Chapter 1

What Is Econometrics

1.1 Economics

1.1.1 The Economic Problem

Economics concerns itself with satisfying unlimited wants with limited resources. As such it proposes solutions that involve the production and consumption of goods using a particular allocation of resources. At the micro-economic level, these solutions are often the results of optimizing behavior by economics agents such as consumers or producers and possibly impacted by regulators. At the macro-economic level, they can also be the results of decisions made by policymakers such as central bankers, the executive branch and the legislative branch.

1.1.2 Some Economic Questions

As economic agents arrive at their decisions they encounter and must answer a host of economic questions: Should the Eurozone countries coordinate their fiscal as well as monetary policies? Relatedly, should Greece default on its' debt and form a new Greek currency? What is the best time for Professor Brown to retire? Should an undergraduate at Rice major in economics or mathematical economics? Why did the savings rate decline during the Great Expansion of the nineties?

As economists we are in the business of finding answers to such questions. Hopefully, the answers are justifiable and replicable so we are in some sense consistent. Moreover, we hope that our answers will be born out by subsequent experience. In this chapter, we discuss a structure for going about constructing such answers. This structure is how we do business in economics.

1.1.3 Economics as a Science

Science can be defined as the intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural
world through observation and experiment. The purpose of the experiment is to control for various factors that might vary when the environment is not managed as in an experiment. This allows the observer to focus on the relationship between a limited number of variables.

So to be a science economics must be founded on the notion of studying and explaining the world through observation and experimentation. Economics is classified as a social science since it is concerned with social phenomena such as organization and operation of the household, firm, or society.

1.4 Providing Scientific Answers

As such, economics would attempt to provide answers to the questions posed above, for example, using scientific approaches that would yield similar answers in similar circumstances. In that sense, it attempts to bring the replicability of scientific experiments into the analysis and formulation of answers. This is accomplished by combining economic theory with the analysis of pre-existing relevant data using appropriate statistical techniques.

1.2 Data

1.2.1 Accounting

Most of the data used in econometrics is accounting data. Accounting data is routinely recorded data (that is, records of transactions) as part of market or institutional activities. Accounting data has many shortcomings, since it is not collected specifically for the purposes of the econometrician. Oftentimes the available data may not measure the phenomenon we are interested in explaining but a related one.

Of course this characterization is not universal. For example, we may be interested in the relationship between birth weight or mortality to the income level. In both cases, the variable we are looking at is not transactional and hence accounting data.

1.2.2 Nonexperimental

Moreover, the data is typically nonexperimental. The econometrician does not have any control over nonexperimental data. In this regard economics is in much the same situation as meteorology and astronomy. This nonexperimental nature of the data causes various problems. Econometrics is largely concerned with how to manage these problems in such a fashion that we can have the same confidence in our findings as when experimental data is used.

Here again, the characterization is not universal. For example there have been income maintenance experiments where the impact on labor effort in response to guaranteeing a base level of income was examined directly with application of a negative income tax to participants. And there are numerous
laboratory experiments that have helped advance our understanding of game theoretic equilibria.

1.2.3 Data Types

It is useful to give a typeology of the different forms that economic data can take:

**Definition 1.1.** *Time-series data* are data that represent repeated observations on some variable in adjacent time periods. A time-series variable is often subcripted with the letter $t$. Examples are the familiar aggregate economic variables such as interest rates, gross domestic product, and the consumer price index, which are available at the annual and quarterly, and monthly intervals. Stock prices are another good example, available at the tic interval.

**Definition 1.2.** *Cross-sectional data* are data that represent a set of observations of some variable at one specific instant for a number of agents or organizations. A cross-sectional variable is often indexed with the letter $i$. Most often, this data is microeconomic. For example, a firm’s employment data at a particular point in time has observations for each employee.

**Definition 1.3.** *Time-series cross-sectional data* are data that are both time-series and cross-sectional. Such data is typically indexed by $t$ for the time period and $i$ for the agent.

**Definition 1.4.** A special case of time-series cross-sectional data is *panel data*. Panel data are observations on the same set of agents or a panel over time. A leading example is the Survey of Income and Program Participation which provides information about the incomes of American individuals and households and their participation in income transfer programs.

1.2.4 Empirical Regularities

We need to look at the data to detect regularities or stylized facts. These regularities that persist across time or individuals beg for explanation of some sort. Why do they occur and why do they persist? An example would be the observation that the savings rate in the U.S. was very low during the record expansion of the nineties and rose during the subsequent great recession. In economics we seek to explain the regularities with models.

1.3 Models

1.3.1 Simplifying Assumptions

Models are simplifications of the real world. As such they introduce simplifying assumptions that hold a number of factors constant and allow the analysis
of the relationship between a limited number of variables. These simplifying assumptions are the factors that would be controlled if we were conducting a controlled experiment to verify the model.

Models are sometimes criticized as unrealistic but if they were realistic they would not be models but in fact represent reality. The real question is whether the simplifications introduced by the assumptions are useful for dealing with the phenomena we attempt to explain. Hopefully, we will provide some metric of what is a useful model.

1.3.2 Economic Theory

As a result of optimizing behavior by consumers, producers, or other agents in the economy, economic theory will suggest how certain variables will be related. For example demand for a consumer good can be represented as a function of its own price, prices of complimentary goods, prices of competing goods, and the income level of the consumer.

1.3.3 Equations

Economic models are usually stated in terms of one or more equations. For example, in a simple macroeconomic model we might represent \( C_t \) (aggregate consumption at time \( t \)) as a linear function of \( Y_t \) (aggregate income):

\[
C_t = \alpha + \beta Y_t
\]

where \( \alpha \) and \( \beta \) are parameters to be interest with their values targets of inquiry. The choice of the linear model is usually a simplification but the variables that go into the model will typically follow directly from economic theory.

1.3.4 Error Component

As an example, consider the following scatter plot of aggregate income versus aggregate consumption for annual data over the period 1946-1958. The data reveal a clear increasing relationship between income and consumption. Moreover, average consumption, which is represented by a ray from the origin declines as income increases. These both suggest that the linear relationship has a positive intercept and a slope less than one. However, the relationship is not exact for any choice of the parameters \( \alpha \) and \( \beta \).

Because none of our models are exact, we include an error component in our equations, usually denoted \( u_t \). Thus we have

\[
C_t = \alpha + \beta Y_t + u_t
\]

This error term could be viewed as an approximation error, a misspecification error, or result from errors in optimization by economic agents. In any event, the linear parametric relationship is not exact for any choice of the parameters.
1.4 Statistics

1.4.1 Stochastic Component

In order to conduct analysis of the model once the error term has been introduced, we assume that it is a random variable and subject to distributional restrictions. For example, we might assume that \( u_t \) is an i.i.d. variable with

\[
E[u_t|Y_t] = 0
\]

\[
E[u_t^2|Y_t] = \sigma^2.
\]

Again, this is a modeling assumption and may be a useful simplification or not. We have combined economic theory to explain the systematic part of the behavior and statistical theory to characterize the unsystematic part.

1.4.2 Statistical Analysis

The introduction of the stochastic assumptions on the error component allows the use of statistical techniques to perform various tasks in analyzing the combined model. This analysis will involve statistical inference whereby we use the distributional behavior implied by our economic model and statistical assumptions to make inferences about properties of the underlying model based on the observed behavior generated by the model.

Implicitly, in the conditional expectations assumptions used in the aggregate consumption example above, the explanatory variable is taken as fixed or non-stochastic. Sometimes the explanatory variables are just as obviously random.
as the dependent variables, in which case we will need to make stochastic assumptions regarding their joint behavior. This will sometimes complicate matters considerably.

1.4.3 Tasks

In the sequel we will concern ourselves with three main tasks in applying statistical analysis to the model in question:

1. Estimating the parameters of the model. These parameters would include the parameters of the systematic part of the model and the parameters of the distribution of the error component. These estimates are typically point estimates though they can be interval estimates.

2. Hypothesis tests concerning the behavior of the model. This is where the approach really becomes scientific. We might be interested in whether accumulation of wealth is a an adequate explanation for the low saving rate situation mentioned above. This could be tested using an hypothesis test.

3. Forecast the behavior of the dependent variable outside the sample using the estimated model. Such forecasts can be point forecasts such as expected behavior or interval forecasts which give a range of values with an attached probability of occurrence. Usually, these forecasts will be conditional on knowledge of the values of the explanatory variables.

1.5 Econometrics

Econometrics involves the integration of these three components: data, economic theory, and statistical theory to model an economic phenomenon. Since our data is usually nonexperimental, econometrics makes use of economic and statistical theory to adjust for the lack of proper data. Economic theory allows us to focus on a reduced set of variables and possible knowledge of how they interact. If we were conducting an experiment, we would be able to control the various other factors and focus on the variables whose relationship interests us. And statistical assumptions allow us to effectively control for the effect of other factors not suggested by economic theory. This integration of data, economic theory, and statistical theory to answer economic questions is the essence of econometrics.

1.5.1 Interactions

Often the phenomenon we are seeking to explain is a set of stylized facts that characterize how the various economic variables move together. This may lead to economic theories that possibly explain these stylized facts. The question of whether or not the theory adequately explains the data is measured through
the application of appropriate statistical techniques to estimate the model based on the available data. Hopefully, there are hypotheses (behaviors) that are predicted by the economic theory that can be tested using statistical techniques. Based on the outcome of this exercise the theory may need modification or additional or different data gathered to address the problem at hand.

Economists speak of “bringing the model to the data” in this exercise. It is the responsibility of the econometrician to choose or develop an appropriate statistical technique that matches the needs of the data and the theory. As was mentioned above, economic data are largely nonexperimental. And our economic models typically have multiple variables determined simultaneously by the model. Consequently, we focus on and develop techniques for dealing with these properties and others that may need to be handled.

Some of this interaction is illustrated in Figure 1.2. Observations on aggregate consumption and aggregate disposable income show a generally positive relationship in the plot in the Data circle. This leads to the Duesenberry-type
linear consumption equation: \( C_t = \alpha + \beta D_t \) in the Theory circle. Since the model does not exactly fit the data an error term \( u_t \) is appended. Assumptions are made about the statistical properties of the observable data and the error term and an appropriate statistical technique is applied to obtain the estimated model which is represented as the fitted green line in the Statistics circle. There are feedbacks between the three circles with the needs or results of the statistical analysis leading to modifications in the data and/or theory. Likewise the theoretical analysis may lead to changes in which and what type of data is gathered.

1.5.2 Scientific Method

In the scientific method, we take an hypothesis, construct an experiment with which to test the hypothesis, and based on the results of the hypothesis we either reject the hypothesis or not. The purpose of the experiment is to control for various factors that might impact the outcome but are not of interest. This allows us to focus our attention on the relationship between the variables of interest that result from the experiment. Based on the outcomes for these variables of interest, we either reject or not.

In econometrics, we let the systematic model and the statistical model or some specified component of them to play the role of the hypothesis. The available data, ex post, plays the role of the outcome of an experiment. Statistical analysis of the data using the combined model enables us to reject the hypothesis or not. So econometrics is what allows us to use the scientific method and hence is critical in economics being treated as a science.

1.5.3 Choosing A Model

In testing an hypothesis above we are implicitly choosing between a model which satisfies the hypothesis and one which doesn’t. So the hypothesis test implicitly is a criterion for choosing between models. If we reject the hypothesis then we are choosing the model which doesn’t satisfy the hypothesis. This a clear-cut decision framework. In any event, we are choosing the model which best agrees with the data.

Sometimes, we are faced with a situation where the data do not give a strong preference between two competing models. In this case, the simplest model that fits the data is often the best choice. This is known as using Ocam’s razor.
Chapter 2

Some Useful Distributions

Usually, interest in an economic phenomenon is piqued from viewing sample moments of some important variable. So naturally, we will give a great deal of attention to sample moments in the sequel. Ultimately, though, to understand the behavior of such sample moments in either large or small samples we will need to resort to distributional theory. We will only consider continuous distributions, in this chapter, but the discrete cases are easily handled in a similar fashion. For reasons that become apparent, the normal distribution and distributions associated with it play a central role in much of our statistical analysis in both small and large samples.

2.1 Introduction

2.1.1 Inference

As statisticians, we are often called upon to answer questions or make statements concerning certain random variables. For example: is a coin fair (i.e. is the probability of heads = 0.5) or what is the expected value of GNP for the quarter.

Typically, answering such questions requires knowledge of the underlying (population) distribution of the random variable. Unfortunately, we usually do not know this distribution (although we may have a strong idea, as in the case of the coin).

In order to gain knowledge of the distribution, we draw several realizations of the random variable. The notion is that the observations in this sample contain information concerning the population distribution.

Definition 2.1. The process by which we make statements concerning the population distribution based on the sample observations is called inference. □

Example 2.1. We decide whether a coin is fair by tossing it several times and observing whether it seems to be heads about half the time. □
2.1.2 Random Samples

**Definition 2.2.** Suppose we draw \( n \) observations of a random variable, denoted \( \{x_1, x_2, ..., x_n\} \) and each \( x_i \) is independent and has the same (marginal) distribution, then \( \{x_1, x_2, ..., x_n\} \) constitute a *simple random sample.* □

**Example 2.2.** We toss a coin three times. Supposedly, the outcomes are independent. If \( x_i \) counts the number of heads for toss \( i \), then we have a simple random sample. □

Note that not all samples are simple random. They can be either non-independent or non-identical or both.

**Example 2.3.** We are interested in the income level for the population in general. The \( n \) observations available in this class are not identical since the higher income individuals will tend to be more variable. □

**Example 2.4.** Consider the aggregate consumption level. The \( n \) observations available in this set are not independent since a high consumption level in one period is usually followed by a high level in the next. □

2.1.3 Sample Statistics

**Definition 2.3.** Any function of the observations in the sample which is the basis for inference is called a *sample statistic.* □

**Example 2.5.** In the coin tossing experiment, let \( S \) count the total number of heads and \( P = \frac{S}{3} \) count the sample proportion of heads. Both \( S \) and \( P \) are sample statistics. □

2.1.4 Sample Distributions

A sample statistic is a random variable — its value will vary from one experiment to another. As a random variable, it is subject to a distribution.

**Definition 2.4.** The distribution of the sample statistic is the *sample distribution* of the statistic. □

**Example 2.6.** The statistic \( S \) introduced above has a multinomial sample distribution. Specifically \( \Pr(S = 0) = 1/8, \Pr(S = 1) = 3/8, \Pr(S = 2) = 3/8, \) and \( \Pr(S = 3) = 1/8. \) □

2.2 The Sample Mean

2.2.1 Sample Mean

**Definition 2.5.** Given a sample \( \{x_1, x_2, ..., x_n\} \) of observations on a random variable \( x \), then

\[
x_n = \frac{1}{n} \sum_{i=1}^{n} x_i
\]
2.2. THE SAMPLE MEAN

is the sample mean or average of the sample. □

This form plays a predominate role in statistics and econometrics. All sample moments can be expressed as sample moments of functions of the data. And, at least in large samples, many of the estimators’ and statistics’ behavior depends on their relationship to sample averages. Consequently, the behavior of the form is well studied and well known.

2.2.2 Sample Sum

Consider the simple random sample \( \{x_1, x_2, ..., x_n\} \), where \( x \) measures the height of an adult female. We will assume that \( E(x_i) = \mu \) and \( \text{Var}(x_i) = \sigma^2 \), for all \( i = 1, 2, ..., n \).

Let \( S = x_1 + x_2 + \cdots + x_n \) denote the sample sum. Now,

\[
E(S) = E(x_1 + x_2 + \cdots + x_n) = E(x_1) + E(x_2) + \cdots + E(x_n) = n\mu
\]

Also,

\[
\text{Var}(S) = E(S - E(S))^2 = E((x_1 + x_2 + \cdots + x_n - n\mu)^2)
\]

\[
= E \left[ \sum_{i=1}^{n} (x_i - \mu)^2 \right]
\]

\[
= E[(x_1 - \mu)^2 + (x_1 - \mu)(x_2 - \mu) + \cdots + (x_n - \mu)^2]
\]

\[
= n\sigma^2,
\]

since \( E((x_i - \mu)(x_j - \mu)) = 0 \) for \( i \neq j \), by independence.

2.2.3 Moments Of The Sample Mean

The mean of the sample mean is

\[
E \bar{x}_n = E\frac{S}{n} = \frac{1}{n}E(S) = \frac{1}{n}n\mu = \mu.
\]
The variance of the sample mean is
\[ \text{Var}(\bar{x}_n) = E((\bar{x} - \mu)^2) \]
\[ = E\left(\frac{S}{n} - \mu\right)^2 \]
\[ = E\left[\frac{1}{n} (S - n\mu)]^2 \right. \]
\[ = \frac{1}{n^2} E(S - n\mu)^2 \]
\[ = \frac{1}{n^2} n\sigma^2 \]
\[ = \frac{\sigma^2}{n}. \quad (2.3) \]

2.2.4 Sampling Distribution

We have been able to establish the mean and variance of the sample mean. However, in order to know its complete distribution precisely, we must know the probability density function (pdf) of the random variable \( x \).

2.3 The Normal Distribution

2.3.1 Density Function

Definition 2.6. A continuous random variable \( x \) with the density function
\[ f(x) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2\sigma^2} (x-\mu)^2} \]
follows the normal distribution, where \( \mu \) and \( \sigma^2 \) are the mean and variance of \( x \), respectively. \( \Box \)

Since the distribution is characterized by the two parameters \( \mu \) and \( \sigma^2 \), we denote a normal random variable by \( x \sim N(\mu, \sigma^2) \).

The normal density function is the familiar “bell-shaped” curve, as is shown in Figure 2.1 for \( \mu = 0 \) and \( \sigma^2 = 1 \). It is symmetric about the mean \( \mu \). Approximately 2/3 of the probability mass lies within \( \pm \sigma \) of \( \mu \) and about .95 lies within \( \pm 2\sigma \). There are numerous examples of random variables that have this shape. Many economic variables are assumed to be normally distributed.

2.3.2 Linear Transformation

Consider the transformed random variable \( Y = a + bx \)

We know that
\[ \mu_Y = EY = a + b\mu_x \]
2.3. THE NORMAL DISTRIBUTION

If $x$ is normally distributed, then $Y$ is normally distributed as well. That is,

$$Y \sim N \left( \mu_Y, \sigma_Y^2 \right)$$

Moreover, if $x_i \sim N \left( \mu_x, \sigma_x^2 \right)$ and $z_i \sim N \left( \mu_z, \sigma_z^2 \right)$ are independent, then

$$Y = a + bx + cz \sim N \left( a + b\mu_x + c\mu_z, b^2\sigma_x^2 + c^2\sigma_z^2 \right)$$

These results will be formally demonstrated in a more general setting in the next chapter. This result means that any linear combination of any number of independent normals is itself normal.

2.3.3 Distribution Of The Sample Mean

If, for each $i = 1, 2, \ldots, n$, the $x_i$’s are independent, identically distributed (i.i.d.) normal random variables, then

$$\bar{x}_n \sim N \left( \mu_x, \frac{\sigma_x^2}{n} \right)$$

(2.5)
2.3.4 The Standard Normal

The distribution of $\bar{x}_n$ will vary with different values of $\mu_x$ and $\sigma^2_x$, which is inconvenient. Rather than dealing with a unique distribution for each case, we perform the following transformation:

$$Z = \frac{\bar{x}_n - \mu_x}{\sigma_x / \sqrt{n}} = \frac{\bar{x}_n}{\sqrt{\sigma_x^2 / n}} - \frac{\mu_x}{\sqrt{\sigma_x^2 / n}}$$  \hspace{1cm} (2.6)

Now,

$$E(Z) = E\left(\frac{\bar{x}_n}{\sqrt{\sigma_x^2 / n}} - \frac{\mu_x}{\sqrt{\sigma_x^2 / n}}\right) = E\left(\frac{\bar{x}_n}{\sqrt{\sigma_x^2 / n}}\right) - E\left(\frac{\mu_x}{\sqrt{\sigma_x^2 / n}}\right) = 0.$$

Also,

$$E(Z^2) = E\left(\frac{\bar{x}_n}{\sqrt{\sigma_x^2 / n}} - \frac{\mu_x}{\sqrt{\sigma_x^2 / n}}\right)^2 = E\left[\left(\frac{\bar{x}_n}{\sqrt{\sigma_x^2 / n}} - \frac{\mu_x}{\sqrt{\sigma_x^2 / n}}\right)^2\right] = E\left[\frac{n}{\sigma_x^2} (\bar{x}_n - \mu_x)^2\right] = \frac{n \sigma^2_x}{\sigma_x^2} \frac{1}{n} = 1.$$

Thus $Z \sim N(0,1)$. If we know the variance $\sigma^2$, then the $z$-transformation can be used to test the mean.

**Definition 2.7.** The $N(0, 1)$ distribution is the *standard normal* and is well-tabulated. The probability density function for the standard normal distribution evaluated at point $z$ is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} = \varphi(z)$$  \hspace{1cm} (2.7)

**Example 2.7.** Suppose that $x_i \sim i.i.d. N(\mu, \sigma^2)$ with $\sigma^2$ known and we seek to test the null hypothesis that $\mu = \mu_0$ against $\mu = \mu_1$, then

$$Z = \frac{\bar{x}_n - \mu_0}{\sigma^2 / \sqrt{n}} \sim N(0, 1)$$

under the null hypothesis but will have a non-zero mean under the alternative. The behavior under the alternatives will be treated carefully in the next chapter.
2.4 The Central Limit Theorem

2.4.1 Normal Theory

The normal density has a prominent position in statistics. This is not only because many random variables appear to be normal, but also because most any sample mean appears normal as the sample size increases. Subject to a few weak conditions, this result applies for the average of even highly non-normal distributions. Specifically, we have

**Theorem 2.1.** Suppose (i) \( \{x_1, x_2, \ldots, x_n\} \) is a simple random sample, (ii) \( E x_i = \mu_x \), and (iii) \( E x_i^2 = \sigma_x^2 \), then as \( n \to \infty \), the distribution of \( Z = \frac{\bar{x} - \mu}{\sqrt{\sigma_x^2/n}} \) becomes standard normal. □

What this means is that for \( f_n(\cdot) \) the probability density function (PDF) of \( Z \), then

\[
\lim_{n \to \infty} f_n(c) = \varphi(c) \tag{2.8}
\]

pointwise for any point of evaluation \( c \). Likewise, for \( F_n(\cdot) \) denoting the analogous cumulative distribution function (CDF) of \( Z \) and \( \Phi(\cdot) \) denoting the CDF of the standard normal, then

\[
\lim_{n \to \infty} F_n(c) = \Phi(c)
\]

pointwise at any point of evaluation \( c \). This is the Lindberg-Levy form of the central limit theorem and will be treated more extensively in Chapter 4. Notationally, we write

\[
Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \to_d N(0, 1)
\]

or equivalently

\[
\sqrt{n}(\bar{x} - \mu) \to_d N(0, \sigma^2).
\]

The panels in Figure 2.2 demonstrate the power of the result for a \( x \) drawn from a triangular distribution, normalized to have mean zero and unit variance. We draw 100,000 replications of samples of size 25. The first panel displays a histogram of the first observation of each replication. This is the triangular distribution we are working with. It has limited support and is highly asymmetric. The second panel is a histogram of the \( z \)-transformation of the means for the 100,000 replications. At a sample size of 25 it is clear that the standard normal distribution is already a very good approximation to the distribution.
CHAPTER 2. SOME USEFUL DISTRIBUTIONS

2.5 Distributions Associated With The Normal Distribution

2.5.1 The Chi-Squared Distribution

Definition 2.8. Suppose that \{Z_1, Z_2, \ldots, Z_m\} is a simple random sample, and $Z_i \sim i.i.d. N(0, 1)$. Then

$$w = \sum_{i=1}^{m} Z_i^2 \sim \chi^2_m,$$

or $w$ has the Chi-squared distribution with $m$ degrees of freedom. □

The probability density function for the $\chi^2_m$ is

$$f_{\chi^2}(x) = \frac{1}{2^{m/2}\Gamma(m/2)}x^{m/2-1}e^{-x/2}, x > 0 \quad (2.9)$$

where $\Gamma(x)$ is the gamma function. See Figure 2.3 for plots of the chi-squared distribution for various degrees of freedom. Note how the distribution begins to attain the familiar bell-shape but is stretched out to the right as the degrees of freedom grows larger. This a manifestation of the central limit theorem since $w/m$ would be an average. In fact, $(w - m)/\sqrt{2m}$ is approximately standard normal for large $m$.

The chi-squared distribution will prove useful in testing hypotheses on both the variance of a single variable and the (conditional) means of several. Typically, we are only interested in the content of the right-hand side tail in conducting tests, for reasons that will become apparent with the multivariate usage, which will be explored in the next chapter. But the left-hand side tail could be interesting if we are testing whether the true variance is a specific value, say $\sigma^2_0$ and are interested in the possibility that the true value is either larger or smaller than this value.
2.5. DISTRIBUTIONS ASSOCIATED WITH THE NORMAL DISTRIBUTION

Example 2.8. Consider the estimate of \( \sigma^2 \)

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

Then we can show that

\[
(n - 1) \frac{s^2}{\sigma^2} \sim \chi^2_{n-1}.
\]  \hspace{1cm} (2.10)

\[\square\]

2.5.2 The \( t \) Distribution

Definition 2.9. Suppose that \( Z \sim N(0,1) \), \( Y \sim \chi^2_m \), and that \( Z \) and \( Y \) are independent. Then

\[
\frac{Z}{\sqrt{Y/m}} \sim t_m,
\]  \hspace{1cm} (2.11)

where \( m \) are the degrees of freedom of the \( t \) distribution. \[\square\]

The probability density function for a \( t \) random variable with \( m \) degrees of freedom is

\[
f_t(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi} \Gamma\left(\frac{m}{2}\right) \left(1 + \frac{x^2}{m}\right)^{(m+1)/2}},
\]  \hspace{1cm} (2.12)
for $-\infty < x < \infty$. See Figure 2.4 for plots of the distribution for various choices of $m$.

The $t$ (also known as Student’s $t$) distribution, is named after William Gossett, who published under the pseudonym “Student.” It is useful in testing hypotheses concerning the (conditional) mean when the variance is estimated. Note that the distribution is symmetric around zero and becomes less fat-tailed as the degrees of freedom increase. In fact, as $m \to \infty$, the distribution approaches the standard normal. This just reflects the fact that the denominator $Y/m$ converges in probability to one.

**Example 2.9.** Consider the sample mean from a simple random sample of normals. We know that $\bar{x} \sim N(\mu, \sigma^2/n)$ and

$$Z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

Also, we know that

$$Y = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{n-1},$$

where $s^2$ is the unbiased estimator of $\sigma^2$. Thus, if $Z$ and $Y$ are independent
2.5. DISTRIBUTIONS ASSOCIATED WITH THE NORMAL DISTRIBUTION

(which, in fact, is the case), then

\[
\frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\frac{x - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{(n-1)\frac{s^2}{\sigma^2}/(n-1)}}
\]

\[
= \frac{(x - \mu)}{\sqrt{\frac{s^2}{n}\frac{m}{\sigma^2}}}
\]

\[
= \frac{x - \mu}{\sqrt{s^2/n}} \sim t_{n-1} \tag{2.13}
\]

\[
\square
\]

2.5.3 The F Distribution

Definition 2.10. Suppose that \(Y \sim \chi^2_m\), \(W \sim \chi^2_n\), and that \(Y\) and \(W\) are independent. Then

\[
\frac{Y/m}{W/n} \sim F_{m,n},
\]

where \(m, n\) are the degrees of freedom of the F distribution. \(\square\)

The probability density function for a \(F\) random variable with \(m\) and \(n\) degrees of freedom is

\[
f_F(x) = \frac{\Gamma \left( \frac{m+n}{2} \right) \left( m/n \right)^{m/2}}{\Gamma \left( \frac{m}{2} \right) \Gamma \left( \frac{n}{2} \right)} \frac{x^{(m/2)-1}}{(1 + mx/n)^{(m+n)/2}} \tag{2.15}
\]

The \(F\) distribution is named after the great statistician Sir Ronald A. Fisher, and is used in many applications, most notably in the analysis of variance. This situation will arise when we seek to test multiple (conditional) mean parameters with estimated variance. Note that when \(x \sim t_n\) then \(x^2 \sim F_{1,n}\). Also note that as the numerator degrees of freedom grows large, which is frequently the case, the denominator converges in probability to one and the ratio is approximated by the behavior of the numerator: a chi-squared divided by its’ degrees of freedom. Some plots of the \(F\) distribution for various choices of numerator and denominator degrees of freedom can be seen in Figure 2.5.
Figure 2.5: Some F Distributions
Chapter 3

Multivariate Distributions

In economics we are typically interested in relationships that involve a number of variables. For example aggregate consumption, disposable income, and wealth in a macroeconomic context or a list of stock prices that make up a portfolio in a microeconomic situation. Consequently, the statistical analysis will of necessity be multivariate in nature. In this chapter we will develop a number of concepts that prove useful in multivariate statistics. As for the univariate case, the normal distribution plays a central role in multivariate statistics and will be fully developed. In this treatment, the list of random variables of interest will be arrayed into a vector for notational convenience.

3.1 Matrix Algebra Of Expectations

3.1.1 Moments of Random Vectors

Let

\[
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_m
\end{bmatrix}
= \mathbf{x}
\]

be an \( m \times 1 \) vector-valued random variable. Each element of the vector is a scalar random variable of the type discussed in the previous chapter.

The expectation of a random vector is

\[
E[\mathbf{x}] = E \begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_m
\end{bmatrix} = \begin{bmatrix}
    \mu_1 \\
    \mu_2 \\
    \vdots \\
    \mu_m
\end{bmatrix} = \mathbf{\mu}.
\] (3.1)

Note that \( \mathbf{\mu} \) is also an \( m \times 1 \) column vector. We see that the mean of the vector is the vector of the means.
Next, we evaluate the following:

\[ E[(\mathbf{x} - \mathbf{\mu})(\mathbf{x} - \mathbf{\mu})'] \]

\[
= \mathbf{E} \begin{bmatrix}
(x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) & \cdots & (x_1 - \mu_1)(x_m - \mu_m) \\
(x_2 - \mu_2)(x_1 - \mu_1) & (x_2 - \mu_2)^2 & \cdots & (x_2 - \mu_2)(x_m - \mu_m) \\
\vdots & \vdots & \ddots & \vdots \\
(x_m - \mu_m)(x_1 - \mu_1) & (x_m - \mu_m)(x_2 - \mu_2) & \cdots & (x_m - \mu_m)^2
\end{bmatrix}
\]

\[
= \begin{bmatrix}
\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\
\sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm}
\end{bmatrix}
\]

\[ = \Sigma. \tag{3.2} \]

\(\Sigma\), the covariance matrix, is an \(m \times m\) matrix of variance and covariance terms. The variance \(\sigma_i^2 = \sigma_{ii}\) of \(x_i\) is along the diagonal, while the cross-product terms represent the covariance between \(x_i\) and \(x_j\).

### 3.1.2 Properties Of The Covariance Matrix

#### 3.1.2.1 Symmetric

The variance-covariance matrix \(\Sigma\) is a symmetric matrix. This can be shown by noting that

\[ \sigma_{ij} = \mathbf{E}(x_i - \mu_i)(x_j - \mu_j) = \mathbf{E}(x_j - \mu_j)(x_i - \mu_i) = \sigma_{ji}. \]

Due to this symmetry \(\Sigma\) will only have \(m(m + 1)/2\) unique elements.

#### 3.1.2.2 Positive Semidefinite

\(\Sigma\) is a positive semidefinite matrix. Recall that any \(m \times m\) matrix is positive semidefinite if and only if it meets any of the following three equivalent conditions:

1. All the principle minors are nonnegative;
2. \(\lambda'\Sigma\lambda \geq 0\), for all \(\lambda \in \mathbb{R}^m\) \(\neq 0\);
3. \(\Sigma = PP'\), for some \(P \in \mathbb{R}^{m \times m}\).

The first condition (actually we use negative definiteness) is useful in the study of utility maximization while the latter two are useful in econometric analysis.
The necessity of second condition is the easiest to demonstrate in the current context. Let $\lambda \neq 0$. Then, we have

$$
\lambda' \Sigma \lambda = \lambda' [E((x - \mu)(x - \mu)')] \lambda \\
= E[\lambda'(x - \mu)(x - \mu)' \lambda] \\
= E[[\lambda'(x - \mu)]^2] \geq 0,
$$

since the term inside the expectation is a quadratic. Hence, $\Sigma$ is a positive semidefinite matrix.

Note that $P$ satisfying the third relationship is not unique. Let $D$ be any $m \times m$ orthonormal matrix, then $DD' = I_m$ and $P^* = PD$ yields $P^*P'' = PDDD'P' = P[I_m]P' = \Sigma$. Usually, we will choose $P$ to be an upper or lower triangular matrix with $m(m+1)/2$ nonzero elements, which will then make such $P$ unique.

### 3.1.2.3 Positive Definite

Since $\Sigma$ is a positive semidefinite matrix, it will be a positive definite matrix if and only if $\det(\Sigma) \neq 0$. Now, we know that $\Sigma = PP'$ for some $m \times m$ matrix $P$. This implies that $\det(P) \neq 0$.

### 3.1.3 Linear Transformations

Let $y = \frac{b}{m \times 1} + \frac{B}{m \times m} x$. Then

$$
E[y] = b + B E[x] \\
= b + B \mu \\
= \mu_y
$$

Thus, the mean of a linear transformation is the linear transformation of the mean.

Next, we have

$$
E[(y - \mu_y)(y - \mu_y)'] = E[[B (x - \mu)][(B (x - \mu))']] \\
= BE[(x - \mu)(x - \mu)']B' \\
= B \Sigma B' \\
= \Sigma_y
$$

where we use the result $(ABC)' = C'B'A'$, if conformability holds.
3.2 Change Of Variables

3.2.1 Univariate

Let \( x \) be a random variable and \( f_x(\cdot) \) be the probability density function of \( x \). Now, define \( y = h(x) \), where

\[
h'(x) = \frac{d}{dx} h(x) > 0.
\]

That is, \( h(x) \) is a strictly monotonically increasing function and so \( y \) is a one-to-one transformation of \( x \). Now, we would like to know the probability density function of \( y \), \( f_y(y) \). To find it, we note that

\[
\Pr(y \leq h(a)) = \Pr(x \leq a), \quad (3.6)
\]

\[
\Pr(x \leq a) = \int_{-\infty}^{a} f_x(x) \, dx = F_x(a), \quad (3.7)
\]

and,

\[
\Pr(y \leq h(a)) = \int_{-\infty}^{h(a)} f_y(y) \, dy = F_y(h(a)), \quad (3.8)
\]

for all \( a \).

Assuming that the cumulative density function is differentiable, we use (3.6) to combine (3.7) and (3.8), and take the total differential, which gives us

\[
dF_x(a) = dF_y(h(a))
\]

\[
f_x(a) \, da = f_y(h(a)) h'(a) \, da
\]

for all \( a \). Thus, for a small perturbation,

\[
f_x(a) = f_y(h(a)) h'(a) \quad (3.9)
\]

for all \( a \). Also, since \( y \) is a one-to-one transformation of \( x \), we know that \( h(\cdot) \) can be inverted. That is, \( x = h^{-1}(y) \). Thus, \( a = h^{-1}(y) \), and we can rewrite (3.9) as

\[
f_x(h^{-1}(y)) = f_y(y) h'(h^{-1}(y)).
\]

Therefore, the probability density function of \( y \) is

\[
f_y(y) = f_x(h^{-1}(y)) \frac{1}{h'(h^{-1}(y))}. \quad (3.10)
\]

Note that \( f_y(y) \) has the properties of being nonnegative, since \( h'(\cdot) > 0 \). If \( h'(\cdot) < 0 \), (3.10) can be corrected by taking the absolute value of \( h'(\cdot) \), which will assure that we have only positive values for our probability density function.
3.2. CHANGE OF VARIABLES

Example 3.1. Let $y = b_0 + b_1 x$, where $x$, $b_0$, and $b_1$ are scalars. Then

$$x = \frac{y - b_0}{b_1}$$

and

$$\frac{dy}{dx} = b_1.$$

Therefore,

$$f_y(y) = f_x\left(\frac{y - b_0}{b_1}\right) \frac{1}{|b_1|}. \quad \square$$

3.2.2 Geometric Interpretation

Consider the graph of the relationship shown in Figure 3.1. We know that

$$\Pr[h(b) > y > h(a)] = \Pr(b > x > a).$$

Also, we know that

$$\Pr[h(b) > y > h(a)] \simeq f_y[h(b)][h(b) - h(a)],$$

and

$$\Pr(b > x > a) \simeq f_x(b)(b - a).$$
CHAPTER 3. MULTIVARIATE DISTRIBUTIONS

So,

\[ f_y[h(b)][h(b) - h(a)] \approx f_x(b)(b - a) \]

\[ f_y[h(b)] \approx f_x(b)\frac{1}{h(b) - h(a)/(b - a)} \] (3.11)

Now, as we let \( a \to b \), the denominator of (3.11) approaches \( h'(\cdot) \). This is then the same formula as (3.10).

3.2.3 Multivariate

Let

\[ \mathbf{x} \sim f_x(\mathbf{x}). \]

Define a one-to-one transformation

\[ \mathbf{y} = \mathbf{h}(\mathbf{x}). \]

Since \( h(\cdot) \) is a one-to-one transformation, it has an inverse:

\[ \mathbf{x} = h^{-1}(\mathbf{y}). \]

We also assume that \( \frac{\partial h(\mathbf{x})}{\partial \mathbf{x}} \) exists. This is the \( m \times m \) Jacobian matrix, where

\[
\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}'} = \begin{bmatrix}
\frac{\partial h_1(\mathbf{x})}{\partial x_1} & \frac{\partial h_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial h_m(\mathbf{x})}{\partial x_1} \\
\frac{\partial h_1(\mathbf{x})}{\partial x_2} & \frac{\partial h_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial h_m(\mathbf{x})}{\partial x_2} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial h_1(\mathbf{x})}{\partial x_m} & \frac{\partial h_2(\mathbf{x})}{\partial x_m} & \cdots & \frac{\partial h_m(\mathbf{x})}{\partial x_m}
\end{bmatrix} = J_x(\mathbf{x}) \quad (3.12)
\]

Given this notation, the multivariate analog to (3.11) can be shown to be

\[ f_y(\mathbf{y}) = f_x[h^{-1}(\mathbf{y})]\frac{1}{|\text{det}(J_x[h^{-1}(\mathbf{y})])|} \] (3.13)

Since \( h(\cdot) \) is differentiable and one-to-one then \( \text{det}(J_x[h^{-1}(\mathbf{y})]) \neq 0. \)
Example 3.2. Let \( \mathbf{y} = \mathbf{b} + \mathbf{Bx} \), where \( \mathbf{y} \) is an \( m \times 1 \) vector and \( \det(\mathbf{B}) \neq 0 \). Then
\[
\mathbf{x} = \mathbf{B}^{-1}(\mathbf{y} - \mathbf{b})
\]
and
\[
\frac{\partial \mathbf{y}}{\partial \mathbf{x}'} = \mathbf{B} = J_x(\mathbf{x}).
\]
Thus,
\[
f_y(\mathbf{y}) = f_x\left(\mathbf{B}^{-1}(\mathbf{y} - \mathbf{b})\right) \frac{1}{|\det(\mathbf{B})|}. \square
\]

3.3 Multivariate Normal Distribution

3.3.1 Spherical Normal Distribution

Definition 3.1. An \( m \times 1 \) random vector \( \mathbf{z} \) is said to be spherically normally distributed if
\[
f(\mathbf{z}) = \frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}. \square
\]

Such a random vector can be seen to be a vector of independent standard normals. Let \( z_1, z_2, \ldots, z_m \), be i.i.d. random variables such that \( z_i \sim N(0, 1) \). That is, \( z_i \) has pdf given in (2.8), for \( i = 1, \ldots, m \). Then, by independence, the joint distribution of the \( z_i \)'s is given by
\[
f(z_1, z_2, \ldots, z_m) = f(z_1)f(z_2) \cdots f(z_m)
\]
\[
= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_i^2}
\]
\[
= \frac{1}{(2\pi)^{m/2}}e^{-\frac{1}{2}\sum_{i=1}^{m} z_i^2}
\]
\[
= \frac{1}{(2\pi)^{m/2}}e^{-\frac{1}{2}\mathbf{z}'\mathbf{z}}, \quad (3.14)
\]
where \( \mathbf{z}' = (z_1 \ z_2 \ \ldots \ z_m) \).

3.3.2 Multivariate Normal

Definition 3.2. The \( m \times 1 \) random vector \( \mathbf{x} \) with density
\[
f_x(\mathbf{x}) = \frac{1}{(2\pi)^{m/2}|\det(\mathbf{\Sigma})|^{1/2}}e^{-\frac{1}{2}(\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)} \quad (3.15)
\]
is said to be distributed multivariate normal with mean vector \( \mu \) and positive definite covariance matrix \( \mathbf{\Sigma} \). \square

Such a distribution for \( \mathbf{x} \) is denoted by \( \mathbf{x} \sim N(\mu, \mathbf{\Sigma}) \). The spherical normal distribution is seen to be a special case where \( \mu = 0 \) and \( \mathbf{\Sigma} = \mathbf{I}_m \).
There is a one-to-one relationship between the multivariate normal random vector and a spherical normal random vector. Let \( z \) be an \( m \times 1 \) spherical normal random vector and

\[ x = \mu + Az, \]

where \( z \) is defined above, and \( \det(A) \neq 0 \). Then,

\[ \text{Ex} = \mu + \text{AE}z = \mu, \quad (3.16) \]

since \( \text{E}[z] = 0 \).

Also, we know that

\[ \text{E}(zz') = \text{E} \begin{bmatrix} z_1^2 & z_1z_2 & \cdots & z_1z_m \\ z_2z_1 & z_2^2 & \cdots & z_2z_m \\ \vdots & \vdots & \ddots & \vdots \\ z_mz_1 & z_mz_2 & \cdots & z_m^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I_m, \quad (3.17) \]

since \( \text{E}[z_i z_j] = 0 \), for all \( i \neq j \), and \( \text{E}[z_i^2] = 1 \), for all \( i \). Therefore,

\[ \text{E}[(x - \mu)(x - \mu)'] = \text{E}(AA'z z') = \text{AE}(zz')A' = \text{AE}I_m A' = \Sigma, \quad (3.18) \]

where \( \Sigma \) is a positive definite matrix (since \( \det(A) \neq 0 \)).

Next, we need to find the probability density function \( f_x(x) \) of \( x \). We know that

\[ z = A^{-1}(x - \mu), \]

\[ z' = (x - \mu)' A^{-1'}, \]

and

\[ J_z(z) = A, \]

so we use (3.13) to get

\[ f_x(x) = f_z[A^{-1}(x - \mu)] \frac{1}{|\det(A)|}, \]

\[ = \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}(x - \mu)' A^{-1'} A^{-1}(x - \mu)} \frac{1}{|\det(A)|}, \]

\[ = \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}(x - \mu)' (AA')^{-1}(x - \mu)} \frac{1}{|\det(A)|}, \quad (3.19) \]

where we use the results \( (ABC)^{-1} = C^{-1}B^{-1}A^{-1} \) and \( A'^{-1} = A^{-1'} \). However, \( \Sigma = AA' \), so \( \det(\Sigma) = \det(A) \cdot \det(A) \), and \( |\det(A)| = [\det(\Sigma)]^{1/2} \). Thus we can rewrite (3.19) as

\[ f_x(x) = \frac{1}{(2\pi)^{m/2}|\det(\Sigma)|^{1/2}} e^{-\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)}, \quad (3.20) \]
and we see that \( x \sim N(\mu, \Sigma) \) with mean vector \( \mu \) and covariance matrix \( \Sigma \). Since this development is completely reversible the relationship is one-to-one.

The multivariate normal distribution has a number of useful special properties, which will be stated without proof: (i) The conditional distribution of a subvector, say \( x_1 \), given the remainder \( x_2 \), is normally distributed, specifically

\[
x_1 \mid x_2 \sim N(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}').
\]

(ii) The unconditional distribution of any subvector, say \( x_1 \), has a multivariate normal distribution, so

\[
x_1 \sim N(\mu_1, \Sigma_{11}).
\]

(iii) Consequently, the elements of any subvector, say \( x_1 \), are jointly normal and independent of the remainder \( x_2 \), if and only if \( \Sigma_{12} = 0 \).

### 3.3.3 Linear Transformations

**Theorem 3.1.** Suppose \( x \sim N(\mu, \Sigma) \) with \( \det(\Sigma) \neq 0 \) and \( y = b + Bx \) with \( B \) square and \( \det(B) \neq 0 \). Then \( y \sim N(\mu_y, \Sigma_y) \).

**Proof:** From (3.3) and (3.4), we have \( \mathbb{E}y = b + B\mu = \mu_y \) and \( \mathbb{E}[(y - \mu_y)'(y - \mu_y)'] = B\Sigma B' = \Sigma_y \). To find the probability density function \( f_y(y) \) of \( y \), we again use (3.13), which gives us

\[
f_y(y) = f_x[B^{-1}(y - b)] \left| \frac{1}{\det(B)} \right| = \frac{1}{(2\pi)^{m/2} |\det(\Sigma)|^{1/2}} e^{-\frac{1}{2}(y - b - B\mu)'\Sigma^{-1}(y - b - B\mu)} \left| \frac{1}{\det(B)} \right| = \frac{1}{(2\pi)^{m/2} |\det(B\Sigma B')|^{1/2}} e^{-\frac{1}{2}(y - b - B\mu)'(B\Sigma B')^{-1}(y - b - B\mu)}
\]

\[
= \frac{1}{(2\pi)^{m/2} |\det(\Sigma_y)|^{1/2}} e^{-\frac{1}{2}(y - \mu_y)'\Sigma_y^{-1}(y - \mu_y)} \tag{3.21}
\]

So, \( y \sim N(\mu_y, \Sigma_y) \).

Thus we see, as asserted in the previous chapter, that linear transformations of multivariate normal random variables are also multivariate normal random variables. And any linear combination of independent normals will also be normal, since by (iii) in the previous subsection, independent normals are also jointly normal.

### 3.3.4 Quadratic Forms

**Theorem 3.2.** Let \( x \sim N(\mu, \Sigma) \), where \( \det(\Sigma) \neq 0 \), then \( (x - \mu)'\Sigma^{-1}(x - \mu) \sim \chi^2_n \).

**Proof** Let \( \Sigma = PP' \), which is possible since \( \Sigma \) positive definite. Then

\[
(x - \mu) \sim N(0, \Sigma),
\]

and

\[
z = P^{-1}(x - \mu) \sim N(0, I_m).
\]
Therefore,

\[ z'z = \sum_{i=1}^{n} z_i^2 \sim \chi^2_m \]

\[ = P^{-1}(x - \mu)'P^{-1}(x - \mu) \]

\[ = (x - \mu)'P^{-1}P^{-1}(x - \mu) \]

\[ = (x - \mu)'\Sigma^{-1}(x - \mu) \sim \chi^2_m. \quad \square \tag{3.22} \]

\( \Sigma^{-1} \) is called the weight matrix. With this result, we can use the \( \chi^2_m \) to make inferences about the vector mean \( \mu \) of \( x \).

### 3.3.5 Singular Multivariate Normal

On occasion we will encounter an \( m \times 1 \) vector of random variables \( x \) with mean vector \( \mu \) that is jointly normal but with a singular covariance matrix \( \Sigma \). Suppose \( \Sigma \) has rank \( m_1 \), then, without loss of generality, we suppose the last \( m_2 \) columns (rows) of \( \Sigma \) are linearly dependent on the first \( m_1 \) and we can write

\[
\Sigma = \begin{pmatrix}
    \Sigma_{11} & \Sigma_{12} \\
    \Sigma_{12}' & \Sigma_{22}
\end{pmatrix} = \begin{pmatrix}
    \Sigma_{11} & \Sigma_{11}'B' \\
    BS_{11} & BS_{11}'B'
\end{pmatrix} = \begin{pmatrix}
    I_{m_1} & 0 \\
    B & I_{m_1}
\end{pmatrix}
\]

for some \( m_2 \times m_1 \) matrix \( B \) and \( \Sigma_{11} \) is \( m_1 \times m_1 \) nonsingular. Define, conformably, \( x' = (x_1', x_2') \) and \( \mu' = (\mu_1', \mu_2') \). This singularity implies that \( (x_2 - \mu_2) = B(x_1 - \mu_1) \) since the covariance of \( v = (x_2 - \mu_2) - B(x_1 - \mu_1) \), which must also be multivariate normal, has covariance matrix \( 0 \) and hence must be identically zero.

Consider a square matrix \( A \), then \( A^+ \) is a generalized inverse if \( AA^+ A = A \). Such generalized inverses are not unique and we typically append additional restrictions to obtain a unique solution such as the Moore-Penrose pseudo-inverse. For the problem at hand, it is easily shown that

\[
\Sigma^+ = \begin{pmatrix}
    I_{m_1} \\
    0
\end{pmatrix} \Sigma^{-1}_{11} \begin{pmatrix}
    I_{m_1} & 0
\end{pmatrix}
\]

is a generalized inverse of \( \Sigma \). Now consider the quadratic form of the previous section but with a generalized inverse used as the weight matrix

\[
(x - \mu)'\Sigma^+(x - \mu) = (x_1 - \mu_1)' (x_2 - \mu_2)' \begin{pmatrix}
    I_{m_1} \\
    0
\end{pmatrix} \Sigma^{-1}_{11} \begin{pmatrix}
    I_{m_1} & 0
\end{pmatrix} \begin{pmatrix}
    (x_1 - \mu_1) \\
    (x_2 - \mu_2)
\end{pmatrix}
\]

\[ = (x_1 - \mu_1)'\Sigma^{-1}_{11}(x_1 - \mu_1) \sim \chi^2_{\text{rank}(\Sigma)}. \quad \tag{3.23} \]

Thus weighting by the generalized inverse reduces the quadratic form to a quadratic form involving the linearly independent sub-vector \( x_1 \). The value of the quadratic form will be the same regardless of which generalized inverse we choose. Note that (3.22), the result of the previous section where \( \Sigma \) is nonsingular with rank \( m \), is a special case of (3.23).
3.4. NORMALITY AND THE SAMPLE MEAN

Writing the probability density function of a singular normal takes some care. The joint density of \( x_1 \), which is multivariate normal with nonsingular covariance, is

\[
f_{x_1}(x_1) = \frac{1}{(2\pi)^{m_1/2}[\det(\Sigma_{11})]^{1/2}} e^{-\frac{1}{2}(x_1-\mu_1)'\Sigma_{11}^{-1}(x_1-\mu_1)}.
\]

For \( x \) satisfying \( B(x_1 - \mu_1) - (x_2 - \mu_2) = 0 \), then the joint probability density function becomes

\[
f_x(x) = \frac{1}{(2\pi)^{m_1/2}[\det^+(\Sigma)]^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^+(x-\mu)}
\]

where \( \det^+(\Sigma) = \det(\Sigma_{11}) \) is the pseudo-determinant. For \( x \) not satisfying this restriction the density is zero. The joint density does not have support everywhere in the \( m \)-space of \( x \), but only on an \( m_1 \)-dimensional manifold. Typically, we will dispense with the linear dependent variables \( x_2 \) and focus on the linear independent variables \( x_1 \). Using the generalized inverse does this automatically with the quadratic forms.

3.4 Normality and the Sample Mean

3.4.1 Moments of the Sample Mean

Consider the \( m \times 1 \) vector \( x_i \sim i.i.d. \) jointly, with \( m \times 1 \) vector mean \( \mathbf{E}[x_i] = \mu \) and \( m \times m \) covariance matrix \( \mathbf{E}[(x_i - \mu)(x_i - \mu)'] = \Sigma \). Define \( \mathbf{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \) as the vector sample mean which is also the vector of scalar sample means. The mean of the vector sample mean follows directly:

\[
\mathbf{E}[\mathbf{x}_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}[x_i] = \mu.
\]

Alternatively, this result can be obtained by applying the scalar results element by element. The second moment matrix of the vector sample mean is given by

\[
\mathbf{E}[(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)'] = \frac{1}{n^2} \mathbf{E} \left[ \left( \sum_{i=1}^{n} x_i - n\mu \right) \left( \sum_{i=1}^{n} x_i - n\mu \right)' \right]
\]

\[
= \frac{1}{n^2} \mathbf{E} \left[ \left( (x_1 - \mu) + (x_2 - \mu) + ... + (x_n - \mu) \right) \left( (x_1 - \mu) + (x_2 - \mu) + ... + (x_n - \mu) \right)' \right]
\]

\[
= \frac{1}{n^2} n\Sigma = \frac{1}{n} \Sigma
\]

since the covariances between different observations are zero.
CHAPTER 3. MULTIVARIATE DISTRIBUTIONS

3.4.2 Distribution of the Sample Mean

Suppose \( x_i \sim i.i.d. N(\mu, \Sigma) \) then they are all jointly normal with zero covariance between observations. It follows from joint multivariate normality that \( \mathbf{x}_n \) must also be multivariate normal since it is a linear transformation. Specifically, we have

\[
\mathbf{x}_n \sim N(\mu, \frac{1}{n} \Sigma)
\]

or equivalently

\[
\mathbf{x}_n - \mu \sim N(0, \frac{1}{n} \Sigma) \\
\sqrt{n}(\mathbf{x}_n - \mu) \sim N(0, \Sigma) \\
\sqrt{n} \Sigma^{-1/2} (\mathbf{x}_n - \mu) \sim N(0, I_m)
\]

where \( \Sigma = \Sigma^{1/2} \Sigma^{1/2} \) and \( \Sigma^{-1/2} = (\Sigma^{1/2})^{-1} \) for \( \Sigma \) nonsingular. This transformation is the multivariate generalization of the "z-transformation" for scalars.

In order to make a probability statement about this vector, we must consider a scalar transformation. Using the results from above, we find that the quadratic form

\[
n \cdot (\mathbf{x}_n - \mu)' \Sigma^{-1} (\mathbf{x}_n - \mu) = n \cdot (\mathbf{x}_n - \mu)'(\Sigma^{1/2} \Sigma^{1/2})^{-1}(\mathbf{x}_n - \mu)
\]

\[
= [n \cdot (\mathbf{x}_n - \mu)' \Sigma^{-1/2} \Sigma^{-1/2}(\mathbf{x}_n - \mu)]
\]

\[
= [\sqrt{n} \Sigma^{-1/2} (\mathbf{x}_n - \mu)]' [\sqrt{n} \Sigma^{-1/2} (\mathbf{x}_n - \mu)]
\]

\[
\sim \chi^2_m.
\]

Under a null hypothesis regarding all the means, say \( H^0 : \mu = \mu^0 \), we have

\[
n \cdot (\mathbf{x}_n - \mu)' \Sigma^{-1} (\mathbf{x}_n - \mu^0) \sim \chi^2_m.
\]

Under an alternative we will have a shift or noncentrality which is studied below. This quadratic form approach is central to testing hypothesis involving multiple random variables.

3.4.3 Multivariate Central Limit Theorem

Theorem 3.3. Suppose that (i) \( x_i \sim i.i.d \) jointly, (ii) \( E[x_i] = \mu \), and (iii) \( E[(x_i - \mu)(x_i - \mu)'] = \Sigma \), then

\[
\sqrt{n}(\mathbf{x}_n - \mu) \rightarrow_d N(0, \Sigma) \nabla.
\]

Equivalently, we can write, analogous to the scalar z-transformation,

\[
z = \sqrt{n} \Sigma^{-1/2} (\mathbf{x}_n - \mu) \rightarrow_d N(0, I_m).
\]

These results apply even if the original underlying distribution is not normal and follow directly from the scalar results applied to any linear combination of \( \mathbf{x}_n \).
Correspondingly, under the assumptions of the theorem, the following quadratic form

\[ n \cdot (\mathbf{x}_n - \mu)' \Sigma^{-1} (\mathbf{x}_n - \mu) \rightarrow_d \chi^2_m. \]

This quadratic form is convenient for testing an hypothesis that specifies the value of all the means, i.e. \( H_0 : \mu = \mu^0 \). If \( \sqrt{n}(\mathbf{x}_n - \mu) \rightarrow_d N(0, \Sigma) \) and \( \Sigma \) is singular, then we can use the generalized inverse as the weight matrix in the quadratic form and the limiting distribution will be \( \chi^2_{\text{rank}(\Sigma)} \).

### 3.5 Noncentral Distributions

When we conduct an hypothesis test we have a stated null hypothesis \( H_0 \) and at least implicitly, stated alternative hypothesis \( H_1 \). As a matter of convention, we carefully establish the behavior of our test statistic under the null, such as the standard normal, for example. The fact is that it is quite easy to develop a statistic with a known distribution and hence known size or probability of rejecting under the null. What is interesting is what happens to the statistic when the alternative applies. This means we need to establish the power of a test, which is the probability of rejecting under the alternative. We choose between equal sized tests on the basis of which has the most power. And we prefer consistent tests whose power improves with sample size and reject false nulls with probability one in the limit.

#### 3.5.1 Noncentral Scalar Normal

**Definition 3.3.** Suppose \( x \sim N(\mu, \sigma^2) \), then, for \( \mu^* = \mu/\sigma \),

\[ z^* = \frac{x}{\sigma} \sim N(\mu^*, 1) \]  

(3.24)

has a noncentral normal distribution with noncentrality parameter \( \mu^* \). □

**Example 3.3.** Suppose \( x_i \sim i.i.d. N(\mu, \sigma^2) \) for \( i = 1, 2, ..., n \). When we do a hypothesis test of the mean, with known variance, we have, under the null hypothesis \( H_0 : \mu = \mu_0 \),

\[ S_n = \frac{x_n - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1) \]  

(3.25)

and, under the alternative \( H_1 : \mu = \mu_1 \neq \mu_0 \),

\[ S_n = \frac{\bar{x}_n - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\bar{x}_n - \mu_1}{\sqrt{\sigma^2/n}} + \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/n}} \]

\[ = N(0, 1) + \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/n}} \sim N \left( \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2/n}}, 1 \right). \]  

(3.26)
Thus, the behavior of $S_n$ under the null hypothesis follows the standard normal distribution but under the alternative hypothesis follows a noncentral normal distribution. Note how the noncentrality is increasing at the rate $\sqrt{n}$ so the test will be consistent and reject a false null with certainty in large samples.

□

As this example makes clear, the noncentral normal distribution is especially useful when carefully exploring the behavior of the alternative hypothesis.

### 3.5.2 Noncentral t

**Definition 3.4.** Suppose $z^* \sim N(\mu^*, 1)$, $w \sim \chi^2_k$, and let $z^*$ and $w$ are independent, then

$$T = \frac{z^*}{\sqrt{w/k}} \sim t_k(\mu^*)$$

(3.27)

has a noncentral $t$ distribution with noncentrality parameter $\mu^*$. □

The noncentral $t$ distribution is used in calculating the power of tests of the mean, when the variance is unknown and must be estimated.

Unlike the noncentral normal, the noncentral $t$ is more than a central $t$, which has zero mean, shifted by $\mu^*$. Let $z = z^* - \mu^*$, then, as before, we can write

$$\frac{z^*}{\sqrt{w/k}} = \frac{z}{\sqrt{w/k}} + \frac{\mu^*}{\sqrt{w/k}}.$$

The distribution after the shift is now a rather complicated function of both $\mu^*$ and $k$. The first term is central $t$ but the second or shift term is now stochastic so the distribution involves more than a shift by the constant $\mu^*$.

In particular, the distribution of the second term is bell-shaped but skewed to the right so the sum is also asymmetric with right-hand tail heavier than the left unless $\mu^* = 0$. Consequently, the median exceeds $\mu^*$ with the positive difference converging toward zero as $k$ becomes large. In this sense, the noncentral $t$ distribution is more right-shifted than the noncentral normal, which has median $\mu^*$. For small values of $k$, however, both the first and second terms will have fatter distributions as will their sum so the net shift in probability to the right of a critical value will not increase. As $k$ grows large, the distribution becomes more symmetric about $\mu^*$ and distribution converges to the noncentral normal since $w/k \to \mu^*$.

The complications that arise with the distribution can be illustrated with the mean and variance. Although $E[w/k] = 1$, due to Jensen’s inequality $E[1/\sqrt{w/k}] > 1/E[w/k] = 1$, so $E[T] > \mu^*$. Specifically, we find, for $k > 1$,

$$E[T] = \mu^* \sqrt{\frac{k}{2}} \frac{\Gamma((k-1)/2)}{\Gamma(k/2)}$$

which obviously depends on the degrees of freedom in a complicated way. Similarly, for $k > 2$, we find

$$\text{Var}[T] = \frac{k^2 \Gamma((k-1)/2)}{2} \left( \frac{\Gamma((k-1)/2)}{\Gamma(k/2)} \right)^2.$$  

An excellent approximation to $\sqrt{\frac{k}{2}} \frac{\Gamma((k-1)/2)}{\Gamma(k/2)}$ in both these moments is $\left( 1 - \frac{3}{4 \pi k} \right)^{-1}$, which clearly approaches 1 as $k$ grows large. Thus
as \( k \) grows large \( \mathbb{E}[T] \to \mu^* \) and \( \text{Var}[T] \to 1 \), which would be the moments of a central \( t \) shifted by \( \mu^* \), which, in turn, converge to the moments of the noncentral normal.

**Example 3.4.** Consider the Example 3.3 above but we don’t know \( \sigma^2 \) and use the unbiased estimator \( s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - 1) \). Now from the last chapter we have

\[
Y = (n - 1) \frac{s^2}{\sigma^2} \sim \chi^2_{n-1}
\]

while from above

\[
Z = \frac{\bar{x}_n - \mu_0}{\sigma^2 / n} \sim N \left( \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2 / n}}, 1 \right),
\]

thus under the alternative hypothesis

\[
S_n = \frac{Z}{Y} = \frac{\bar{x}_n - \mu_0}{s^2 / n} \sim t_{n-1} \left( \frac{\mu_1 - \mu_0}{\sqrt{\sigma^2 / n}} \right).
\]

Note that the alternative distribution uses \( \sigma^2 \) rather than \( s^2 \) in the noncentrality parameter, which increases at the rate \( \sqrt{n} \).

It is instructive to examine the power of a two-sided 5% test implied for this example for various values of the noncentrality as given in Figure 3.2. Let \( d = (\mu_1 - \mu_0) / \sigma \) then the noncentrality is \( \sqrt{n}d \) and we can plot the probability of exceeding the critical value \( \tau_{0.025, n-1} \) in absolute value for various choices of \( d \) and \( n \). Notice for each choice of \( n \) the curves have a minimum around \( d = 0 \) at the ostensible size under the null of .05. And they asymptote to 1 as \( d \) diverges from zero in the positive or negative direction. This is the typical shape of the power curve for a two-sided test. Also note that the curves become more compact around zero as the sample size \( n \) increases. In the limit it will approach a spike at zero which reflects the fact that the test is consistent.

### 3.5.3 Noncentral Chi-Squared

**Definition 3.5.** Suppose \( z^* \sim N(\mu^*, I_m) \), then

\[
V = z^* z^* \sim \chi^2_m(\delta),
\]

has a noncentral chi-squared distribution, where \( \delta = \mu^{**} \mu^* \) is the noncentrality parameter.

As with the noncentral \( t \), the noncentral chi-squared is more than a central chi-squared shifted by a constant. Let \( z = z^* - \mu^* \), then

\[
z^* z^* = (z + \mu^*)' (z + \mu^*) = z' z + 2z' \mu^* + \mu^{**} \mu^*.
\]

The first term is a central chi-squared and the last is the noncentrality parameter \( \delta \) which shifts the distribution to the right, but the middle term is \( N(0, 4\delta) \) and
can shift the distribution to the right or left with the sum restricted to be non-negative. The density is given by
\[ f_V(v; k, \delta) = \sum_{i=0}^{\infty} \frac{e^{-\delta/2} (\delta/2)^i}{i!} f_{Y_k+2i}(v), \]
where \( f_{Y_q}(\cdot) \) is the density of a central chi-square with \( q \) degrees of freedom. The weights in the summation are the weights of a Poisson distribution with mean and variance \( \delta/2 \). For a given \( k \), increases in \( \delta \) will shift more and more of the distribution to the right as the larger weights of the Poisson, which has mode at floor(\( \delta/2 \)), shift to the right. The mean of \( v \) is \( k + \delta \) and the variance is \( 2(k + 2\delta) \).

**Example 3.5.** Consider a test of the means for \( x_i \sim i.i.d. N(\mu, \Sigma) \) for \( \Sigma \) non-singular. For \( H_0 : \mu = \mu_0 \), we have
\[ n \cdot (\bar{x}_n - \mu_0)' \Sigma^{-1} (\bar{x}_n - \mu_0) \sim \chi^2_m \]
Let \( z^* = P^{-1} \sqrt{n}(\bar{x}_n - \mu_0) \) for \( \Sigma = PP' \), then for \( H_1 : \mu = \mu_1 \neq \mu_0 \), we have
\[ n \cdot (\bar{x}_n - \mu_0)' \Sigma^{-1} (\bar{x}_n - \mu_0) = \sqrt{n}(\bar{x}_n - \mu_0)'P^{-1}P^{-1}\sqrt{n}(\bar{x}_n - \mu_0) = z^* z^* \sim \chi^2_m \left( (n \cdot \mu_1 - \mu_0)' \Sigma^{-1} (\mu_1 - \mu_0) \right) \]
Note that the noncentrality is always non-negative so the probability mass of the distribution under the alternative is shifted to the right. Consequently, we focus only on the right-hand side tail of the null distribution. Moreover, the noncentrality will increase at the rate \( n \), so the test will be consistent. □
3.5. NONCENTRAL DISTRIBUTIONS

3.5.4 Noncentral F

Definition 3.6. Let \( Y \sim \chi^2_m(\delta) \), \( W \sim \chi^2_n \), and let \( Y \) and \( W \) be independent random variables. Then

\[
\frac{Y}{m} \sim F_{m,n}(\delta),
\]

has a noncentral F distribution, where \( \delta > 0 \) is the noncentrality parameter. \( \square \)

The density is given by

\[
f_{F_{m,n,\delta}}(x) = \sum_{i=0}^{\infty} \frac{e^{-\delta/2} (\delta/2)^i}{i!} \left[ \frac{\Gamma \left( \frac{2i+m+n}{2} \right) (m/n)^{(2i+m)/2}}{\Gamma \left( \frac{2i+m}{2} \right) \Gamma \left( \frac{n}{2} \right)} \right] \frac{x^{(2i+m/2)-1}}{(1 + mx/n)^{(2i+m+n)/2}},
\]

which simplifies to the central \( F_{m,n} \) for \( \delta = 0 \). Notice that the summation has the same Poisson weights as appeared in the noncentral chi-squared density. For \( i = 0 \), the term in brackets in the summation is the density of the central \( F_{m,n} \), so this density reduces to the central when \( \delta = 0 \). For \( i > 0 \) the additional terms shift the density to the right when \( \delta > 0 \).

The noncentral F distribution is used in tests of mean vectors, where the scale of the covariance matrix is unknown and must be estimated. As with the other noncentral distributions the noncentrality parameter determines the amount the distribution is shifted (to the right in this case) when nonzero. Such a shift would arise under an alternative hypothesis as with the chi-squared example above. Examples for both the central and noncentral F distribution are deferred until Chapter 8.
Chapter 4

Asymptotic Theory

In Chapter 2 we found, via the central limit theorem that the behavior of an average, under very general assumptions and when properly normalized, tends toward the standard normal distribution even when the underlying distribution is non-normal. In Chapter 3, we extended this result to multivariate averages. These are both examples of depending on the large-sample limiting or asymptotic behavior of the average to simplify our analysis. As we develop appropriate tools for more complicated models in later chapters we will encounter numerous cases of estimators that, as a result of non-normality of the underlying variables or nonlinearity with respect these variables, have distributions sufficiently complicated to render finite sample inference infeasible. In most of these cases, however, the large-sample limiting or asymptotic behavior is tractable. Accordingly, in this chapter, we develop a number of asymptotically appropriate tools that will be useful for these cases.

4.1 Convergence Of Random Variables

4.1.1 Sequences Of Random Variables

In the following we will consider the behavior of an estimator or statistic, based on a sample of size \( n \), as the sample size grows large. Let \( y_n \) denote a realization of this statistic. This realization is a random draw from the underlying distribution of \( y_n \) which may well change as the sample size \( n \) increases. For the various values of \( n \), these draws form a sequence of random variables \( \{y_1, y_2, ..., y_n, ...\} \).

An example of such a sequence is the sample average \( y_n = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \). Each realization from this sequence is a random draw and cannot be known \textit{a priori}, however, probability measures based on the distribution of each realization are real numbers and hopefully exhibit regularity and even simplicity as the sample size increases. Since such probability measures map each draw in the sequence onto the real line, they form a sequence of real numbers, whose behavior we now examine.
4.1. CONVERGENCE OF RANDOM VARIABLES

4.1.2 Convergence and Limits

We first examine the familiar concept of the limit from calculus.

Definition 4.1. A sequence of real numbers \( a_1, a_2, \ldots, a_n, \ldots \), is said to be convergent to or have a limit of \( \alpha \) if for every \( \delta > 0 \), there exists a positive real number \( N \) such that for all \( n > N \), \( |a_n - \alpha| < \delta \). This is denoted as

\[
\lim_{n \to \infty} a_n = \alpha.
\]

Example 4.1. Let \( \{ a_n \} = 3 + 1/n \), then clearly

\[
\lim_{n \to \infty} a_n = 3,
\]

since \( |a_n - 3| = |1/n| \) can be made arbitrarily small for sufficiently large \( n \).

We frequently encounter sequences of functions of variables whose value at convergence depends on the value of the variables. For example, the sequence of functions \( \{ a_n(x) \} = x(3 + 1/n) \) converges to \( \alpha(x) = 3x \) for any given \( x \). Here, the value of \( n \) needed to make \( |a_n(x) - \alpha(x)| = |x/n| < \delta \) will depend on \( x \). This dependence of the limiting process on the point of evaluation necessitates the introduction of the following uniformity definition

Definition 4.2. A sequence of functions \( \{ f_n \} \), \( n = 1, 2, 3, \ldots \), is said to be uniformly convergent to the function \( f \) for a set \( E \) of values of \( x \) if, for each \( \delta > 0 \), an integer \( N \) can be found such that \( |f_n(x) - f(x)| < \delta \) for \( n > N \) and \( x \in E \).

Notationally, this means that we can define \( a_n = \sup_{x \in E} |f_n(x) - f(x)| \) and \( \alpha = 0 \) and apply the standard definition of convergence to choose \( \delta \) and the corresponding \( n \). For the simple example \( \{ a_n(x) \} = x(3 + 1/n) \) we define \( \alpha_n = \sup_{x \in E} |x/n| = \sup_{x \in E} |x|/n \) which will be finite for bounded \( x \) so \( E \) can simply be a closed interval on the real line.

4.1.3 Orders of Sequences

We now introduce the notion of orders of a sequence, which will play an important role in the sequel. A sequence of interest may not have a limit and, in fact, may be divergent but a simple transformation may yield a limit and thereby yield insight into the rate at which the sequence diverges.

Definition 4.3. A sequence of real numbers \( \{ a_n \} \) is said to be of order at most \( n^k \), and we write \( \{ a_n \} \) is \( O(n^k) \), if

\[
\lim_{n \to \infty} \frac{1}{n^k} a_n = c,
\]

where \( c \) is any real constant.
Example 4.2. Let \( \{ a_n \} = 3 + 1/n \), and \( \{ b_n \} = 4 - n^2 \). Then, \( \{ a_n \} \) is \( O(1) = O(n^0) \), since, from the previous example,

\[
\lim_{n \to \infty} a_n = 3
\]

and \( \{ b_n \} \) is \( O(n^2) \), since \( \lim_{n \to \infty} \frac{1}{n^2}b_n = -1 \). □

**Definition 4.4.** A sequence of real numbers \( \{ a_n \} \) is said to be of order smaller than \( n^k \), and we write \( \{ a_n \} \) is \( o(n^k) \), if

\[
\lim_{n \to \infty} \frac{1}{n^k}a_n = 0. \quad □
\]

Example 4.3. Let \( \{ a_n \} = 6/n \). Then, \( \{ a_n \} \) is \( o(1) \), since \( \lim_{n \to \infty} \frac{1}{n^2}a_n = 0 \). More informatively, we see that \( \{ a_n \} \) is also \( O(n^{-1}) \). Note that negative values of the exponent are possible, which indicates the rate at which the sequence converges to its’ limit point. □

The order of a sequence, like its’ limit, is a simplification relative to the sequence itself but more informative than the limit since it reveals the rate at which the divergence or convergence occurs. There is an algebra for orders of sequences that will be useful later. Suppose \( a_n = O(n^k) \) and \( b_n = O(n^l) \), then it is straightforward to show that \( c_n = a_n \cdot b_n = O(n^k) \cdot O(n^l) = O(n^{k+l}) \) and \( c_n = a_n + b_n = O(n^k) + O(n^l) = O(n^{\max(k,l)}) \). Similarly, for \( a_n = o(n^k) \) and \( b_n = o(n^l) \), then \( c_n = a_n \cdot b_n = o(n^k) \cdot o(n^l) = o(n^{k+l}) \) and \( c_n = a_n + b_n = o(n^k) + o(n^l) = o(n^{\max(k,l)}) \). Finally, if \( a_n = O(n^k) \) and \( b_n = o(n^l) \), then \( c_n = a_n \cdot b_n = O(n^k) \cdot o(n^l) = o(n^{k+l}) \) however \( c_n = a_n + b_n = O(n^k) + o(n^l) = 1(l > k) \cdot o(n^l) + 1(l \leq k) \cdot O(n^k) \), where \( 1(\cdot) \) is the indicator function.

### 4.1.4 Convergence In Probability

We now turn to the large-sample limiting behavior of sequences of random variables such as the average. A probability measure on such a random sequence maps each realization onto the real line and forms a sequence of real numbers such as those above. We first look at a sequence of random variables whose distribution is collapsing about some limit point.

**Definition 4.5.** A sequence of random variables \( y_1, y_2, \ldots, y_n, \ldots \), with distribution functions \( F_1(\cdot), F_2(\cdot), \ldots, F_n(\cdot), \ldots \), is said to converge (weakly) in probability to some constant \( c \) if

\[
\lim_{n \to \infty} \Pr[|y_n - c| > \epsilon] = 0 \quad (4.1)
\]

for every real number \( \epsilon > 0 \). □

Weak convergence in probability is denoted by

\[
\text{plim } y_n = c,
\]

\[
\lim_{n \to \infty} y_n = c.
\]
or sometimes,
\[ y_n \xrightarrow{p} c, \]
or
\[ y_n \xrightarrow{} c. \]

This definition is equivalent to saying that we have a sequence of tail probabilities (of being greater than \( c + \epsilon \) or less than \( c - \epsilon \)), and that the tail probabilities approach 0 as \( n \to \infty \), regardless of how small \( \epsilon \) is chosen. Equivalently, the probability mass of the distribution of \( y_n \) is collapsing about the point \( c \). The leading example of such behavior is given by the laws of large numbers which are discussed in a subsequent subsection.

**Definition 4.6.** A sequence of random variables \( y_1, y_2, \ldots, y_n, \ldots \), is said to **converge strongly in probability** to some constant \( c \) if

\[
\Pr[\lim_{n \to \infty} y_n = c] = 1. \quad (4.2)
\]

Strong convergence is also called **almost sure convergence**, which is denoted

\[ y_n \xrightarrow{a.s.} c, \]
or
\[ y_n \xrightarrow{a.s.} c. \]

Another obvious terminology is to say that \( y_n \) **converges to \( c \) with probability one**.

This definition would seem to involve a probability statement about a non-probabilistic sequence of reals with the limit statement. We can think about all the possible realizations of the sequence of random variables. Each realization has a probability measure attached to it. We can place the possible outcomes into the set of those that converge to a limit and another set of those that do not. Strong convergence in probability states that the total probability measure of the set of realizations that do not possess a limit is zero.

The relationship between weak and strong convergence is best seen by comparison to an alternative and equivalent (but not obviously) definition for strong convergence. Suppose

\[
\lim_{N \to \infty} \Pr[\sup_{n > N} |y_n - c| > \epsilon] = 0 \quad (4.3)
\]
for any \( \epsilon > 0 \). Consider the equivalent statement

\[
\lim_{N \to \infty} \Pr[\sup_{n > N} |y_n - c| < \epsilon] = 1
\]
for any \( \epsilon > 0 \), which, using the definition of a limit, can be rewritten as

\[
\left| \Pr[\sup_{n > N} |y_n - c| < \epsilon] - 1 \right| < \delta
\]
for any $\delta > 0$ and $N > N^*$ for $N^*$ sufficiently large. Now for any $\epsilon > 0$ and $N$ sufficiently large, the condition within the $Pr$ for this formulation can, by the definition of a limit, also be written $\lim_{n \to \infty} y_n = c$. Thus for $N$ sufficiently large and any $\delta > 0$ we have

$$\left| Pr\left[ \lim_{n \to \infty} y_n = c \right] - 1 \right| < \delta,$$

which implies (4.2). Since these steps can be reversed, then (4.3) implies and is implied by the definition of strong convergence (4.2).

Comparison of the weak convergence condition (4.1) with the alternative strong convergence condition (4.3), reveals that, within the limit, weak convergence concerns itself with the probability of a single realization $y_n$ meeting the condition $|y_n - c| < \epsilon$ while strong convergence involves the joint probability of $y_n$ and all subsequent realizations meeting the condition. Obviously, if a sequence of random variables converges strongly in probability, it converges weakly in probability. Equally obviously, the reverse is not necessarily true. Most of our attention in the sequel will focus on weak convergence so when a sequence is said to converge in probability it can be taken as weak convergence unless otherwise noted.

**Definition 4.7.** A sequence of random variables $y_1, y_2, \ldots, y_n, \ldots$, is said to converge in quadratic mean if

$$\lim_{n \to \infty} E[y_n] = c$$

and

$$\lim_{n \to \infty} \text{Var}[y_n] = 0. \quad \square$$

For a random variable $x$ with mean $\mu$ and variance $\sigma^2$, Chebyshev’s inequality states $Pr(|x - \mu| \geq k\sigma) \leq \frac{1}{k^2}$ for $k > 0$. For arbitrary $\epsilon > 0$, let $\mu_n$ and $\sigma_n^2$ denote the mean and variance of $y_n$, and for $n$ sufficiently large that $|c - \mu_n| < \epsilon$, consider

$$Pr[|y_n - c| > \epsilon] = 1 - Pr[\epsilon - c < y_n < c + \epsilon]$$

$$\leq 1 - Pr[|c - \mu_n| - \epsilon < y_n - \mu_n < -|c - \mu_n| + \epsilon], \text{ since the interval is narrower}$$

$$= 1 - Pr[\mu_n - (\alpha_n + \epsilon) < y_n - \mu_n < \mu_n + (\alpha_n + \epsilon)], \text{ where } \alpha_n = -|c - \mu_n|$$

$$= Pr[|y_n - \mu_n| > \alpha_n + \epsilon]$$

$$= Pr[|y_n - \mu_n| > \frac{\alpha_n + \epsilon}{\sigma_n} \sigma_n]$$

$$\leq \left( \frac{1}{\frac{\alpha_n + \epsilon}{\sigma_n}} \right)^2 = \frac{\sigma_n^2}{(\alpha_n + \epsilon)^2} \to 0$$

by Chebyshev’s inequality. Thus convergence in quadratic mean implies convergence in probability. The reverse is not necessarily true. There is no direction of implication between strong convergence and convergence in quadratic mean.
4.1. CONVERGENCE OF RANDOM VARIABLES

The random sequences we consider will sometimes be a function of a variable (and/or parameter) and the constants $\varepsilon$ and $\delta$ will depend on the point of evaluation of the function. Analogous to uniform convergence of a function we introduce the following:

**Definition 4.8.** A sequence of random functions $\{f_n = f_n(x)\}$, $n = 1, 2, 3, \ldots$, is said to be *uniformly convergent in probability* to the function $f$ for a set $E$ of values of $x$ if, for each $\varepsilon > 0$ and $1 > \delta > 0$, an integer $N$ can be found such that $\Pr(|f_n(x) - f(x)| > \varepsilon) < \delta$ for $n > N$ and $x \in E$. □

This definition differs from pointwise convergence in probability in that $\varepsilon$ and $\delta$ do not depend on the value of $x$. In pointwise convergence these values can differ depending on the point of evaluation for $x$. Uniform convergence will sometimes be denoted by

$$\sup_{x \in E} |f_n(x) - f(x)| \longrightarrow_p 0,$$

which assures we are using the same constants $\varepsilon$ and $\delta$ across the set $E$.

### 4.1.5 Orders In Probability

Beyond knowing that the distribution is collapsing around a limit point as the sample size grows large we may be interested in the rate at which it collapses. This is sometimes important in comparing estimators and also in determining the appropriate normalization for finding the limiting distribution in the next subsection.

**Definition 4.9.** Let $y_1, y_2, \ldots, y_n, \ldots$ be a sequence of random variables. This sequence is said to be *bounded in probability* if for any $1 > \delta > 0$, there exist a $\Delta < \infty$ and some $N$ sufficiently large such that

$$\Pr(|y_n| > \Delta) < \delta,$$

for all $n > N$. □

The value $\delta$ provides a bound for the probability content of the tail to the right of $\Delta$ and the left of $-\Delta$ once we are far enough into the sequence. These conditions require that the tail behavior of the distributions of the sequence not be pathological. Specifically, the tail mass of the distributions cannot be drifting away from zero as we move out in the sequence. The leading case occurs with the central limit theorem discussed below which shows that the sample average, suitably normalized, has a limiting normal distribution, which satisfies the conditions for being bounded in probability.

**Definition 4.10.** The sequence of random variables $\{y_n\}$ is said to be of *order in probability at most* $n^\lambda$, and is denoted $O_p(n^\lambda)$, if $n^{-\lambda}y_n$ is bounded in probability. □
Example 4.4. Suppose \( z \sim N(0, 1) \) and \( y_n = 3 + n \cdot z \), then \( n^{-1}y_n = 3/n + z \) is a bounded random variable since the first term is asymptotically negligible and we see that \( y_n = O_p(n) \). □

Example 4.5. Suppose \( z_i \sim i.i.d. N(0, 1) \) for \( i = 1, 2, \ldots, n \) then \( y_n = \sum_{i=1}^{n} z_i^2 \sim \chi_n^2 \), which has \( \mathbb{E}[y_n] = n \) and \( \text{var}[y_n] = 2n \). The sequence \( y_n \) does not satisfy the definition of bounded in probability since the probability mass is being strung out to the right as \( n \to \infty \). However, by the univariate CLT \( \sqrt{n}(y_n/n - 1) \to_d N(0, 2) \), which is bounded in probability and hence \( O_p(1) \). Thus \( (y_n/n - 1) = O_p(1/\sqrt{n}) \), \( (y_n - n) = O_p(\sqrt{n}) \), and \( y_n = (y_n - n) + n = O_p(\sqrt{n}) + O(n) = O_p(n) \). □

Definition 4.11. The sequence of random variables \( \{ y_n \} \) is said to be of order in probability smaller than \( n^\lambda \), and is denoted \( o_p(n^\lambda) \), if \( n^{-\lambda}y_n \xrightarrow{p} 0 \). □

Example 4.6. Convergence in probability can be represented in terms of order in probability. Suppose that \( y_n \xrightarrow{p} c \) or equivalently \( y_n - c \xrightarrow{p} 0 \), then \( n^0(y_n - c) \xrightarrow{p} 0 \) and \( y_n - c = o_p(1) \). □

As with the order of sequences of real numbers, there is an algebra for the order in probability of sequences of random variables. Specifically, for the random sequences \( y_n = O_p(n^k) \) and \( z_n = O(n^\ell) \), then \( w_n = y_n \cdot z_n = O_p(n^k \cdot n^\ell) = O_p(n^{k+\ell}) \). Similarly, \( z_n = o_p(n^k) \) and \( w_n = o_p(n^\ell) \), then \( y_n = z_n \cdot w_n = o_p(n^k) \cdot o_p(n^\ell) = o_p(n^{k+\ell}) \). Finally, if \( y_n = O_p(n^k) \) and \( z_n = o_p(n^\ell) \), then \( w_n = z_n \cdot w_n = O_p(n^k) \cdot o_p(n^\ell) = o_p(n^{k+\ell}) \) and \( y_n = z_n + w_n = O_p(n^k) + o_p(n^\ell) = 1(l > k) \cdot o_p(n^\ell) + 1(l \leq k) \cdot O_p(n^k) \). Moreover, for mixtures of sequences of reals and randoms, a similar algebra holds except the results are always orders in probability since the mixtures will be random.

4.1.6 Convergence In Distribution

The previous sections have provided tools for describing the limiting point of the distributions of a random sequence and also the rate at which the sequence is tending toward a bounded random variable. Inevitably, we also want to know the specific distribution toward which suitably transformed random sequence is tending, if such is the case.

Definition 4.12. A sequence of random variables \( y_1, y_2, \ldots, y_n, \ldots \), with cumulative distribution functions \( F_1(\cdot), F_2(\cdot), \ldots, F_n(\cdot), \ldots \), is said to converge in distribution to a random variable \( y \) with a cumulative distribution function \( F(y) \), if

\[
\lim_{n \to \infty} F_n(\cdot) = F(\cdot),
\]

for every point of continuity of \( F(\cdot) \). The distribution \( F(\cdot) \) is said to be the limiting distribution of this sequence of random variables. □
4.1. CONVERGENCE OF RANDOM VARIABLES

For notational convenience, we often write $y_n \xrightarrow{d} F(\cdot)$ or $y_n \rightarrow_d F(\cdot)$ if a sequence of random variables converges in distribution to $F(\cdot)$. Note that the moments of elements of the sequence do not necessarily converge to the moments of the limiting distribution. There are well-known examples where the limiting distribution has moments of all orders but certain moments of the finite-sample distribution do not exist.

Note that the pointwise convergence in distribution above allows for the constant $\delta$ used in the limit to depend on the point of evaluation, say $c$, and (possibly) the values of the underlying parameters of the generating process, say $\theta$. It will be sometimes useful to utilize cases where the same constant can be utilized across the set of possible $c$ and (possibly) $\theta$. For simplicity, we ignore the possible dependence on $\theta$ since the definition easily generalizes.

**Definition 4.13.** A sequence of random variables $y_1, y_2, \ldots, y_n, \ldots$, with cumulative distribution functions $F_1(\cdot), F_2(\cdot), \ldots, F_n(\cdot), \ldots$, is said to converge uniformly in distribution to a random variable $y$ with a cumulative distribution function $F(y)$, if

$$
\sup_{c \in C} |F_n(c) - F(c)| \xrightarrow{p} 0,
$$

for $C$ being the set of feasible points $c$ of evaluation for the function $F(\cdot)$. □

Uniform convergence implies pointwise convergence but the reverse is not generally true. One important exception occurs when $F(c)$ is continuous over compact $C$, whereupon pointwise convergence of $F_n(c)$ to $F(c)$ implies uniform convergence.

### 4.1.7 Some Useful Propositions

In the sequel, we will sometimes need to work with combinations of two or more sequences or transformations of a sequence. Consequently, the following propositions will prove useful, where $x_n$ and $y_n$ denote two sequences of random vectors.

**Proposition 4.1.** If $x_n - y_n$ converges in probability to zero, and $y_n$ has a limiting distribution, then $x_n$ has a limiting distribution, which is the same. □

**Proposition 4.2.** If $y_n$ has a limiting distribution and plim $x_n = c$, then for $z_n = y_n' x_n$, plim $z_n = 0$. □

**Proposition 4.3.** Suppose that $y_n$ converges in distribution to a random variable $y$, and plim $x_n = c$. Then $x_n' y_n$ converges in distribution to $c'y$. □

**Proposition 4.4.** If $g(\cdot)$ is a continuous function, and if $x_n - y_n$ converges in probability to zero, then $g(x_n) - g(y_n)$ converges in probability to zero. □
Proposition 4.5. If $g(\cdot)$ is a continuous function, and if $x_n$ converges in probability to a constant $c$, then $z_n = g(x_n)$ converges in distribution to the constant $g(c)$. □

Proposition 4.6. If $g(\cdot)$ is a continuous function, and if $x_n$ converges in distribution to a random variable $x$, then $z_n = g(x_n)$ converges in distribution to a random variable $g(x)$. □

4.2 Estimation Theory

4.2.1 Estimators

Typically, our interest centers on certain properties of the data that can be characterized as parameters of the model used. An example is the marginal propensity to consume in a model of aggregate consumption. Our econometric analysis of this property usually begins with a point estimate of the parameter based on the available data. The formula or function that was used to obtain the point estimate is an estimator. The estimate is a realization drawn from the distribution that follows when the estimator is applied to the sample distribution of the underlying variables of the model. Thus as the sample size is increased, the estimator forms a sequence of random variables. The simple example given before was the sample mean $\bar{x}_n$ as the sample size $n$ is increased.

In the sequel, we will entertain various estimators that result from optimization of a function of the data with respect to parameters of interest. Such estimators comprise the class of extremum estimators. Examples are: least squares, which finds a minimum; maximum likelihood, which finds the maximum; and generalized method of moments (GMM), which finds a minimum. In each case, we obtain first order conditions for the objective function, in terms of the unknown parameters, and use the solution of these as estimators.

Without loss of generality, we will consider estimators that result from minimizing a function of the data. Denote this function by $\psi_n(\theta) = \psi(z_1, z_2, \ldots, z_n; \theta)$, where $z_1, z_2, \ldots, z_n$ are $n$ observations on an $m \times 1$ vector of random variables $z$ and $\theta$ is the $p \times 1$ vector of unknown parameters defined on a set $\Theta$. Formally, our estimator is

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \psi_n(\theta). \quad (4.4)$$

Typically, this estimator will be obtained by setting the first-order conditions (FOC), namely the partial derivatives of the function to be minimized, to zero

$$0 = \frac{\partial \psi_n(\theta_n)}{\partial \theta}$$

and solving for $\hat{\theta}_n$. In some cases such a solution is easy to obtain and in others it is not. It is possible for the FOC to be nonlinear in $\theta$ and the resulting solution to the FOC to be very complicated functions of the data.
In Appendix A to this chapter, we will develop results which establish the large-sample asymptotic properties of such estimators under very general conditions.

### 4.2.2 Properties Of Estimators

**Definition 4.14.** An estimator $\hat{\theta}_n$ of the $p \times 1$ parameter vector $\theta$ is a function of the sample observations $z_1, z_2, ..., z_n$. □

It follows that $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_n$ form a sequence of random variables.

**Definition 4.15.** The estimator $\hat{\theta}_n$ is said to be **unbiased** if $E \hat{\theta}_n = \theta$, for all $n$. □

**Definition 4.16.** The estimator $\hat{\theta}_n$ is said to be **asymptotically unbiased** if

$$\lim_{n \to \infty} E \hat{\theta}_n = \theta.$$

Note that an estimator can be biased in finite samples, but asymptotically unbiased.

**Definition 4.17.** The estimator $\hat{\theta}_n$ is said to be **consistent** if

$$\text{plim}_{n \to \infty} \hat{\theta}_n = \theta.$$ □

Asymptotic unbiasedness implies that a measure of central tendency of the estimator, the mean, is the target value so in some sense the distribution of the estimator is centered around the target in large samples. Consistency implies that the distribution of the estimator is collapsing about the target as the sample size increases. Consistency neither implies nor is implied by asymptotic unbiasedness, as demonstrated by the following examples.

**Example 4.7.** Let

$$\tilde{\theta}_n = \begin{cases} \theta, & \text{with probability } 1 - 1/n \\ \theta + nc, & \text{with probability } 1/n \end{cases}$$

We have $E \tilde{\theta}_n = \theta + c$, so $\tilde{\theta}_n$ is a biased estimator, and $\lim_{n \to \infty} E \tilde{\theta}_n = \theta + c$, so $\tilde{\theta}_n$ is asymptotically biased as well. However, $\lim_{n \to \infty} \Pr(|\tilde{\theta}_n - \theta| > \epsilon) = 0$, so $\tilde{\theta}_n$ is a consistent estimator. □

**Example 4.8.** Suppose $x_i \sim i.i.d. N(\mu, \sigma^2)$ for $i = 1, 2, ..., n, ...$ and let $\bar{x}_n = x_n$ be an estimator of $\mu$. Now $E[\bar{x}_n] = \mu$ so the estimator is unbiased but

$$\Pr(|\bar{x}_n - \mu| > 1.96\sigma) = .05$$

so the probability mass is not collapsing about the target point $\mu$ so the estimator is not consistent. □
4.2.3 Laws Of Large Numbers And Central Limit Theorems

Most of the large sample properties of the estimators considered in the sequel derive from the fact that the estimators involve sample averages and that the asymptotic behavior of averages is well studied. In addition to the central limit theorems presented in the previous two chapters we have the following two laws of large numbers:

**Theorem 4.1.** If \( x_1, x_2, \ldots, x_n \) is a simple random sample, that is, the \( x_i \)'s are i.i.d., and \( E x_i = \mu \) and \( \text{Var}(x_i) = \sigma^2 \), then by Chebyshev’s Inequality,

\[
\lim_{n \to \infty} \frac{x_n - \mu}{\sqrt{n}} = 0
\]

**Theorem 4.2.** Suppose that \( x_1, x_2, \ldots, x_n \) are i.i.d. random variables, such that for all \( i = 1, \ldots, n \), \( E x_i = \mu \), then,

\[
\lim_{n \to \infty} \frac{x_n - \mu}{\sqrt{n}} = 0
\]

Both of these results apply element-by-element to vectors of estimators. Note that Kharichev’s Theorem does not require second moments for the underlying random variable.

For sake of completeness we repeat the following scalar central limit theorem.

**Theorem 4.3.** Suppose that \( x_1, x_2, \ldots, x_n \) are i.i.d. random variables, such that for all \( i = 1, \ldots, n \), \( E x_i = \mu \) and \( \text{Var}(x_i) = \sigma^2 \), then

\[
\sqrt{n}(x_n - \mu) \xrightarrow{d} N(0, \sigma^2)
\]

This result is easily generalized to obtain the multivariate version given in Theorem 3.3.

**Theorem 4.4.** Suppose that \( m \times 1 \) random vectors \( x_1, x_2, \ldots, x_n \) are (i) jointly i.i.d., (ii) \( E x_i = \mu \), and (iii) \( \text{Cov}(x_i) = \Sigma \), then

\[
\sqrt{n}(x_n - \mu) \xrightarrow{d} N(0, \Sigma)
\]

4.2.4 Asymptotic Linearity

The importance of the average and its’ asymptotic regularity as given by the laws of large numbers and the central limit theorem arise because a large number of estimators and most of those considered in the sequel are asymptotically linear.

**Definition 4.18.** An estimator is said to be asymptotically linear if, for \( z_i \) i.i.d., there exists \( v(\cdot) \) such that

\[
\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} v(z_i; \theta) + o_p(1)
\]
Typically, the form of $v(z_i; \theta)$ arises from setting first-order conditions to zero that might arise from optimization of an objective function. Accordingly, we will have $E(v(z_i; \theta) = 0$, and $\text{Cov}(v(z_i; \theta)) = V$ at $\theta = \theta^0$, whereupon we can apply the central limit theorem to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, V).$$

In order to obtain asymptotic linearity we will typically have to impose additional assumptions on the data and $v(\cdot)$, as with the general consistency and asymptotic normality theorems given in the appendix to this chapter.

**Example 4.9.** Suppose $x_i \sim \text{i.i.d.}\ N(\mu, \sigma^2)$ for $i = 1, 2, ..., n$ and we obtain $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X}_n)^2$, and form the feasible statistic

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

$$= \frac{\bar{X}_n - \mu \sqrt{\sigma^2/n}}{\sqrt{\sigma^2/n}}$$

$$= \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \sqrt{\frac{\sigma^2}{\hat{\sigma}^2}}$$

$$= \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} + \left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}\right) \left(\sqrt{\frac{\sigma^2}{\hat{\sigma}^2}} - 1\right).$$

Now the first term converges in distribution to a standard normal and is hence $O_p(1)$. It can be shown that $\hat{\sigma}^2 \rightarrow_p \sigma^2$, so the second term is the product of $O_p(1)$ and $o_p(1)$ factors and hence $o_p(1)$. Thus,

$$T_n = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} + o_p(1)$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} + o_p(1)$$

and the statistic is seen to be asymptotically linear with the limiting behavior governed by the central limit theorem applied to the normalized average of $v(z_i; \theta) = (x_i - \mu)/\sigma$. □

### 4.2.5 CUAN And Efficiency

We often encounter more than one candidate estimator for a particular model. It behooves us to have some metric, which will motivate a choice between the alternatives. Such a comparison requires a little more smoothness of the estimator.

**Definition 4.19.** An estimator is said to be consistently uniformly asymptotically normal (CUAN) if it is consistent, and if $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in
distribution to $N(0, V)$, and if the convergence is uniform over some compact subset of the parameter space $\Theta$. □

Uniformity is introduced in order to rule out pathological cases where the limiting distribution changes as we move away from the limit point $\theta$. In such pathological cases, the preferred estimator can change as we move away from the pathological point.

Suppose that $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to $N(0, V)$. Let $\tilde{\theta}_n$ be an alternative estimator such that $\sqrt{n}(\tilde{\theta}_n - \theta)$ converges in distribution to $N(0, W)$.

**Definition 4.20.** If $\hat{\theta}_n$ is CUAN with asymptotic covariance $V$ and $\tilde{\theta}_n$ is CUAN with asymptotic covariance $W$, then $\hat{\theta}_n$ is *asymptotically efficient relative to $\tilde{\theta}_n$* if $W - V$ is a positive semidefinite matrix. □

Among other properties asymptotic relative efficiency implies that the diagonal elements of $V$ are no larger than those of $W$, so the asymptotic variances of $\hat{\theta}_{n,i}$ are no larger than those of $\tilde{\theta}_{n,i}$. And a similar result applies for the asymptotic variance of any linear combination.

Sometimes an estimator can be shown to be efficient relative to all alternative.

**Definition 4.21.** A CUAN estimator $\hat{\theta}_n$ is said to be *asymptotically efficient* if it is asymptotically efficient relative to any other CUAN estimator. □

In the next chapter we find that when the Maximum Likelihood Estimator (MLE) is CUAN then it is efficient within the class.

### 4.3 Asymptotic Inference

#### 4.3.1 Inference

Strictly speaking, inference is the process by which we infer properties of the underlying model and distribution from the behavior of the observable variables. Point estimation of parameters of interest is certainly a leading example of such inference. Beyond point estimation, we are interested in making probability statements concerning the properties of interest or choosing between alternative models. Such statements or choices are usually based on hypothesis tests, interval estimates, or probability values, all of which require the distribution of the estimator. In cases where the small sample distribution of the estimator is unavailable or formidably complex, we must rely on the large-sample limiting or asymptotic distribution of the estimator to provide answers.

#### 4.3.2 Normal Ratios

To fix ideas, we begin with inference in the simple mean model. Under the conditions of the central limit theorem, this model yields $\sqrt{n}(\bar{x}_n - \mu) \overset{d}{\to} N(0, \sigma^2)$,
4.3. ASYMPTOTIC INFERENCE

so

$$\frac{(\overline{x}_n - \mu)}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0, 1) \quad (4.5)$$

Suppose that $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2$. Then, by Proposition 4.3, we also have

$$\frac{(\overline{x}_n - \mu)}{\sqrt{\hat{\sigma}^2/n}} = \frac{\sqrt{\sigma^2/n}}{\sqrt{\hat{\sigma}^2/n}} \frac{(\overline{x}_n - \mu)}{\sqrt{\sigma^2/n}} \xrightarrow{d} N(0, 1)$$

since the term under the square root converges in probability to one and the remainder converges in distribution to $N(0, 1)$.

Most typically, such ratios will be used for inference in testing a hypothesis. Now, for $H_0 : \mu = \mu_0$, we have

$$\frac{(\overline{x}_n - \mu_0)}{\sqrt{\hat{\sigma}^2/n}} \xrightarrow{d} N(0, 1), \quad (4.6)$$

while under $H_1 : \mu = \mu_1 \neq \mu_0$, we find that

$$\frac{(\overline{x}_n - \mu_0)}{\sqrt{\sigma^2/n}} = \frac{\sqrt{n}((\overline{x}_n - \mu_1) + \sqrt{n}(\mu_1 - \mu_0))}{\sqrt{\sigma^2/n}} = N(0, 1) + O_p(\sqrt{n})$$

In conducting an hypothesis test, we choose the critical value $c$ so that the size $\alpha$, or probability our statistic exceeding $c$ in absolute value, is sufficiently small that we prefer the alternative hypothesis where such extreme events are more likely, particularly in large samples.

Such ratios are of interest in estimation and inference with regard to more general parameters. Suppose that $\hat{\theta}$ is an estimator of the parameter vector $\theta$ and

$$\sqrt{n}(\hat{\theta}_{p \times 1} - \theta) \xrightarrow{d} N(0, \mathbf{V}).$$

Then, if $\theta_j$ is the parameter of particular interest, we consider

$$\frac{(\hat{\theta}_j - \theta_j)}{\sqrt{\hat{v}_{jj}/n}} \xrightarrow{d} N(0, 1),$$

and duplicating the arguments for the sample mean

$$\frac{(\overline{\theta}_j - \theta_j)}{\sqrt{\hat{v}_{jj}/n}} \xrightarrow{d} N(0, 1), \quad (4.7)$$

where $\hat{\mathbf{V}}$ is a consistent estimator of $\mathbf{V}$. This ratio will have a similar behavior under a null and alternative hypothesis with regard to $\theta_j$. 
4.3.3 Asymptotic Chi-Square

The ratios just discussed are only appropriate for inference with respect to a single parameter. Often, some question to be answered using inference involves more than one parameter at a time. In a leading example, we have one model embedded in another as a special case then the more complex model reduces to the special case if a set of parameters take on a specific value, such as zero.

Suppose that
\[ \sqrt{n} (\hat{\theta} - \theta) \xrightarrow{d} N(0, V), \]
where \( \hat{\Psi} \) is a consistent estimator of the nonsingular \( p \times p \) matrix \( \Psi \). Then, from the previous chapter we have
\[ \sqrt{n} (\hat{\theta} - \theta) V^{-1} \sqrt{n} (\hat{\theta} - \theta) \xrightarrow{d} \chi^2_p \]
and
\[ n \cdot (\hat{\theta} - \theta) \hat{V}^{-1} (\hat{\theta} - \theta) \xrightarrow{d} \chi^2_p \]
for \( \hat{V} \) a consistent estimator of \( V \).

A similar result can also be obtained for any subvector of \( \theta \). Let
\[ \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \]
where \( \theta_2 \) is a \( p_2 \times 1 \) vector. Then
\[ \sqrt{n} (\hat{\theta}_2 - \theta_2) \xrightarrow{d} N(0, V_{22}), \]
where \( V_{22} \) is the lower right-hand \( p_2 \times p_2 \) submatrix of \( \Psi \) and,
\[ n \cdot (\hat{\theta}_2 - \theta_2) \hat{V}_{22}^{-1} (\hat{\theta}_2 - \theta_2) \xrightarrow{d} \chi^2_{p_2}, \]
where \( \hat{V}_{22} \) is the analogous lower right-hand submatrix.

This result can be used to conduct inference by testing the sub-vector. If \( H_0 : \theta_2 = \theta_2^0 \), then
\[ n \cdot (\hat{\theta}_2 - \theta_2^0) \hat{V}_{22}^{-1} (\hat{\theta}_2 - \theta_2^0) \xrightarrow{d} \chi^2_{p_2}, \]
and large positive values are rare events. while for \( H_1 : \theta_2 = \theta_2^1 \neq \theta_2^0 \), we can show (later)
\[ n \cdot (\hat{\theta}_2 - \theta_2^0) \hat{V}_{22}^{-1} (\hat{\theta}_2 - \theta_2^0) = n \cdot ((\hat{\theta}_2 - \theta_2^1) + (\theta_2^1 - \theta_2^0)) \hat{V}_{22}^{-1} ((\hat{\theta}_2 - \theta_2^1) + (\theta_2^1 - \theta_2^0)) \]
\[ = n \cdot (\hat{\theta}_2 - \theta_2^1) \hat{V}_{22}^{-1} (\hat{\theta}_2 - \theta_2^1) + 2n \cdot (\theta_2^1 - \theta_2^0) \hat{V}_{22}^{-1} (\hat{\theta}_2 - \theta_2^1) \]
\[ + n \cdot (\theta_2^1 - \theta_2^0) \hat{V}_{22}^{-1} (\theta_2^1 - \theta_2^0) = \chi^2_{p_2} + O_p(\sqrt{n}) + O_p(n) \]
Thus, if we obtain a large value of the statistic, we may take it as evidence that the null hypothesis is incorrect.

The ratio approach of the previous section is a special case of the subvector approach. If the hypothesis to be tested only concerns the last element of the parameter vector so $H_0: \theta_p = \theta^0_p$, then we can test using the ratio

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{v}_{jj}/n}} \xrightarrow{d} N(0,1),$$

or the quadratic form, which for this case is

$$n \cdot (\hat{\theta}_p - \theta^0_p)^\prime \hat{\Omega}^{-1} (\hat{\theta}_p - \theta^0_p) = n \cdot (\hat{\theta}_p - \theta^0_p)^2/\hat{v}_{pp} \xrightarrow{d} \chi^2_1.$$

But the second is just the square of the first and the distribution consulted is the chi-squared with 1 degree of freedom which results from a standard normal squared. Note that large negative and positive excursions of $\hat{\theta}_p$ from $\theta^0_p$ impact the quadratic statistic equally so it can only be used for two-sided tests.

### 4.3.4 Tests Of General Restrictions

We can use similar results to test general nonlinear restrictions. Suppose that $r(\cdot)$ is a $q \times 1$ continuously differentiable function, and

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V).$$

By the intermediate value theorem we can obtain the exact Taylor’s series representation

$$r(\hat{\theta}) = r(\theta) + \frac{\partial r(\theta^*)}{\partial \theta^*} (\hat{\theta} - \theta)$$

or equivalently

$$\sqrt{n}(r(\hat{\theta}) - r(\theta)) = \frac{\partial r(\theta^*)}{\partial \theta^*} \sqrt{n}(\hat{\theta} - \theta)$$

$$= R(\theta^*) \sqrt{n}(\hat{\theta} - \theta)$$

where $R(\theta^*) = \frac{\partial r(\theta^*)}{\partial \theta^*}$ and $\theta^*$ lies between $\hat{\theta}$ and $\theta$. Now $\hat{\theta} \rightarrow_p \theta$ so $\theta^* \rightarrow_p \theta$ and $R(\theta^*) \rightarrow_p R(\theta)$. Thus, we have

$$\sqrt{n}[r(\hat{\theta}) - r(\theta)] \xrightarrow{d} N(0, R(\theta)VR'(\theta)).$$

Consider a test of $H_0: r(\theta) = 0$. Assuming $R(\theta)VR'(\theta)$ is nonsingular, we have

$$n \cdot r(\hat{\theta})[R(\theta)VR'(\theta)]^{-1} r(\theta) \xrightarrow{d} \chi^2_q,$$

where $q$ is the length of $r(\cdot)$. In practice, we substitute the consistent estimates $\hat{R}(\hat{\theta})$ for $R(\theta)$ and $\hat{V}$ for $V$ to obtain, following the arguments given above

$$n \cdot r(\hat{\theta})[\hat{R}(\hat{\theta})\hat{V}R'(\theta)]^{-1} r(\theta) \xrightarrow{d} \chi^2_q,$$  \hspace{1cm} (4.11)

The behavior under the alternative hypothesis will be $O_p(n)$ as above.
4.A Appendix: Limit Theorems for Extremum Estimators

In this section, we will develop results which establish the consistency and asymptotic normality of extremum estimators as defined by (4.4) under very general conditions. These conditions are high-level, in the sense that they encompass many models and must be verified for the specific model of interest. The assumptions on the specific model that lead to verification of the high-level assumptions are low-level or primitive assumptions.

We first consider a result for consistency of the extremum estimator. For notational simplicity we will suppress the subscript \( n \) on the estimator \( \hat{\theta}_n \).

**Theorem 4.5.** Suppose (C1) \( \Theta \subset \mathbb{R}_p \) compact, (C2) \( \exists \psi_0(\theta) : \theta \rightarrow \mathbb{R} \) s.t. \( \psi_n(\theta) \rightarrow_p \psi_0(\theta) \) uniformly over \( \theta \in \Theta \), (C3) \( \psi_0(\theta) \) continuous over \( \Theta \), (C4) for \( \theta \in \Theta \), \( \psi_0(\theta) \) attains a unique minimum for \( \theta = \theta_0 \), then \( \hat{\theta} \rightarrow_p \theta_0 \) □

**Proof:** The following proof uses the properties that the inf and sup of a compact set must exist, that a closed subset of a compact set is also compact, and that the image set of a continuous function on a compact set is also compact. Consider any open neighborhood \( \Theta_0 \) of \( \theta_0 \), then the complement \( \Theta_1 = \Theta - \Theta_0 \) is a closed subset of \( \Theta_0 \) and hence compact by (C1). For \( \theta \in \Theta_1 \) we have \( \psi_0(\theta) - \psi_0(\theta_0) > 0 \), by (C4). So, by continuity of \( \psi_0(\theta) \) and compactness of \( \Theta_1 \), we find

\[
\delta^* = \inf_{\theta \in \Theta_1} (\psi_0(\theta) - \psi_0(\theta_0))
\]

exists and positive. Therefore, \( \exists \delta \) s.t. \( \delta^* > \delta > 0 \) and \( \psi_0(\theta) - \psi_0(\theta_0) > \delta \) for \( \theta \in \Theta_1 \). Next, for any \( \varepsilon > 0 \), we find, by (C2), that \( |\psi_n(\theta) - \psi_0(\theta)| \leq \varepsilon \) for \( \theta \in \Theta \) with \( \Pr \) approaching 1. Thus

\[
\psi_0(\hat{\theta}) - \psi_0(\theta_0) = [\psi_0(\hat{\theta}) - \psi_n(\hat{\theta})] + [\psi_n(\hat{\theta}) - \psi_0(\theta_0)]
\]

\[
= [\psi_0(\hat{\theta}) - \psi_n(\hat{\theta})] + [\psi_n(\hat{\theta}) - \psi_n(\theta_0)] + [\psi_n(\theta_0) - \psi_0(\theta_0)]
\]

\[
\leq [\psi_0(\hat{\theta}) - \psi_n(\hat{\theta})] + [\psi_n(\theta_0) - \psi_0(\theta_0)] \text{ since } [\psi_n(\hat{\theta}) - \psi_n(\theta)] \leq 0 \text{ by definition}
\]

\[
\leq 2\varepsilon,
\]

with \( \Pr \) approaching 1. But if we choose \( \varepsilon = \delta/2 \), then with \( \Pr \) approaching 1, \( \psi_0(\hat{\theta}) - \psi_0(\theta_0) \leq \delta \) and \( \hat{\theta} \notin \Theta_1 \), which means \( \hat{\theta} \in \Theta_0 \), for any neighborhood of \( \theta_0 \). □

The assumptions for this theorem are very high level and verification is sometimes easy and sometimes not. The compactness assumption applies directly to the parameter space and hence should not be difficult to verify. The continuity assumption will usually be met if the functions of the underlying model are themselves smooth but should be verified nonetheless. The global minimum assumption is not so restrictive as it might seem since we can always restrict \( \theta \) to a smaller space where there is only one minimum. In practice, the uniform convergence in probability assumption can become the most difficult to verify when the underlying model is complicated.
4.3. ASYMPTOTIC INFERENCE

With a little more smoothness, we can give a geometric proof of the consistency theorem for the single parameter case. Consider Figure 4.1 above. For simplicity, we consider the limiting function \( \psi_0(\theta) \) to be approximately quadratic (more will be said on this below). For an arbitrary choice of \( \delta > 0 \), we know by the uniform convergence assumption that the function we are minimizing \( \psi_n(\theta) \), represented here by the slightly wiggly line, will lie in the sleeve between \( \psi_0(\theta) - \delta \) and \( \psi_0(\theta) + \delta \) with probability approaching one as \( n \) grow large. Now, if \( \psi_n(\theta) \) lies in this region its minimum cannot exceed \( \psi_{\text{max}} \) which is the minimum of \( \psi_0(\theta) + \delta \) since \( \psi_n(\theta) \) lies below \( \psi_0(\theta) + \delta \). At the same time, this minimum must also lie above \( \psi_0(\theta) - \delta \), which means that the value of \( \theta \) that minimizes \( \psi_n(\theta) \), namely \( \hat{\theta} \), must lie between \( \theta_a \) and \( \theta_b \). For sufficiently large \( n \), we can make \( \delta \) arbitrarily small, whereupon the sleeve formed by \( \psi_0(\theta) - \delta \) and \( \psi_0(\theta) + \delta \) about \( \psi_0(\theta) \) grows tighter and by continuity of \( \psi_0(\theta) \), \( \theta_a \) and \( \theta_b \) must be closer and closer to \( \theta_0 \), with probability approaching one, which proves consistency.

If the function \( \psi_0(\theta) \) is twice differentiable then in the neighborhood of the minimum it will be approximately quadratic and this proof will apply. Since there is a unique global minimum by assumption, the global minimum will be in the neighborhood of \( \theta_0 \) and the proof applies.

We now present a normality result for the extremum estimator.

**Theorem 4.6.** Suppose \((N1)\) \( \hat{\theta} \stackrel{p}{\rightarrow} \theta_0 \), \( \theta_0 \) interior \( \Theta \), \((N2)\) \( \psi_n(\theta) \) twice
continuously differentiable in neighborhood $\mathcal{N}$ of $\theta_0$, (N3) $\exists \Psi(\theta)$, continuous at $\theta^*$, s.t. $\partial^2 \psi_n(\theta)/\partial \theta \partial \theta' \rightarrow_p \Psi(\theta)$ uniformly over $\theta \in \mathcal{N}$, (N4) $\Psi = \Psi(\theta^*)$ nonsingular, (N5) $\sqrt{n} \partial \psi_n(\theta^0)/\partial \theta \rightarrow_d N(0, V)$, then $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \Psi^{-1}V\Psi^{-1})$ □

**Proof:** By (N1) $\hat{\theta} \in \mathcal{N}$ with probability one, and we can use (N2) to expand the first-order conditions in a Taylor’s series and add and subtract terms to yield

$$0 = \frac{\partial \psi_n(\hat{\theta})}{\partial \theta} = \frac{\partial \psi_n(\theta^0)}{\partial \theta} + \frac{\partial^2 \psi_n(\theta^*)}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0)$$

$$= \frac{\partial \psi_n(\theta^0)}{\partial \theta} + \Psi(\theta_0)(\hat{\theta} - \theta_0) + [\Psi(\theta^*) - \Psi(\theta_0)](\hat{\theta} - \theta_0) + \left[\frac{\partial^2 \psi_n(\theta^*)}{\partial \theta \partial \theta'} - \Psi(\theta^*)\right](\hat{\theta} - \theta_0)$$

for $\theta^*$ between $\hat{\theta}$ and $\theta_0$. It follows that $\theta^* \rightarrow_p \theta_0$ so, by continuity and the Slutsky Theorem, $\Psi(\theta^*) - \Psi(\theta_0) = o_p(1)$, and by (N3) $\frac{\partial^2 \psi_n(\theta^*)}{\partial \theta \partial \theta'} - \Psi(\theta^*) = o_p(1)$. Thus

$$-\sqrt{n} \frac{\partial \psi_n(\theta^0)}{\partial \theta} = \Psi(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0) + o_p\left(\sqrt{n}(\hat{\theta} - \theta_0)\right)$$

which, using (N4), can be multiplied through by $\Psi(\theta_0)^{-1}$ and solving to obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\Psi(\theta_0)^{-1}\sqrt{n} \frac{\partial \psi_n(\theta^0)}{\partial \theta} + o_p\left(\sqrt{n}(\hat{\theta} - \theta_0)\right)$$

$$\rightarrow_d N(0, \Psi^{-1}V\Psi^{-1}) \text{ by (N5). □}$$

The assumptions of this theorem are fairly standard but should be verified for the model in question. The nonsingularity assumption will usually follow from the identification assumption introduced for the consistency theorem. In order to perform inference based on the estimator $\hat{\theta}$, we need a consistent estimator of the covariance matrix $\Psi^{-1}V\Psi^{-1}$. The uniform convergence result for the second derivatives yields $\hat{\Psi} = \partial^2 \psi_n(\hat{\theta})/\partial \theta \partial \theta'$ as an obvious consistent estimator of $\Psi$. An appropriate consistent estimator of $V$ will depend on the specifics of the model in question and may require additional assumptions.

### 5.B Appendix: Limit Theorems for Dependent Processes

In Section 4.2, we presented laws of large numbers (LLN) and central limit theorems (CLT) for i.i.d. processes. Economic outcomes frequently depend on forward-looking decision-making which involves optimization over a future time horizon that is predicted on the basis of past behaviour. To study such models
4.3. ASYMPTOTIC INFERENCE

involves analysis of data series that are repeated measures of the same process over time. The assumption of \textit{i.i.d.} is questionable or sometimes obviously inappropriate for many of these models. Consequently, we will present limit theorems in this section that are appropriate for what are called dependent processes. We will only look at processes that involve discrete time so the time index is informative. A sequence of such random variables \( x_1, x_2, ... \) is called a \textit{time-series} in the statistical literature and a \textit{dependent process} because the random variables are not necessarily stochastically independent.

Since we will relax the \textit{i.i.d.} assumption it will need to be replaced by some alternative assumptions. The secret of the limit theorems for \textit{i.i.d.} processes was that each additional observation brought more information to bear on the process, whose distribution was constant. In order to mimick these properties in a dependent process, we will need some notion of constancy and also additional information from each observation.

\textbf{Definition 4.22.} A process is \textit{strictly stationary} if for any fixed integer \( m \) the joint distribution of \((x_{n+1}, x_{n+2}, ..., x_{n+m})\) has the same distribution for all nonnegative integers \( n \).

Thus if we have repeated observations on \( m \) contiguous random variables, we have to notion of a constant target in the distribution.

\textbf{Definition 4.23.} A process having second moments is \textit{covariance stationary} if \( \mu = E[x_n] \) is the same for all positive integers \( n \) and for each nonnegative integer \( m \), \( \gamma_m = E[(x_n - \mu)(x_{n+m} - \mu)] \) is the same for all \( n \).

Note that strict stationarity implies covariance stationarity if the moments exist.

Now we need some notion that additional information about the distribution is obtained with additional time-series observations. Accordingly, we introduce the concept of ergodicity.

\textbf{Definition 4.24.} A stationary process is \textit{ergotic} if the sample moments over time converge in probability to the population moments at one point in time, which are always the same, due to stationarity.

This definition can also be stated as the following theorem:

\textbf{Theorem 4.7.} Suppose \( x_1, x_2, ... \) is strictly stationary and ergotic, then \( n^{-1} \sum_{t=1}^{n} a(x_t) \xrightarrow{a.s.} E[a(x_1)] \) for any measurable functions \( a(x) \) with \( E[|a(x_1)|] < \infty \).

There is a sense in which this theorem assumes away the problem. That is, ergodicity is equivalent to consistency, which is more or less true. Finding conditions for ergodicity to hold is the same as finding conditions for consistency to hold so it is often just assumed.

More insight can be gained by looking at a \textit{mean ergotic} process which is a stationary process with consistency of the first sample moment. We further restrict our attention to covariance stationary processes. Note that \( E[x_t] = \mu \) for all \( t \), so for \( \bar{x}_n = \frac{1}{n} \sum_{t=1}^{n} x_t \), we find

\[ E[\bar{x}_n] = \mu \]
CHAPTER 4. ASYMPTOTIC THEORY

and the sample mean is unbiased. If we can further show the variance converges
to zero then we will have convergence in quadratic mean and hence consistency.
Accordingly, consider the variance of the normalized and centered average
\( \sqrt{n}(\bar{x}_n - \mu) \), which has a more complicated variance than before, due to the
dependence,

\[
\sigma_n^2 = \text{var}[\sqrt{n}(\bar{x}_n - \mu)] = E \left[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (x_t - \mu) \right]
\]

\[
= \frac{1}{n} \sum_{t=1}^{n} \sum_{s=1}^{n} E[(x_t - \mu)(x_s - \mu)]
\]

\[
= \frac{1}{n} \left( \sum_{t=1}^{n} E[(x_t - \mu)^2] + \sum_{t=1}^{n} \sum_{s \neq t}^{n} E[(x_t - \mu)(x_s - \mu)] \right).
\]

From the definition above, we have \( \gamma_{t-s} = E[(x_t - \mu)(x_s - \mu)] \) for \( t > s \), so

\[
\sigma_n^2 = \frac{1}{n} \left[ n\gamma_0 + 2 \sum_{t=2}^{n} \sum_{s=1}^{t-1} \gamma_{t-s} \right]
\]

\[
= \frac{1}{n} \left[ n\gamma_0 + 2 \sum_{s=1}^{n-1} (n-s)\gamma_s \right]
\]

\[
= \gamma_0 + 2 \sum_{s=1}^{n-1} \frac{n-s}{n} \gamma_s
\]

\[
= \gamma_0 + 2 \sum_{s=1}^{n-1} (1 - \frac{s}{n})\gamma_s.
\]

Since \( 2 > 2(1 - s/n) > 0 \), we can apply the triangle inequality to find that

\[
\sigma_n^2 \leq \gamma_0 + 2 \sum_{s=1}^{n-1} \left| \frac{1 - s}{n} \right| \gamma_s
\]

\[
\leq \gamma_0 + 2 \sum_{s=1}^{n-1} \left| \gamma_s \right|.
\]

Thus, the absolute summability condition \( \sum_{s=1}^{n} \left| \gamma_s \right| < \infty \) assures that \( \sigma_n^2 < \infty \),
so \( \text{var}(\bar{x}_n) = \sigma_n^2/n \to 0 \), whereupon we have proved the following result:

**Theorem 4.8.** Suppose \( x_1, x_2, \ldots \) is covariance stationary and \( \sum_{s=1}^{n} \left| \gamma_s \right| < \infty \),
then \( \bar{x}_n = \frac{1}{n} \sum_{t=1}^{n} x_t \xrightarrow{p} \mu = E[x_1] \).

This condition means the covariances are declining much faster that \( s \), in some
sense. The memory of the process is declining, in the second moment sense,
4.3. ASYMPTOTIC INFERENCE

so adding observations adds information that is not correlated with previous realizations in the process. Existence of higher moments, say fourth, and strict stationarity enables us to apply this result to similarly consistently estimate second moments of the process such as \( \gamma_4 = E[(x_t - \mu)(x_{t-s} - \mu)] \) for finite \( s \).

Asymptotic normality requires a stronger notion of independence such as the sequence being mixing of some form. Mixing places explicit restrictions on the degree of dependence between two collections of random variables that are sufficiently far apart. The easiest form of mixing to understand is \( \rho \)- or rho-mixing which requires that the sequence \( \{x_t\}_{t=-\infty}^{\infty} \) be strictly stationary. Let \( A^n_m = g(x_1, x_2, ..., x_m) \) represent a family of measurable functions of the first \( m \) random variables in the stochastic process and \( A^\infty_m = h(x_{m+n}, x_{m+n+1}, ...) \) represent a family of measurable functions of the remainder of the variables starting at \( m+n \), where both families are square integrable random variables. Define \( \rho_n = \sup_{w \in A^n_m, v \in A^\infty_{m+n}} |\text{cor}(w, v)| \), then the sequence is \( \rho \)-mixing if \( \lim_{n \to \infty} \rho_n = 0 \). So the elements of the one collection are uniformly uncorrelated with the elements of the other collection when they are sufficiently far apart. Since \( g(\cdot) \) and \( h(\cdot) \) can be any measurable functions, the two collections are asymptotically independent.

Given this form of mixing we can now state a central limit theorem. Specifically, we have, from Ibragamov [1975],

**Theorem 4.9.** Suppose \( x_1, x_2, ... \) is strictly stationary, \( \rho \)-mixing, \( E[|x_t - \mu|^{2+\delta}] < \infty \) for some \( \delta > 0 \), and \( n \sigma_n^2 \to \infty \), then \( \sigma_n^2 \) converges with \( \sigma^2 = \lim_{n \to \infty} \sigma_n^2 \) and \( \sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \).

Thus replacing the \textit{i.i.d.} assumption with strict stationarity and \( \rho \)-mixing still requires a little more than existence of mean and variance. Specifically, we need a little more than second moments. This is a condition that also crops up in central limit theorems for independent but not identical processes.
Chapter 5

Maximum Likelihood Methods

5.1 Maximum Likelihood Estimation (MLE)

5.1.1 Motivation

Suppose we have a model for the random variable \( y_i \), for \( i = 1, 2, \ldots, n \), with unknown \((p \times 1)\) parameter vector \( \theta \). In many cases, the model will imply a distribution \( f(y_i|\theta) \) for each realization of the variable \( y_i \).

A basic premise of statistical inference is to avoid unlikely or rare models, for example, in hypothesis testing. If we have a realization of a statistic that exceeds the critical value then it is a rare event under the null hypothesis. Under the alternative hypothesis, however, such a realization is much more likely to occur and we reject the null in favor of the alternative. Thus in choosing between the null and alternative, we select the model that makes the realization of the statistic more likely to have occurred.

Carrying this idea over to estimation, we select values of \( \theta \) such that the corresponding values of \( f(y_i|\theta) \) are not unlikely. After all, we do not want a model that disagrees strongly with the data. Maximum likelihood estimation is merely a formalization of this notion that the model chosen should not be unlikely. Specifically, we choose the values of the parameters that make the realized data most likely to have occurred. This approach does, however, require that the model be specified in enough detail to imply a distribution for the variable of interest.
5.1. MAXIMUM LIKELIHOOD ESTIMATION (MLE)

5.1.2 The Likelihood Function

Suppose that the random variables \( y_1, y_2, \ldots, y_n \) are i.i.d. Then, the joint density function for \( n \) realizations is

\[
f(y_1, y_2, \ldots, y_n | \theta) = f(y_1 | \theta) \cdot f(y_2 | \theta) \cdot \ldots \cdot f(y_n | \theta)
= \prod_{i=1}^{n} f(y_i | \theta) \tag{5.1}
\]

Given values of the parameter vector \( \theta \), this function allows us to assign local probability measures for various values of the random variables \( y_1, y_2, \ldots, y_n \). This is the function which must be integrated to make probability statements concerning the joint outcomes of \( y_1, y_2, \ldots, y_n \).

Given a set of realized values of the random variables, we use this same function to establish the probability measure associated with various choices of the parameter vector \( \theta \).

**Definition 5.1.** The *likelihood function* of the parameters, for a particular sample of \( y_1, y_2, \ldots, y_n \), is the joint density function considered as a function of \( \theta \) given the \( y_i \)'s. That is,

\[
L(\theta | y) = \prod_{i=1}^{n} f(y_i | \theta) \quad \square \tag{5.2}
\]

5.1.3 Maximum Likelihood Estimation

For a particular choice of the parameter vector \( \theta \), the likelihood function gives a probability measure for the realizations that occurred. Consistent with the approach used in hypothesis testing, and using this function as the metric, we choose \( \theta \) that make the realizations most likely to have occurred.

**Definition 5.2.** The *maximum likelihood estimator* of \( \theta \) is the estimator obtained by maximizing \( L(\theta | y_1, y_2, \ldots, y_n) \) with respect to \( \theta \). That is,

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta | y), \tag{5.3}
\]

where \( \hat{\theta} \) is called the MLE of \( \theta \). \( \square \)

Equivalently, since \( \ln(\cdot) \) is a strictly monotonic transformation, we may find the MLE of \( \theta \) by solving

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta | y), \tag{5.4}
\]

where

\[
\mathcal{L}(\theta | y) = \ln L(\theta | y)
\]
denotes the log-likelihood function. In practice, we obtain \( \hat{\theta} \) by solving the first-order conditions (FOC)

\[
\frac{\partial \mathcal{L}(\theta; y)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i; \hat{\theta})}{\partial \theta} = 0.
\]
The motivation for using the log-likelihood function is apparent since the summation form will result, after division by \( n \), in estimators that are approximately averages, about which we know a lot. This advantage is particularly clear in the following example.

**Example 5.1.** Suppose that \( y_i \sim \text{i.i.d.} \mathcal{N}(\mu, \sigma^2) \), for \( i = 1, 2, \ldots, n \). Then,

\[
f(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} (y_i-\mu)^2},
\]

for \( i = 1, 2, \ldots, n \). Using the likelihood function (5.2), we have

\[
L(\mu, \sigma^2|y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} (y_i-\mu)^2}.
\] (5.5)

Next, we take the logarithm of (5.5), which gives us

\[
\ln L(\mu, \sigma^2|y) = \sum_{i=1}^{n} \left[ -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_i-\mu)^2 \right] = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i-\mu)^2
\] (5.6)

We then maximize (5.6) with respect to both \( \mu \) and \( \sigma^2 \). That is, we solve the following first order conditions:

(A) \( \frac{\partial \ln L(\cdot)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}) = 0; \)

(B) \( \frac{\partial \ln L(\cdot)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 0. \)

By solving (A), we find that \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}_n \). Solving (B) gives us \( \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 \).

Note that \( \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \neq s^2 \), where \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \). \( s^2 \) is the familiar unbiased estimator for \( \sigma^2 \), so \( \hat{\sigma}^2 \) is a biased estimator.

**Example 5.2.** Some notion of the attraction to maximum likelihood can be gained by a numerical example of the same problem just considered, where \( y_i \sim \text{i.i.d.} \mathcal{N}(\mu, \sigma^2) \). Choosing \( \mu = 10 \) and \( \sigma = 5 \) we randomly generate a sample of size \( n = 100 \). In the absence of an explicit choice of a distribution we are left with non-parametric approaches such as the histogram to assess probabilities. For the sample generated the resulting histogram given below is skewed to the right and does not appear normal. We then estimate the model by MLE and plot the true pdf and the estimated pdf with \( \hat{\mu} = 9.446 \) and \( s = 5.0816 \). Obviously, the maximum likelihood estimate of the distribution is much closer to the truth than the histogram. One could adopt more sophisticated nonparametric techniques that would result in a smoother estimated distribution than the histogram but they could not begin to approach the accuracy of the MLE approach when we have chosen the correct distribution to estimate.
5.1. MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Figure 5.1: Histogram and MLE of Normal Distribution
5.2 Asymptotic Behavior of MLE

5.2.1 Assumptions

For the asymptotic results we will derive in the following sections, we will make five assumptions. These assumptions are specific to maximum likelihood and are introduced for pedagogic reasons. They are shown to satisfy the higher-level assumptions given in the Appendix to the previous chapter.

(M1) The \( y_i \)'s are iid random variables with density function \( f(y_i|\theta) > 0 \) for \( i = 1, 2, \ldots, n \);

(M2) \( \ln f(y|\theta) \) and hence \( f(y|\theta) \) possess continuous derivatives with respect to \( \theta \) up to the third order for \( \theta \in \Theta \);

(M3) The range of \( y_i \) is independent of \( \theta \) and differentiation with respect to \( \theta \) to second order under the integral of \( f(y|\theta) \) is possible;

(M4) The parameter vector \( \theta \in \Theta \) is globally identified by the density function at the true parameter point \( \theta^0 \), which is interior to \( \Theta \), and \( \mathcal{I} = E[\partial^2 \ln f(y|\theta^0)/\partial \theta \partial \theta'] \) nonsingular.

(M5) \( \partial^3 \ln f(y_i|\theta)/\partial \theta \partial \theta_j \partial \theta_k \) is bounded in absolute value by some positive function \( H_{ijk}(y) \) for all \( y \) and \( \theta \in \Theta \), which, in turn, has a finite expectation for all \( \theta \in \Theta \).

The first assumption is fundamental and the basis of the estimator. If it is not satisfied then we are misspecifying the model and there is little hope for obtaining correct inferences, at least in finite samples. The second assumption is a regularity condition that is usually satisfied and easily verified. The first part of the third assumption is also easily verified and guaranteed to be satisfied in models where the dependent variable has smooth and infinite support. The fourth assumption must be verified, which is easier in some cases than others. The last assumption is crucial and bears a cost and really should be verified before MLE is undertaken but is usually ignored. Conditions for interchanging the order of differentiation and integration of \( f(y|\theta) \) to satisfy the second part of the third assumption require boundedness of the integrated norms of \( \partial f(y|\theta)/\partial \theta \) and \( \partial^2 f(y|\theta)/\partial \theta \partial \theta' \) and are discussed in the Appendix to this chapter.

Example 5.3. It is instructive to examine these assumptions with respect to the simple example just given where we estimated the mean \( \mu \) and variance \( \sigma^2 \) for \( y_i \sim i.i.d.N(\mu, \sigma^2) \). The first two assumptions are obviously satisfied and the third is met since the random variable has infinite support. Since the normal distribution is completely characterized by its mean and variance knowing the latter means knowing the former and the fourth assumption is satisfied.
5.2. ASYMPTOTIC BEHAVIOR OF MLE

The four unique third derivatives are:

\[
\begin{align*}
\frac{\partial^3 \ln f(y_i|\theta)}{\partial \mu^3} &= 0 \\
\frac{\partial^3 \ln f(y_i|\theta)}{\partial (\sigma^2)^3} &= -\frac{1}{(\sigma^2)^3} + \frac{3}{(\sigma^2)^4} (y_i - \mu)^2 \\
&= -\frac{1}{(\sigma^2)^3} + \frac{3}{(\sigma^2)^4} ((y_i - \mu_0)^2 + 2(\mu_0 - \mu)(y_i - \mu_0) + (\mu_0 - \mu)^2) \\
\frac{\partial^3 \ln f(y_i|\theta)}{\partial \mu^2 \partial (\sigma^2)} &= \frac{1}{(\sigma^2)^2} (y_i - \mu) \\
&= \frac{1}{(\sigma^2)^3} (\mu_0 - \mu) + \frac{1}{(\sigma^2)^4} (y_i - \mu_0) \\
\frac{\partial^3 \ln f(y_i|\theta)}{\partial \mu \partial (\sigma^2)^2} &= \frac{1}{(\sigma^2)^2}.
\end{align*}
\]

For \(\mu\) and \(\sigma^2\) bounded (above and below) these are all bounded in absolute value by a polynomial in \(|y_i - \mu_0|\), each of which has a finite expectation, so Assumption (M5) is also met. From this we see that Assumption (M5) may implicitly place restrictions on the parameter space \(\Theta\) such as the compactness assumption introduced in the Appendix. □

5.2.2 Some Preliminaries

In the following results we exploit Assumptions (M2) and (M3) to take differentiation inside the integral. Let \(\mathcal{N}\) be a neighborhood about the true parameter value \(\theta^0\). Now, we know that, for any value of the parameter vector \(\theta \in \mathcal{N}\),

\[
\int f(y|\theta)dy = 1 \tag{5.7}
\]

and hence

\[
0 = \frac{\partial}{\partial \theta} \int f(y|\theta)dy
\]

where integration is understood to be the multiple integral and the limits of integration are ±∞. Interchanging the order of differentiation and integration and multiplying and dividing by the nonzero density yields

\[
\begin{align*}
0 &= \int \frac{\partial f(y|\theta)}{\partial \theta} dy \\
&= \int \frac{\partial f(y|\theta)}{f(y|\theta)} f(y|\theta) dy \\
&= \int \frac{\partial \ln f(y|\theta)}{\partial \theta} f(y|\theta) dy, \tag{5.8}
\end{align*}
\]

for any \(\theta \in \mathcal{N}\) and

\[
0 = E \left[ \frac{\partial \ln f(y|\theta^0)}{\partial \theta} \right]
\]
for any value of the true parameter vector \( \theta^0 \in \mathcal{N} \), where \( \mathbb{E}[] \) indicates integration of the expression with respect to the true density \( f(y|\theta^0) \). The first derivative of the log-likelihood function is known as the score function and the result given as the mean zero property of scores at the true values of the parameters.

Differentiating (5.8) again and bringing the differentiation inside the integral yields

\[
0 = \int \left[ \frac{\partial^2 \ln f(y|\theta)}{\partial \theta \partial \theta'} f(y|\theta) + \frac{\partial \ln f(y|\theta)}{\partial \theta} \frac{\partial f(y|\theta)}{\partial \theta'} \right] dy. \tag{5.9}
\]

Since

\[
\frac{\partial f(y|\theta)}{\partial \theta'} = \frac{\partial \ln f(y|\theta)}{\partial \theta'} f(y|\theta), \tag{5.10}
\]

then, for \( \theta = \theta^0 \), we can rewrite (5.9) as

\[
0 = \mathbb{E}\left[ \frac{\partial^2 \ln f(y|\theta^0)}{\partial \theta \partial \theta'} \right] + \mathbb{E}\left[ \frac{\partial \ln f(y|\theta^0)}{\partial \theta} \frac{\partial \ln f(y|\theta^0)}{\partial \theta'} \right] \tag{5.11}
\]

and define

\[
\psi(\theta^0) = \mathbb{E}\left[ \frac{\partial \ln f(y|\theta^0)}{\partial \theta} \frac{\partial \ln f(y|\theta^0)}{\partial \theta'} \right] = -\mathbb{E}\left[ \frac{\partial \ln f(y|\theta^0)}{\partial \theta} \right] \tag{5.12}
\]

for any value of the true parameter vector \( \theta^0 \in \mathcal{N} \). The matrix \( \psi(\theta^0) \) is called the information matrix and the relationship given in (5.12) the information matrix equality.

Finally, we note, under Assumption (M1) that

\[
\mathbb{E}\left[ \frac{\partial^2 \ln L(\theta^0|y)}{\partial \theta \partial \theta'} \right] = \mathbb{E}\left[ \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta \partial \theta'} \right] = \sum_{i=1}^{n} \mathbb{E}\left[ \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta \partial \theta'} \right] = -n \psi(\theta^0), \tag{5.13}
\]

since the distributions are identical, and

\[
\mathbb{E}\left[ \frac{\partial \ln L(\theta^0|y)}{\partial \theta} \frac{\partial \ln L(\theta^0|y)}{\partial \theta'} \right] = \mathbb{E}\left[ \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta'} \right] = \sum_{i=1}^{n} \mathbb{E}\left[ \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta'} \right] = n \psi(\theta^0), \tag{5.14}
\]

since the covariances between different observations is zero. These last results will suggest two alternative estimators for \( \psi(\theta^0) \).
5.2. ASYMPTOTIC BEHAVIOR OF MLE

5.2.3 Asymptotic Properties

5.2.3.1 Consistent Root Exists

We will first establish consistency of the maximum likelihood estimator under Assumptions (M1)-(M5) given above. These assumptions and the proof below appear different from the general approach given in the Appendix of Chapter 4. However, as will be shown in the Appendix to this chapter these assumptions, suitably modified, satisfy those for the more general approach. Moreover, the approach taken here is particularly instructive in revealing the role of nonlinearity in estimation.

For notational simplicity, consider the case where $p = 1$, so $\theta$ is scalar. Then, dividing by $n$, Assumption (M2) allows us to expand in a Taylor’s series, and use the mean value theorem on the quadratic term to yield

$$\frac{1}{n} \frac{\partial \ln L(\hat{\theta}|y)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2} (\hat{\theta} - \theta^0)$$

$$+ \frac{1}{2} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3 \ln f(y_i|\theta^*)}{\partial \theta^3} (\hat{\theta} - \theta^0)^2$$

(5.15)

where $\theta^*$ lies between $\hat{\theta}$ and $\theta^0$. Since, by Assumption (M5), every term in the left summation that follows is smaller in absolute value than the corresponding term in the right summation, we have

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3 \ln f(y_i|\theta^*)}{\partial \theta^3} = k \frac{1}{n} \sum_{i=1}^{n} H(y_i),$$

(5.16)

for some $|k| \leq 1$. So,

$$\frac{1}{n} \frac{\partial \ln L(\theta|y)}{\partial \theta} = a\delta^2 + b\delta + c,$$

(5.17)

where

$$\delta = \hat{\theta} - \theta^0,$$

$$a = k \frac{1}{2} \frac{1}{n} \sum_{i=1}^{n} H(y_i),$$

$$b = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2},$$

and

$$c = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta}.$$
We can apply a law of large numbers to the averages in (5.18) to obtain \( \text{plim} \left( \frac{1}{n} \sum_{i=1}^{n} H(y_i) \right) = \frac{1}{2} E[H(y_0)] + O_p(1) \) under Assumption (M5).

Now, since \( \partial \ln L(\hat{\theta}|y)/\partial \theta = 0 \), we have \( a \delta^2 + b \delta + c = 0 \). There are two possibilities. If \( a = 0 \), then the FOC are linear in \( \delta \) whereupon \( \delta = -\frac{c}{b} \) and \( \delta \xrightarrow{p} 0 \). If \( a \neq 0 \) with probability 1, which will occur when the FOC are nonlinear in \( \delta \), then

\[
\delta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.
\] (5.19)

Since \( ac = o_p(1) \), then \( \delta \xrightarrow{p} 0 \) for the plus root while \( \delta \xrightarrow{p} \vartheta(\theta_0)/\alpha \) for the negative root if \( \text{plim}_{n \to \infty} a = \alpha \neq 0 \) exists. If the F.O.C. are nonlinear but asymptotically linear then \( a \xrightarrow{p} 0 \) so \( ac \) in the numerator of (5.19) will go to zero faster than \( a \) in the denominator and again \( \delta \xrightarrow{p} 0 \) for the plus root but will be unbounded for the negative root, which is ruled out if we consider a compact parameter space. Thus there exits at least one consistent solution \( \hat{\theta} \) which satisfies

\[
\text{plim}_{n \to \infty} (\hat{\theta} - \theta_0) = 0.
\] (5.20)

and in the asymptotically nonlinear case there is also a possibly inconsistent solution.

For the case of \( \theta \) a vector, we can apply a similar style proof to show there exists a solution \( \hat{\theta} \) to the FOC that satisfies \( \text{plim}_{n \to \infty} (\hat{\theta} - \theta_0) = 0 \). And, in the event of asymptotically nonlinear FOC, at least one other inconsistent root is possible.

5.2.3.2 Consistent Root Is Global Maximum

In the event of multiple roots, we are left with the problem of selecting between them. By Assumption (M4), the parameter \( \theta \) is globally identified by the density function. Formally, this means that

\[
f(y|\theta) = f(y,\theta^0),
\] (5.21)

for all \( y \) implies that \( \theta = \theta^0 \). Now,

\[
E \left[ \frac{f(y, \theta)}{f(y, \theta^0)} \right] = \int_{-\infty}^{+\infty} \frac{f(y, \theta)}{f(y, \theta^0)} f(y, \theta^0) = 1.
\] (5.22)

Thus, by Jensen’s Inequality, we have

\[
E \left[ \ln \frac{f(y, \theta)}{f(y, \theta^0)} \right] < \ln E \left[ \frac{f(y, \theta)}{f(y, \theta^0)} \right] = 0,
\] (5.23)
5.2. ASYMPTOTIC BEHAVIOR OF MLE

unless \( f(y, \theta) = f(y, \theta^0) \) for all \( y \), or \( \theta = \theta^0 \). Therefore, \( E[\ln f(y, \theta)] \) achieves a maximum if and only if \( \theta = \theta^0 \). This well-known result is called the information inequality.

However, we are seeking to maximize

\[
\frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\theta) \longrightarrow E[\ln f(y|\theta)],
\]

(5.24)

where the convergence in probability is uniform for \( \theta \in \Theta \). If we choose an inconsistent root, the objective function will converge in large samples to a smaller value. The consistent root will converge to a larger value. Accordingly, we choose the root with the larger value of the objective function, which will be appropriate asymptotically. This choice of the global root has added appeal since it is, in fact, the MLE among the possible alternatives and hence the choice that makes the realized data most likely to have occurred.

There are complications in finite samples since the value of the likelihood function for alternative roots may cross over as the sample size increases. That is, the global maximum in small samples may not be the global maximum in larger samples. An added problem is to identify all the alternative roots so we can choose the global maximum. Sometimes a solution is available in a simple consistent estimator which may be used to start the nonlinear MLE optimization.

5.2.3.3 Asymptotic Normality

We now establish the asymptotic normality of the maximum likelihood estimator. We continue to focus on the scalar case \( (p = 1) \). Recall that \( a\delta^2 + b\delta + c = 0 \), so

\[
\delta = \hat{\theta} - \theta^0 = \frac{-c}{a\delta + b}
\]

(5.25)

and

\[
\sqrt{n}(\hat{\theta} - \theta^0) = \frac{-\sqrt{n}c}{a(\hat{\theta} - \theta^0) + b}
\]

(5.26)

\[
= \frac{-1}{a(\hat{\theta} - \theta^0) + b} \sqrt{n}c.
\]

Now since \( a = O_p(1) \) and \( \hat{\theta} - \theta^0 = o_p(1) \), then \( a(\hat{\theta} - \theta^0) = o_p(1) \) and

\[
a(\hat{\theta} - \theta^0) + b \overset{p}{\to} -\vartheta(\theta^0).
\]

(5.27)

And by the CLT we have

\[
\sqrt{n}c = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} \overset{d}{\to} N(0, \vartheta(\theta^0)).
\]

(5.28)
Substituting these two results in (5.26), we find
\[ \sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, \vartheta(\theta^0)^{-1}). \]

In general, for \( p > 1 \), we can apply the same scalar proof to show \( \sqrt{n}(\lambda^t \hat{\theta} - \lambda^t \theta^0) \xrightarrow{d} N(0, \lambda^t \vartheta(\theta^0) \lambda) \) for any vector \( \lambda \), which means
\[ \sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{d} N(0, \vartheta^{-1}(\theta^0)), \]
(5.29)
if \( \hat{\theta} \) is the consistent root.

5.2.3.4 Cramér-Rao Lower Bound

In addition to being the covariance matrix of the MLE, \( \vartheta^{-1}(\theta^0) \) defines a lower bound for covariance matrices with certain desirable properties. Let \( \tilde{\theta}_n(y) \) be any unbiased estimator, then
\[ E\tilde{\theta}_n(y) = \int_{-\infty}^{+\infty} \tilde{\theta}_n(y) f(y; \theta^0) dy = \theta^0, \]
(5.30)
for any underlying \( \theta^0 \). Note that the expectation is with respect to the joint density and involves integrating with respect to all the \( y_i \)'s. Differentiating both sides of this relationship with respect to \( \theta \) yields
\[ I_p = \frac{\partial E\tilde{\theta}_n(y)}{\partial \theta'} = \int_{-\infty}^{+\infty} \tilde{\theta}_n(y) \frac{\partial f(y; \theta^0)}{\partial \theta} dy 
= \int_{-\infty}^{+\infty} \tilde{\theta}_n(y) \frac{\partial \ln f(y; \theta^0)}{\partial \theta^t} f(y; \theta^0) dy 
= E \left[ \tilde{\theta}_n(y) \frac{\partial \ln f(y; \theta^0)}{\partial \theta^t} \right] 
= E \left[ (\tilde{\theta}_n(y) - \theta^0) \frac{\partial \ln f(y; \theta^0)}{\partial \theta^t} \right]. \]
(5.31)

Next, we define
\[ C(\theta^0) = E \left[ (\tilde{\theta}_n(y) - \theta^0)(\tilde{\theta}_n(y) - \theta^0) \right] \]
(5.32)
as the covariance matrix of \( \tilde{\theta}_n(y) \) and note from (5.14) that \( E \left[ \frac{\partial \ln L}{\partial \theta} \frac{\partial \ln L}{\partial \theta} \right] = n(\vartheta(\theta^0)) \), whereupon,
\[ \text{Cov} \left( \frac{\partial \ln L}{\partial \theta} \right) = \begin{pmatrix} C(\theta^0) & I_p \\ I_p & n(\theta^0) \end{pmatrix} \]
(5.33)
which, as a covariance matrix, is positive semidefinite.

Now, for any \((p \times 1)\) vector \(a\), we have

\[
\left( a' - a' \frac{1}{n} \vartheta_1(\theta^0)\right) \left( \begin{array}{cc} C(\theta^0) & I_p \\ I_p & n \vartheta(\theta^0)\end{array} \right) \left( -a' \frac{1}{n} \vartheta_1(\theta^0)\right) = a' \left[C(\theta^0) - \frac{1}{n} \vartheta(\theta^0)^{-1}\right] a \geq 0.
\]

Thus, any unbiased estimator \(\tilde{\theta}_n(y)\) has a covariance matrix that exceeds \(\frac{1}{n} \vartheta(\theta^0)^{-1}\) by a positive semidefinite matrix. So if the MLE estimator is unbiased and has \(\frac{1}{n} \vartheta(\theta^0)^{-1}\) as its covariance matrix, then it is efficient within the class of unbiased estimators.

It is entirely possible that the MLE will be biased in small samples but will still have optimality properties in large samples. As we have shown, under the Assumptions (M1)-(M5), \(\sqrt{n} (\hat{\theta} - \theta^0) \rightarrow_d N(0, \vartheta(\theta^0)^{-1})\) and is, moreover, CUAN. In much the same fashion as above for the class of unbiased estimators, we can demonstrate that, in large samples, any other CUAN estimator will have limiting distribution with a covariance exceeding \(\vartheta(\theta^0)^{-1}\). Thus MLE is asymptotically efficient and \(\vartheta(\theta^0)^{-1}\) is called the Camér-Rao lower bound, within this class.

Example 5.4. Consider again the maximum likelihood problem posed in Example 5.1, where \(y_i \sim i.i.d. N(\mu, \sigma^2)\), for \(i = 1, 2, \ldots, n\). Then,

\[
\ln f(y|\mu, \sigma^2) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y - \mu)^2,
\]

with first derivatives

\[
\frac{\partial \ln f(\cdot)}{\partial \mu} = \frac{1}{\sigma^2} (y - \mu)
\]
\[
\frac{\partial \ln f(\cdot)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} (y - \mu)^2.
\]

For \(\theta' = (\mu, \sigma^2)\), then

\[
\vartheta(\theta^0) = E \left[ \frac{\partial \ln f(y|\theta^0)}{\partial \theta} \frac{\partial \ln f(y|\theta^0)}{\partial \theta'} \right] = \begin{pmatrix} 1/\sigma_0^2 & 0 \\ 0 & 1/(2\sigma_0^4) \end{pmatrix}
\]

and

\[
\vartheta(\theta^0)^{-1} = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix}.
\]

Thus \(\sigma_0^2/n\) is the lower bound for unbiased estimation of the mean, and is obtained by the sample average \(\overline{y}_n\). The estimator \(\hat{\sigma}^2\) is biased but asymptotically unbiased with \(\sqrt{n} (\hat{\sigma}^2 - \sigma_0^2) \rightarrow_d N(0, 2\sigma_0^4)\) and hence attains the bound asymptotically. \(\Box\)
CHAPTER 5. MAXIMUM LIKELIHOOD METHODS

5.3 Maximum Likelihood Inference

5.3.1 Likelihood Ratio Test

Suppose we wish to test $H_0 : \theta = \theta^0$ against $H_1 : \theta \neq \theta^0$. Then, we define

$$L_u = \max_{\theta} L(\theta | y) = L(\hat{\theta} | y)$$

and

$$L_r = L(\theta^0 | y),$$

where $L_u$ is the unrestricted likelihood and $L_r$ is the restricted likelihood. We then form the likelihood ratio

$$\lambda = \frac{L_r}{L_u}.$$  \hspace{1cm} (5.36)

Note that the restricted likelihood can be no larger than the unrestricted which maximizes the function.

As with estimation, it is more convenient to work with the logs of the likelihood functions. It will be shown below that, under $H_0$,

$$LR = -2 \ln \lambda = -2 \left[ \ln \frac{L_r}{L_u} \right]$$

$$= 2[\mathcal{L}(\hat{\theta} | y) - \mathcal{L}(\theta^0 | y)] \xrightarrow{d} \chi^2_p,$$  \hspace{1cm} (5.37)

where $\hat{\theta}$ is the unrestricted MLE, and $\theta^0$ is the restricted MLE. If $H_1$ applies, then $LR = O_p(n)$. Large values of this statistic indicate that the restrictions make the observed values much less likely than the unrestricted and we prefer the unrestricted and reject the restrictions.

In general, for $H_0 : r(\theta) = 0$, and $H_1 : r(\theta) \neq 0$, we have

$$L_u = \max_{\theta} L(\theta | y) = L(\hat{\theta} | y),$$

and

$$L_r = \max_{\theta} L(\theta | y) \text{ s.t. } r(\theta) = 0$$

$$= L(\bar{\theta} | y).$$

where $\bar{\theta}$ is the restricted maximum likelihood estimator. Under $H_0$,

$$LR = 2[\mathcal{L}(\bar{\theta} | y) - \mathcal{L}(\bar{\theta} | y)] \xrightarrow{d} \chi^2_q,$$  \hspace{1cm} (5.40)

where $q$ is the length of $r(\cdot)$.

Note that, in the general case, the likelihood ratio test requires calculation of both the restricted and the unrestricted MLE.
5.3. MAXIMUM LIKELIHOOD INFERENCE

5.3.2 Wald Test

The asymptotic normality of MLE may be used to obtain a test based only on the unrestricted estimates.

Now, under \( H_0 : \theta = \theta^0 \), we have

\[
\sqrt{n} (\hat{\theta} - \theta^0) \xrightarrow{d} N(0, \theta^{-1}(\theta^0)).
\] (5.41)

Thus, using the results on the asymptotic behavior of quadratic forms from the previous chapter, we have

\[
W = n (\hat{\theta} - \theta^0)' \vartheta(\theta^0)(\hat{\theta} - \theta^0) \xrightarrow{d} \chi^2_p,
\] (5.42)

which is the Wald test. As we discussed for quadratic tests, in general, under \( H_1 : \theta^1 \neq \theta^0 \), we would have \( W = O_p(n) \).

The statistic \( W \), is generally infeasible since it requires knowledge of \( \vartheta(\theta^0) \).

In practice, since under Assumptions (M2) and (M5), we can show

\[
\frac{1}{n} \frac{\partial^2 L(\hat{\theta}|y)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} \frac{1}{n} \frac{\partial^2 \ln f(\hat{\theta}|y)}{\partial \theta \partial \theta'} p - \theta(\theta^0),
\] (5.43)

then we instead use the feasible statistic

\[
W^* = -n (\hat{\theta} - \theta^0)' \frac{1}{n} \frac{\partial^2 L(\hat{\theta}|y)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta^0)
\] (5.44)

\[
= - (\hat{\theta} - \theta^0)' \frac{\partial^2 L(\hat{\theta}|y)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta^0) \xrightarrow{d} \chi^2_p.
\]

Aside from having the same asymptotic distribution, the Likelihood Ratio and Wald tests are asymptotically equivalent in the sense that

\[
\lim_{n \to \infty} (LR - W^*) = 0.
\] (5.45)

This is shown by expanding \( L(\theta^0|y) \) in a Taylor’s series about \( \hat{\theta} \). That is,

\[
L(\theta^0) = L(\hat{\theta}) + \frac{\partial L(\hat{\theta})}{\partial \theta} (\theta^0 - \hat{\theta})
\]

\[
+ \frac{1}{2} (\theta^0 - \hat{\theta})' \frac{\partial^2 L(\hat{\theta})}{\partial \theta \partial \theta'} (\theta^0 - \hat{\theta})
\]

\[
+ \frac{1}{6} \sum_i \sum_j \sum_k \left( \frac{\partial^3 L(\theta^*)}{\partial \theta_i \partial \theta_j \partial \theta_k} \right) (\theta^0_i - \hat{\theta}_i)(\theta^0_j - \hat{\theta}_j)(\theta^0_k - \hat{\theta}_k).
\] (5.46)

where the third line applies the intermediate value theorem for \( \theta^* \) between \( \hat{\theta} \) and \( \theta^0 \). Now \( \frac{\partial L(\hat{\theta})}{\partial \theta} = 0 \), and the third line can be shown to be \( O_p(1/\sqrt{n}) \) under Assumptions (M2) and (M5), whereupon we have

\[
L(\hat{\theta}) - L(\theta^0) = -\frac{1}{2} (\theta^0 - \hat{\theta})' \frac{\partial^2 L(\hat{\theta})}{\partial \theta \partial \theta'} (\theta^0 - \hat{\theta}) + O_p(1/\sqrt{n})
\] (5.47)
and
\[ LR = W^* + O_p(1/\sqrt{n}). \] (5.48)
This means that, in large samples, the two tests will reject and accept together under the null hypothesis.

In general, we may test \( H_0 : r(\theta) = 0 \) with
\[ W^* = -r(\hat{\theta})' \left[ R(\hat{\theta}) \left( \frac{\partial^2 L(\hat{\theta})}{\partial \theta \partial \theta'} \right)^{-1} R'(\hat{\theta}) \right]^{-1} r(\hat{\theta}) \xrightarrow{d} \chi^2_p. \] (5.49)

where \( R(\theta) = \frac{\partial r(\theta)}{\partial \theta} \).

5.3.3 Lagrange Multiplier
Alternatively, but in the same fashion, we can expand \( L(\hat{\theta}) \) about \( \theta^0 \) to obtain
\[ L(\hat{\theta}) = L(\theta^0) + \frac{1}{n} \frac{\partial L(\theta^0)}{\partial \theta} (\hat{\theta} - \theta^0) + \frac{1}{2} (\hat{\theta} - \theta^0)' \frac{\partial^2 L(\theta^0)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta^0) + O_p(1/\sqrt{n}). \] (5.50)
Likewise, we can also expand \( \frac{1}{n} \frac{\partial L(\hat{\theta})}{\partial \theta} \) about \( \theta^0 \), which yields
\[ 0 = \frac{1}{n} \frac{\partial L(\hat{\theta})}{\partial \theta} = \frac{1}{n} \frac{\partial L(\theta^0)}{\partial \theta} + \frac{1}{n} \frac{\partial^2 L(\theta^0)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta^0) + O_p(1/n), \] (5.51)
or
\[ (\hat{\theta} - \theta^0) = - \left( \frac{\partial^2 L(\theta^0)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial L(\theta^0)}{\partial \theta} + O_p(1/n). \] (5.52)
Substituting (5.52) into (5.50) gives us
\[ L(\hat{\theta}) - L(\theta^0) = - \frac{1}{2} \frac{\partial L(\theta^0)}{\partial \theta'} \left( \frac{\partial^2 L(\theta^0)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial L(\theta^0)}{\partial \theta} + O_p(1/\sqrt{n}), \] (5.53)
and \( LR = LM + O_p(1/\sqrt{n}) \), where
\[ LM = - \frac{\partial L(\theta^0)}{\partial \theta'} \left( \frac{\partial^2 L(\theta^0)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial L(\theta^0)}{\partial \theta}, \] (5.54)
is the Lagrange Multiplier test.
Thus, under \( H_0 : \theta = \theta^0 \),
\[ \lim_{n \to \infty} (LR - LM) = 0. \] (5.55)
and
\[ LM \xrightarrow{d} \chi^2_p. \] (5.56)
Note that the Lagrange Multiplier test only requires the restricted values of the parameters.

In general, we may test $H_0: r(\theta) = 0$ with

$$LM = -\frac{\partial L(\tilde{\theta})}{\partial \theta} \left( \frac{\partial^2 L(\tilde{\theta})}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial L(\tilde{\theta})}{\partial \theta} d \rightarrow \chi^2_q,$$

where $L(\cdot)$ is the unrestricted log-likelihood function, and $\tilde{\theta}$ is the restricted MLE. Thus we are testing the unrestricted derivatives of the log-likelihood or scores for mean zero at the restricted estimates. Consequently, the Lagrange Multiplier test is sometimes called the score test.

### 5.3.4 Comparing the Tests

In the scalar case the relationship between the three tests can be seen graphically in Figure 5.2. With the parameter values on the horizontal axis and the log-likelihood values on the vertical, we plot an example log-likelihood function. The maximum occurs at $\hat{\theta}$ with the corresponding log-likelihood at $L(\hat{\theta})$. The restricted parameter value, given by the null hypothesis is $\theta_0$ with corresponding log-likelihood $L(\theta_0)$, which is smaller. The score function $\frac{\partial L(\theta_0)}{\partial \theta}$ is represented by the slope of the tangent line at $L(\theta_0)$.

The Wald test is based on the (squared) difference $\hat{\theta} - \theta_0$ with an adjustment made for the variability of $\theta$ as measured by the inverse of the estimated information matrix (ignoring the $n$ factor). The likelihood ratio is based on the difference $L(\hat{\theta}) - L(\theta_0)$ with no adjustment needed. Obviously, for a particular value of $(\hat{\theta} - \theta_0)^2$, the more pronounced the peak of the log-likelihood the more precise the estimates and the smaller the estimated variances we divide by in the Wald statistic, which makes the adjusted statistic larger as would be the case with the log-likelihood difference. The Lagrange multiplier test is based on the slope of the log-likelihood function at the restricted value of the parameter, which will be larger for the more pronounced peak case. This (squared) slope will be adjusted by dividing by second-order curvature, which is inversely related to the variance and hence larger for the more peaked case. In any event, larger values of $(\hat{\theta} - \theta_0)^2$ will occur with larger values of $(\frac{\partial L(\theta_0)}{\partial \theta})^2$, and $L(\hat{\theta}) - L(\theta_0)$.

The above analysis demonstrates that the three tests: likelihood ratio, Wald, and Lagrange multiplier are asymptotically equivalent. In large samples, not only do they have the same limiting distribution, but they will accept and reject together under $H_0$. This in not the case in finite samples where one can reject when the other does not. This might lead a cynical analyst to use one rather than the other by choosing the one that yields the results (s)he wants to obtain. Making an informed choice based on their finite sample behavior is beyond the scope of this course.

In many cases, however, one of the tests is a much more natural choice than the others. Recall that the Wald test only requires the unrestricted estimates while the Lagrange multiplier test only requires the restricted estimates. In some
CHAPTER 5. MAXIMUM LIKELIHOOD METHODS

\[
\frac{\partial \ell(\theta)}{\partial \theta} = \log L(\theta) = L(\hat{\theta}) - L(\theta_0)
\]

Figure 5.2: Comparison of LR, W, and LM Statistics

cases the unrestricted estimates are much easier to obtain than the restricted and in other cases the reverse is true. In the first case we might be inclined to use the Wald test while in the latter we would prefer to use the Lagrange multiplier.

Another issue is the possible sensitivity of the test results to how the restrictions are written. For example, \( \theta_1 + \theta_2 = 0 \) can also be written \( -\theta_2/\theta_1 = 1 \). The Wald test, in particular, is sensitive to how the restriction is written. This is yet another situation where a cynical analyst might be tempted to choose the “normalization” of the restriction to force the desired result. The Lagrange multiplier test, as presented above, is also sensitive to the normalization of the restriction but can be modified to avoid this difficulty. The likelihood ratio test however will be unimpacted by the choice of how to write the restriction.
5.3. MAXIMUM LIKELIHOOD INFERENCE

5.A Appendix

We will now establish that the conditions for Theorem 4.5, are met under the
maximum likelihood Assumptions (M1)-(M5), with one addition. We need to
explicitly add the condition (C1), that the parameter space is compact, which
for the scalar case and the space being the real line means that the space is closed
and bounded. As was demonstrated in the simple example such an assumption
may be implicit in bounding the third derivatives.

Again, for simplicity, we restrict attention to the scalar parameter case.
Define
\[ \psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\theta) \]
and
\[ \psi_0(\theta) = E[\ln f(y_i|\theta)] \]
then the condition (C4) follows from Assumption (M4) by the arguments given
regarding the consistency of the global maximum given above.

This leaves us with the continuity (C3) and uniform convergence in proba-
bility (C2) to establish. In proving these conditions are met, we will utilize the
following uniform law of large numbers from Newey and McFadden (N-M):

**Lemma 2.4 (Newey and McFadden)** If the data \( z_i \) are i.i.d., \( a(z_i, \theta) \)
is continuous at each \( \theta \in \Theta \) with probability one, and there is \( d(z) \) with
\[ \|a(z_i, \theta)\| \leq d(z) \] for all \( \theta \in \Theta \) and \( E[d(z)] < \infty \), then \( E[a(z_i, \theta)] \) is continuous
and \( \sup_{\theta \in \Theta} \left\| n^{-1} \sum_{i=1}^{n} a(z_i, \theta) - E[a(z, \theta)] \right\| \overset{p}{\to} 0. \)

Assumption (M2) allows us to expand in a Taylor’s series to obtain

\[ \psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\theta^0) + \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} (\theta - \theta^0) + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2} (\theta - \theta^0)^2 \]
\[ + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{6} \frac{\partial^3 \ln f(y_i|\theta^*)}{\partial \theta^3} (\theta - \theta^0)^3 \]

where \( \theta^* \) lies between \( \theta \) and \( \theta^0 \). As was shown in the preliminaries of this
chapter, expectations of the first two terms being averaged exist under (M2) and
(M3). And, by (M5) the third is bounded by a function with finite expectation
so its’ expectation exists for all \( \theta^* \). Thus taking expectations of \( \psi_n(\theta) \) and
subtracting yields

$$
\psi_n(\theta) - \psi_0(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{ \ln f(y_i|\theta^0) - E[\ln f(y_i|\theta^0)] \} \\
+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta'} - E \left[ \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} \right] \right\} (\theta - \theta^0) \\
+ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left\{ \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2} - E \left[ \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2} \right] \right\} (\theta - \theta^0)^2 \\
+ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{6} \left\{ \frac{\partial^3 \ln f(y_i|\theta^0)}{\partial \theta^3} - E \left[ \frac{\partial^3 \ln f(y_i|\theta^0)}{\partial \theta^3} \right] \right\} (\theta - \theta^0)^3.
$$

Applying the sup operator to this for $\theta \in \Theta$ yields

$$
\sup_{\theta \in \Theta} ||\psi_n(\theta) - \psi_0(\theta)|| \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \{ \ln f(y_i|\theta^0) - E[\ln f(y_i|\theta^0)] \} \right| \\
+ \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta'} - E \left[ \frac{\partial \ln f(y_i|\theta^0)}{\partial \theta} \right] \right\} (\theta - \theta^0) \right| \\
+ \frac{1}{2} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2} - E \left[ \frac{\partial^2 \ln f(y_i|\theta^0)}{\partial \theta^2} \right] \right\} (\theta - \theta^0)^2 \right| \\
+ \frac{1}{6} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial^3 \ln f(y_i|\theta^0)}{\partial \theta^3} - E \left[ \frac{\partial^3 \ln f(y_i|\theta^0)}{\partial \theta^3} \right] \right\} (\theta - \theta^0)^3 \right|.
$$

Now the average in the first line is a function only of $y_i$ and converge in probability to zero. Likewise the averages in the second and third line plus the fact that $(\theta - \theta^0)$ is bounded means each line converges in probability to zero. For the last line, we define $a(z, \theta) = \frac{\partial \ln f(x|\theta)}{\partial \theta}$, which by Assumption (M5) is bounded in absolute value by $H(y)$ with finite expectation. Applying the Lemma to this result together with $(\theta - \theta^0)$ being bounded means the last line will converge in probability to zero. But this means that $\sup_{\theta \in \Theta} ||\psi_n(\theta) - \psi_0(\theta)|| \to 0$ and moreover $\psi_0(\theta)$ is continuous, which means (C2) and (C3) are satisfied.

A similar approach can be used to verify that Assumptions (M1)-(M5) satisfy the conditions of Theorem 4.6.

Using Lemma 2.4, it is possible to obtain a consistency result for maximum likelihood that does not require existence of derivatives.

**Theorem 5.1.** Suppose (i) $y_i$, $i = 1, 2, \ldots, n$, are i.i.d. with p.d.f. $f(y_i|\theta)$, (ii) $f(y_i|\theta) \neq f(y_i|\theta_0)$ for $\theta \neq \theta_0$, (iii) $\Theta$ is compact, (iv) $\ln f(y_i|\theta)$ continuous at each $\theta \in \Theta$ with probability one, and (v) $E[\sup_{\theta \in \Theta} |\ln f(y_i|\theta)|] < \infty$ then $\hat{\theta} \to_{p} \theta^0$.

**Proof:** We proceed by verifying the conditions of Theorem 4.5. (C1) is assured by (iii). (C4) is guaranteed by (ii) and arguments yielding (5.23). Define
a(z_i, \theta) = \ln f(y_i|\theta) and \psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\theta), then (iv) and (v) allow us to apply Lemma 2.4 of N-M to show (C2) and (C3) are satisfied. □

We now turn to sufficient conditions for interchanging differentiation and integration of a function g(y, \theta). These conditions are given as Theorem 1.3.2 in Amemiya(1985). Suppose (i) \frac{\partial}{\partial \theta} g(y, \theta) continuously in \theta \in \mathcal{N} and y, where \mathcal{N} is an open set, (ii) \int g(y, \theta) dy exits, and (iii) \int \left\| \frac{\partial}{\partial \theta} g(y, \theta) \right\| dy < M < \infty for \theta \in \mathcal{N}, then \frac{\partial}{\partial \theta} \int g(y, \theta) dy = \int \frac{\partial}{\partial \theta} g(y, \theta) dy for \theta \in \mathcal{N}. This means the integrated norm of \frac{\partial}{\partial \theta} g(y, \theta) is bounded by a constant. Verification of (M3) in the Chapter requires verification of these conditions for f(y, \theta) and is a similar exercise to verifying (M5). Note that these conditions are on f(y, \theta) rather than ln f(y, \theta).

With these sufficient conditions we can adapt the asymptotic normality results of Theorem 4.6 to the specific case of maximum likelihood and obtain a result with more primitive assumptions.

**Theorem 5.2.** Suppose (i) y_i i.i.d f(y, \theta), (ii) \hat{\theta} \rightarrow_p \theta, \theta_0 interior \Theta, (iii) f(y, \theta) twice continuously differentiable, f(y, \theta) > 0 in a neighborhood \mathcal{N} of \theta_0, (iv) \sup_{\theta \in \mathcal{N}} \int \left\| \frac{\partial^2}{\partial y \partial \theta} f(y, \theta) \right\| dy < \infty, sup_{\theta \in \mathcal{N}} \int \left\| \frac{\partial^2}{\partial \theta^2} f(y, \theta) \right\| dy < \infty, (v) \theta = E \left[ \frac{\partial^2}{\partial y \partial \theta} f(y, \theta) \right] exists and nonsingular, (vi) E[\sup_{\theta \in \mathcal{N}} \left\| \frac{\partial^2}{\partial \theta^2} f(y, \theta) \right\|] < \infty, then \sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, \theta^{-1}) □

**Proof:** We proceed by verifying the conditions of Theorem 4.6. Define \psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\theta) then (N1) and (N2) are assured (ii) and (iii). Now (iv) assures that the zero mean score and information matrix equality results apply so we have E \left[ \frac{\partial \ln f(y, \theta)}{\partial \theta} \right] = 0 and E \left[ \frac{\partial \ln f(y, \theta)}{\partial \theta} \frac{\partial \ln f(y, \theta)}{\partial \theta} \right] = \theta \neq E \left[ \frac{\partial^2 \ln f(y, \theta)}{\partial \theta^2} \right].

Define a(z_i, \theta) = \frac{\partial^2}{\partial \theta^2} \ln f(y_i|\theta) then by (i), (iii), and (vi), we can apply Lemma 2.4 to find that \Psi(\theta) = E \left[ \frac{\partial^2}{\partial \theta^2} \ln f(y, \theta) \right] continuous and sup_{\theta \in \mathcal{N}} \left\| \frac{\partial}{\partial \theta} \ln f(y_i|\theta) - \Psi(\theta) \right\| \rightarrow_p 0 so (N3) is satisfied. But \Psi = \Psi(\theta^0) = -\theta by the information matrix equality and nonsingular by (v) so (N4) is met. And by (i) we can apply Lindberg-Levy to find \sqrt{n} \psi_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(y_i|\theta) \rightarrow_d N(0, \theta) and (N5) is satisfied with V = \theta. Since the conditions of Theorem 4.6 are met and \Psi^{-1} V \Psi^{-1} = \theta^{-1} we have the final result. □