observations, a financial analyst wants to regress the percentage rate of return Y of the common stock of a corporation on the percentage rate of return X of the Standard & Poor's 500 Index. The following information is available:

$$\sum_{i=1}^{20} y_i = 22.6 \quad \sum_{i=1}^{20} x_i = 25.4 \quad \sum_{i=1}^{20} x_i^2 = 145.7 \quad \sum_{i=1}^{20} x_i y_i = 150.5$$

1. Estimate the linear regression of Y on X .
2. Interpret the slope of the sample regression line.
3. Interpret the intercept of the sample regression line.

Solution

$$\bar{x} = \frac{\sum x_i}{n} = \frac{25.4}{20} = 1.27$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{22.6}{20} = 1.13$$

Estimate $\hat{\beta}_1$

Using the formula:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

Calculate numerator:

$$\sum x_i y_i - n \bar{x} \bar{y} = 150.5 - 20(1.27)(1.13) = 150.5 - 20(1.4351) = 150.5 - 28.702 = 121.798$$

Calculate denominator:

$$\sum x_i^2 - n \bar{x}^2 = 145.7 - 20(1.27)^2 = 145.7 - 20(1.6129) = 145.7 - 32.258 = 113.442$$

Therefore:

$$\hat{\beta}_1 = \frac{121.798}{113.442} \approx 1.0737$$

Estimate $\hat{\beta}_0$

Using the formula:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = 1.13 - 1.0737(1.27) = 1.13 - 1.3636 = -0.2336$$

The estimated regression line is:

$\hat{Y} = -0.2336 + 1.0737X$

1. **Slope interpretation** ($\hat{\beta}_1 = 1.0737$):

- For every 1 percentage point increase in the S&P 500 Index return (X), the common stock's return (Y) is expected to increase by approximately 1.0737 percentage points.

2. **Intercept interpretation** ($\hat{\beta}_0 = -0.2336$):

- When the S&P 500 Index return (X) is 0%, the expected return of the common stock is -0.2336%.

Variance of Least Squares Estimators and Hypothesis Testing

Consider the linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n$$

where $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.

As we have shown above, our estimators are the following ones in a model with one explanatory variable, X

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

What about the variances of these estimators? We will not cover how to derive the variances in this course yet, as this is material for your econometrics course. But for now it is sufficient to know that the formulas for $Var(\hat{\beta}_0)$ and $Var(\hat{\beta}_1)$ are as follows:

$$\begin{aligned}Var(\hat{\beta}_1) &= \frac{\sigma^2}{S_{XX}}, \quad \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{XX}} \\ Var(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right), \quad \widehat{Var}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)\end{aligned}$$

where;

$$\begin{aligned}S_{XX} &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ \hat{\sigma}^2 &= \text{MSE} = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2\end{aligned}$$

Calculating Standard Errors

$$\begin{aligned}\text{SE}(\hat{\beta}_1) &= \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} \\ \text{SE}(\hat{\beta}_0) &= \sqrt{\widehat{\text{Var}}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}\end{aligned}$$

An important note: unlike before, we call the square root of variance of the least squares estimator, *standard deviation* and NOT standard deviation. The reason is as follows:

- **Purpose:** Measures the spread or dispersion of *individual observations/data points*
- **What it describes:** Natural variability in the population or sample
- **Formula for sample standard deviation:**

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- **Interpretation:** How far individual data points typically deviate from the mean
- **Example:** If $s = 10$ for test scores, individual scores typically vary by about 10 points from the class average

Standard Error: Describing Estimator Precision

- Standard errors measures the precision or reliability of an *estimator/statistic*
- **What it describes:** Sampling variability of the estimator across different samples
- **Formula for standard error of $\hat{\beta}_1$:**

$$\text{SE}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

- **Interpretation:** How much $\hat{\beta}_1$ would vary from sample to sample
- It's the *standard deviation of the sampling distribution* of the estimator
- "Error" refers to "error in estimation" or "estimation error"
- Represents the typical magnitude of error we make when using $\hat{\beta}_1$ to estimate β_1
- If we repeated the study infinitely many times, the standard error tells us how much our estimates would typically vary from the true parameter

Hypothesis Testing

Test for slope: $H_0 : \beta_1 = c$ vs. $H_1 : \beta_1 \neq c$

$$t = \frac{\hat{\beta}_1 - c}{\text{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - c}{\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}}, \quad t \sim t_{n-2} \text{ under } H_0$$

Test for intercept: $H_0 : \beta_0 = c$ vs. $H_1 : \beta_0 \neq c$

$$t = \frac{\hat{\beta}_0 - c}{\text{SE}(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - c}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}}, \quad t \sim t_{n-2} \text{ under } H_0$$

Confidence Intervals

100(1 - α)% **Confidence Interval for β_1**

$$\hat{\beta}_1 \pm t_{1-\alpha/2, n-2} \cdot \text{SE}(\hat{\beta}_1) = \hat{\beta}_1 \pm t_{1-\alpha/2, n-2} \cdot \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$$

100(1 - α)% **Confidence Interval for β_0**

$$\hat{\beta}_0 \pm t_{1-\alpha/2, n-2} \cdot \text{SE}(\hat{\beta}_0) = \hat{\beta}_0 \pm t_{1-\alpha/2, n-2} \cdot \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}$$

Example: Study Hours vs. Test Scores

A researcher wants to test if there's a relationship between study hours and test scores. They collect data from 5 students:

Study Hours (X)	Test Score (Y)
1	65
2	75
3	85
4	90
5	95

Solution

$$n = 5, \quad \bar{X} = 3, \quad \bar{Y} = 82$$

$$S_{XX} = \sum (X_i - \bar{X})^2 = (1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 40$$

Calculate OLS Estimates

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{40}{10} = 4.0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 82 - 4.0 \times 3 = 70$$

So our regression line is: $\hat{Y} = 70 + 4.0X$

Standard Error

$$\text{Fitted values: } \hat{Y}_1 = 74, \hat{Y}_2 = 78, \hat{Y}_3 = 82, \hat{Y}_4 = 86, \hat{Y}_5 = 90$$

$$\text{Residuals: } e_1 = -9, e_2 = -3, e_3 = 3, e_4 = 4, e_5 = 5$$

$$\text{Sum of squared residuals: } \sum e_i^2 = 81 + 9 + 9 + 16 + 25 = 140$$

$$\text{MSE: } \hat{\sigma}^2 = \frac{140}{5 - 2} = \frac{140}{3} = 46.67$$

Standard error of $\hat{\beta}_1$:

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} = \sqrt{\frac{46.67}{10}} = \sqrt{4.667} = 2.16$$

Hypothesis Test for Slope: Does study time affect test scores?

Formal hypothesis test:

$$H_0 : \beta_1 = 0 \quad (\text{No relationship})$$

$$H_1 : \beta_1 \neq 0 \quad (\text{There is a relationship})$$

Test statistic:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{4.0 - 0}{2.16} = 1.85$$

Degrees of freedom: $df = n - 2 = 5 - 2 = 3$

Critical value (at $\alpha = 0.05$, two-tailed):

$$t_{0.975,3} = 3.182 \quad (\text{from t-table})$$

Decision and Interpretation

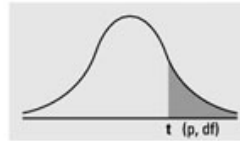
Since $|t| = 1.85 < 3.182$, we **fail to reject** H_0 at the 5% significance level.

Interpretation: We don't have enough evidence to conclude that study hours significantly affect test scores ($t = 1.85$, $df = 3$, $p > 0.05$).

Appendix

Figure 23: Student's T-Distribution Table

Numbers in each row of the table are values on a t -distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p) .



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

Figure 24: Z Table

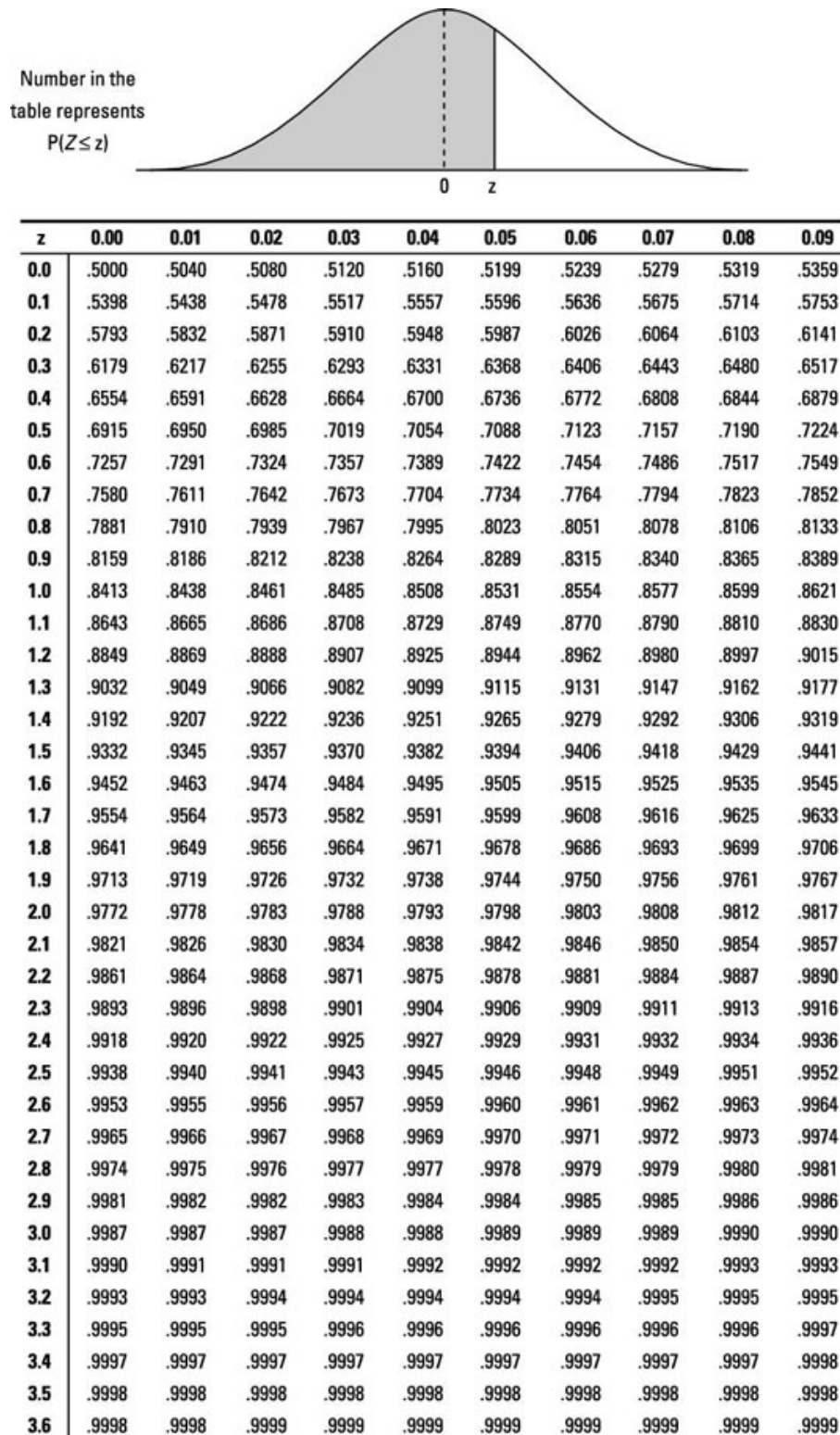


Figure 25: Chi-Square Distribution Table (Right-tail probabilities)

Degrees of freedom (df)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-----	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

References

This note has been prepared based on the following textbooks and you are highly encouraged to read the relevant chapters.

1. Hogg, R.V., McKean, J.W. and Craig, A.T., 2010. Introduction to mathematical statistics. Pearson Education USA. 8th Edition.
2. Newbold, P., Carlson, W.L. and Thorne, B., 2019. Statistics for business and economics. Pearson UK. 8th Edition.
3. Karademir, S. and Yücel, M.E., 2024. Probability Theory and Statistics Lecture Notes <https://sites.google.com/view/erayyucel/probability-and-statistics>
4. Işlak, Ü. 2022. Temel Olasılık Teorisi ve İstatistik-I. Nesin Yayıncılık.
5. Wackerly, D.D., Mendenhall, W. and Scheaffer, R.L., 2008. Mathematical statistics with applications (Vol. 7). Belmont, CA: Thomson Brooks/Cole.