PROBABILITY AND SAMPLING THEORY
MSc. in MTF, QF & FM Induction Programme
Lecture Notes (Academic Year 2010-2011)

Giovanni Urga
Professor of Finance and Econometrics
Faculty of Finance, Cass Business School, London (UK) &
Director, Centre for Econometric Analysis (CEA@Cass)
and
Professor of Econometrics
Dipartimento Scienze Economiche, Bergamo (Italy)
## Contents

0.1 Random Variables and Probability Distributions: A Review ................................................................. iii

0.1.1 Introduction ........................................... iii

0.1.2 Probability theory ........................................ vii

0.1.3 Random Variables and Probability Distribution .............. xi

0.1.4 Mathematical Expectation and Moments ................. xviii

0.1.5 Some Special Distributions .............................. xxvi

0.2 Sampling Theory ......................................................... xxxii

0.2.1 Introduction ..................................................... xxxiii

0.2.2 Methods for Finding Point Estimators ..................... xxxiv

0.2.3 Properties of Point Estimators .............................. xl

0.2.4 Interval estimation and hypothesis testing ............... lii

### 0.1 Random Variables and Probability Distributions: A Review

#### 0.1.1 Introduction

We plan to revise the foundations of probability and statistical theory with a view to applying them to econometric methodology.

Because econometrics generally deal with the study of several unknown parameters, the discipline of econometrics uses statistical techniques to develop and apply tools for:

- a.) estimating economic/financial relationships
- b.) testing hypotheses involving economic behaviour
- c.) forecasting the behaviour of economic variables.

We will stress the relationship between ECONOMIC/FINANCIAL THEORY/IES and ECONOMETRIC METHODS later on in the programme. At
this stage we focus on the relationship between ECONOMETRICS and STATISTICS.

An econometrician typically formulates a theoretical model that is a framework for analysing economic behaviour using some underlying logical structure. For the time being we do not want to know how his model arose.

What we want to stress is the fact that a well-specified model would be a reasonable approximation to the actual process that generates the observed data. This process (known as the Data Generating Process or DGP) would involve the interaction of behaviour amongst numerous economic agents.

Once the model has been formulated and the data gathered, the next step in an empirical study is

- to estimate the unknown parameters of the model
- and
- to subject the model to a variety of diagnostic tests to make sure that one obtain robust conclusions, that is conclusions that are not sensitive to model specification.

To achieve this goal, an investigator may have to reformulate the models and perhaps use alternative techniques to estimate them. Methods of hypothesis testing would be useful not only at this diagnostic testing stage but also to test the validity of a body of theories.

Because economics/finance deals with non experimental data generated by a complicated process involving the interaction of the behaviour of numerous economic and political agents, this imposes a great deal of uncertainty in the models and methods used by the econometricians.

In particular, estimated relations are not precise, hypothesis testing lead to the error of rejecting a true hypothesis or that of accepting a false hypothesis, and forecasts of variables often turn out to be far from their actual values. This uncertainty makes statistical methodology very important in econometrics.

The mathematical and statistical model that an econometrician formulates is meant to be a description of the DGP. The actual data, however, are treated as one of several possible realisations of events.

The theory of probability is useful to construct a framework that portrays the likelihood of one type of realisation or another. Not surprisingly, the probability framework invariably depends on a number of parameters. In order to estimate the parameters, one typically obtains a sample of observations and uses them in conjunction with a probability model.
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

STATISTICAL INFERENCE deals with the methods of obtaining these estimates, measuring their precision, and testing hypotheses on the parameters in the equations.

STATISTICS is a body of knowledge that is useful for collecting, organising, presenting, analysing, and interpreting data and numerical facts.

- DESCRIPTIVE STATISTICS: deals with the presentation and organisation of data. Measures of central tendency, such as mean and median, measures of dispersion, such as standard deviations and range are descriptive statistics.

- INFERENTIAL STATISTICS: deals with the use of sample data to infer, or reach, general conclusions about a much larger population. In stats, we define a population as the entire group of individuals we are interested in studying. A sample is any subset of such a population.
We also encounter another dichotomy in statistical analysis: DEDUCTION vs. INDUCTION.

- **DEDUCTION**: is the use of general information to draw conclusions about specific cases. The use of probability to determine the chance of obtaining a particular kind of sample result is known as deductive statistical analysis.

- **INDUCTION**: involves drawing general conclusions from specific information. This means that on the strength of a specific sample, we infer something about a general population. The sample is all that is known. We must determine the uncertainty characteristics of the population from the incomplete information available. This strand of statistical analysis is called inductive statistics analysis.

PROBABILITY THEORY: We will learn how to apply deductive techniques when we know everything about the population in advance and are concerned with studying the characteristics of the possible samples that may arise from that known population.
0.1.2 Probability theory
Basic Probability

1. Sample space, sample points and events

The term probability is loosely used by many persons to indicate the measure of one’s belief in the occurrence of an uncertain future event. There is no single definition of the term probability that is universally accepted. Various definitions will be introduced.

EXPERIMENTS (as simple as rolling a pair of dice or as complicated as conducting a large-scale survey of household or firms).

RANDOM EXPERIMENTS: if it satisfies the following conditions:
1. all possible distinct outcomes are known ahead of time;
2. the outcome of a particular trial is not known a priori;
3. the experiment can be duplicated in principle under ideal conditions.

The totality of all possible outcomes of the experiments is referred to as the SAMPLE SPACE (denoted by \( S \)) and its distinct individual elements are called the SAMPLE POINTS or ELEMENTARY EVENTS.

An EVENT is a subset of a sample space and is a set of sample points that represents several possible outcomes of an experiments (denoted by \( A \)).

The IMPOSSIBLE EVENT or NULL EVENT is denoted by \( 0 \). A sample space with a finite or COUNTABLE INFINITE sample points (with a one to one correspondence to positive integers) is called a DISCRETE SPACE.

A CONTINUOUS SPACE is one with an UNCOUNTABLY INFINITE number of points, i.e. it has as many elements as there are real numbers.

2. Some results from set theory

- The SAMPLE SPACE is denoted by \( S \). \( A = S \) implies that the events in \( A \) must always occur. The EMPTY SET is a set with no elements and is denoted by \( 0 \). \( A = 0 \) implies that the events in \( A \) do not occur.

- The set of all elements not in \( A \) is called the COMPLEMENT of \( A \) and is denoted by \( A^c \) or \( \overline{A} \). Thus \( A^c \) occurs if and only if \( A \) does not occur.
• The set of all points in either $A$ or a set $B$ or both is called the UNION of the two sets and is denoted by $\cup$. $A \cup B$ means that either the event $A$ or the event $B$ or both occur. Note that $A \cup A^c = S$.

• The set of all elements in both $A$ and $B$ is called the INTERSECTION of the two sets and is represented by $\cap$. $A \cap B$ means that both the events $A$ and $B$ occur simultaneously.

• $A \cap B = 0$ implies that $A$ and $B$ cannot occur together. $A$ and $B$ are then said to be DISJOINT or MUTUALLY EXCLUSIVE. Note that $A \cap A^c = 0$.

• $A \subset B$ means that $A$ is contained in $B$ or that $A$ is a subset of $B$, that is, every element of $A$ is an element of $B$. If an event $A$ has occurred, then $B$ must have occurred also.

3. Probability: definitions and concepts

The probability of an event is defined in a number of ways, all of which are useful in calculating probabilities.

**AXIOMATIC DEFINITION**
The probability of an event $A \subset F$ is a real number such that

1. $P(A) \geq 0; \forall A \subset F$

2. The probability of the entire sample space $S$ is 1, $P(S) = 1$

3. If $A_1, A_2, \ldots, A_n$ are mutually exclusive events (i.e. $A_1 \cup A_2 \cup \ldots \cup A_n = 0$ for all $i \neq j$), then $P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum P(A_i)$ and this holds for $n = \infty$ also.

Note that the triple $(S, F, P)$ is referred to as the PROBABILITY SPACE and $P$ is a PROBABILITY MEASURE.

Although the axiomatic definition is rigorous, it does not directly tell us how to assign probabilities to elementary events. That is accomplished by two other definitions given below. All three definitions are then used to calculate probabilities of various events.

**CLASSICAL DEFINITION**
If an experiments has $n$ ( $n < \infty$) mutually exclusive and equally likely outcomes, and if the event $A$ occurs in $n_A$ possible ways, then the probability of $A$ is $n_A/n$, denoted as $P(A) = n_A/n$. 

© Giovanni Urga, 2009
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

**FREQUENCY DEFINITION**
Let $n_A$ be the number of times the event $A$ occurs in $n$ trials of an experiment. If there exists a real number $p$ such that

$$p = \lim_{n \to \infty} \frac{n_A}{n}$$

then $p$ is called the probability of $A$ and is denoted by $P(A)$.

**SUBJECTIVE PROBABILITY**
In many instances, individuals use personal judgements to assess the relative likelihoods of various outcomes. For instance, one often hear the expression “the odds are 2 to 1 that candidate $X$ will be elected”.

This is an example of subjective (or personal) probability which is based on “educated guesses” or intuition. In statistical inference, the practicability of this approach stems from using prior beliefs or new information in updating previous model specification.

This provides the foundation for the Bayesian approach.

In sum, the axiomatic approach to probability theory is rich enough to include each of the interpretations of probability and using the axioms (1) - (3) the following properties of probabilities can be proved:

- $P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- $P(0) = 0$
- If $A \subset B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

**Conditional Probability**
There is the case when we wish to calculate the probabilities of events when it is known that another event occurred. It would be useful to know how the probability of an event alters as a result of the occurrence of another event.

**Definition:** Let $A$ and $B$ be two events in a probability space $(S, F, P)$ such that $P(B) \geq 0$. The conditional probability of $A$ given that $B$ will occur or has occurred, denoted by $P(A \mid B)$, is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Note that we can re-write the expression above as:

© Giovanni Urga, 2009 ix Adv Stats Induction, 2009-2010
\[ P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]

More generally, if

\[ A_1, A_2, \ldots, A_n \]

are events such that

\[ P(A_1) > 0, P(A_1 \cap A_2) > 0, \ldots, P(A_1 \cap A_2 \cap A_3 \ldots \cap A_{n-1}) > 0 \]

then

\[ P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \ldots P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \]

This definition of conditional probability leads to the definition of INDEPENDENCE OF EVENTS.

**Independence of Events**

Suppose that \( P(A \mid B) = P(A) \). This would imply that knowing that event \( B \) occurred has no effect on the chance of event \( A \) occurring. It would be reasonable then to say that events \( A \) and \( B \) are independent event. So, two events \( A \) and \( B \) are independent when:

\[ P(A \cap B) = P(A \mid B)P(B) = P(A)P(B) \]

Otherwise \( A \) and \( B \) are said to be dependent. Clearly if the expression above is true, then \( P(B \mid A)P(B) \) as well. This definition extends to more than two events, for example events \( A, B, \) and \( C \) are independent if

\[ P(A \cap B \cap C) = P(A)P(B)P(C) \]

**Bayes’ Rule**

Let \( S \) denote the sample space of some experiment. The disjoint events

\[ A_1, A_2, \ldots, A_n \]

are a partition of \( S \) if

\[ A_1 \cup \ldots \cup A_n = S \]
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

and \( P(A_i) > 0 \) for each \( i \). If \( B \) is any event in \( S \) so that \( P(B) > 0 \), then it will be true that

\[
B = (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)
\]

\[\text{Draw Venn diagram to check!!!}\]

Since the events \( A_i \cap B \) are disjoint

\[
P(B) = \sum_{i=1}^{n} P(A_i \cap B)
\]

Also, if \( P(A_i) > 0 \), the

\[
P(A_i \cap B) = P(B | A_i)P(A_i)
\]

so

\[
P(B) = \sum_{i=1}^{n} P(B | A_i)P(A_i)
\]

which is called the law of total probability

The formulation of a partition of \( S \) means that when the experiment is performed then exactly one of the events \( A_i \) will occur. Situations arise in which we would like to determine the probability of the event \( A_i \) given that the event \( B \) occurs. The solution to this problem is given by the Bayes’ theorem

\[
P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B | A_i)P(A_i)}{P(B)}
\]

\[
P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{i=1}^{n} P(B | A_i)P(A_i)}
\]

0.1.3 Random Variables and Probability Distribution

RANDOM VARIABLES

RANDOM VARIABLES: is a real-valued function that assigns a number to each sample point (outcome) in the sample space of an experiment

It is important to distinguish between the rule or function that assigns the numbers to each sample point and the numerical values themselves.

© Giovanni Urga, 2009
In general, we distinguish between the two by assigning random variables as upper case letter, $X$, and the values of the random variable by lower case letter $x$. So $P(X = x)$ or $P(X < 3)$ mean that the probability that the random variable $X$ takes the value $x$ or a value less than 3, respectively.

**Discrete:** whose set of all possible values is finite or countable infinite

**Continuous:** is one that may assume all values in at least one interval of the real number line

We will assume that we have a probability measure $P$ defined on the sample space $S$ that is associate with some experiments. It should not be surprising that we can use the assigned probability measure to make probabilistic statements about the possible values of a random variable. That is, we can construct a probability distribution or probability density function (p.d.f.) for $X$ that will literally tell us how the probability mass (total mass = 1) is distributed (i.e. allocated or spread out) across the values $x$ that $X$ can take. This p.d.f. can be used for example to compute the probability that $X$ takes values or falls in a given range.

**PROBABILITY DISTRIBUTION FOR DISCRETE RANDOM VARIABLES:**

For a discrete random variable $X$, a probability density function (or Probability Function) is defined to be the function $f(x)$ such that for any real number, $x$, which is the value that $X$ can take, $f(x) = P(X = x)$. Thus $0 \leq f(x) \leq 1$.

If $x$ is not one of the values that $X$ can take, then $f(x) = 0$. Also, if the sequence

$$x_1, x_2, \ldots$$

includes all values that $X$ can take then

$$\sum_{i=1}^{\infty} f(x_i) = 1$$

A function closely related to the probability density function of a random variable is the corresponding DISTRIBUTION FUNCTION or CUMULATIVE DISTRIBUTION FUNCTION (c.d.f.)

The c.d.f. of a random variable $X$ is defined for each real number $x$ as $F(x) = P(X \leq x)$ for $-\infty < x < +\infty$. That is, $F(X)$ is the probability that the random variable $X$ takes a value less than or equal to $x$. 

© Giovanni Urga, 2009 xii Adv Stats Induction, 2009-2010
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

For a discrete random variable $X$,

$$F(X) = \sum_{t \leq x} f(t)$$

where this summation notation means sum all the values of the p.d.f. $f(t)$ for the values the random variable can take less than or equal to the specified value, $x$. 

© Giovanni Urga, 2009 xiii Adv Stats Induction, 2009-2010
A continuous random variable $X$ can take any value in at least one interval on the real number line. Suppose $X$ can take all possible values $a \leq x \leq b$. Since the possible values of $X$ are uncountable, the probability that $X$ takes any particular value is zero! So, unlike the situation for discrete random variable, the p.d.f of the continuous random variable, say $f(x)$, will not give the probability that $X$ takes the value $x$.

Instead, the p.d.f, $f(x)$ of a continuous random variable $X$ will be such that the area under $f(x)$ will give probabilities associated with the corresponding intervals on the horizontal axis. More specifically, a function with values $f(x)$, defined over all real $x$, is a p.d.f. for the continuous random variable $X$ if

(i) $f(x) \geq 0$
(ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$
(iii) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$ for any $a, b$ such that $-\infty < a < b < \infty$

Graphically:
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

The cumulative distribution for a continuous random variable \( X \) is given by

\[
F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \quad \text{for } -\infty < x < \infty
\]

where \( f(t) \) is the value of the p.d.f at \( t \). Note that \( F(-\infty) = 0 \), \( F(\infty) = 1 \)

Furthermore,

\[
P(a \leq X \leq b) = F(b) - F(a) \quad \text{and} \quad f(x) = dF(x)/dx = F'(x) \quad \text{where the derivative exists}
\]

MULTIVARIATE DISTRIBUTIONS

So far we have been concerned only with a single random variable \( X \) and functions of \( X \). In most cases, however, the outcome of an experiment may be characterized by more than one random variable. For instance \( X \) may be Income and \( Y \) the total Expenditure of a household. Thus we may observe a pair of random variables \( (X, Y) \). If family size \( Z \) is also added, we observe \( (X, Y, Z) \). We would thus be interested in the study of Multivariate Distribution.

Let us consider with the bivariate case:

If \( X \) and \( Y \) are discrete random variables \( f(x, y) = P(X = x, Y = y) \), defined in each pair of values \((x, y)\) that \( X \) and \( Y \) can take, is called the joint probability distribution.

A joint probability function of \( X \) and \( Y \) \( F(x, y) \geq 0 \) for \((x, y)\) in its domain \( \sum_x \sum_y f(x, y) = 1 \)

If \( A \) and \( B \) are sets of values of \( x \) and \( y \) respectively then \( P(X \in A, Y \in B) = \sum_{x \in A} \sum_{y \in B} f(x, y) \quad \text{for } (x, y) \in R^2 \) where \( R^2 \) is a real two dimensional plane

If \( X \) and \( Y \) are continuous random variables, then the bivariate function \( f(x, y) \) with values defined for \((x, y)\) is a joint p.d.f for \( X \) and \( Y \) if

\[
P[(X, Y) \in A] = \int \int_A f(x, y)dx \quad \text{for any region } A \in R^2 \text{ that represent an event.}
\]

The function \( f(x, y) \) must have the properties that \( f(x, y) \geq 0 \) and

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)dxdy = 1
\]

The correspondent joint cumulative distribution function for the continuous random variables \( X \) and \( Y \) is for \((x, y)\) \( \in R^2 \). Analogous to \( f(x) = \delta F(x)/\delta x \) is \( f(x, y) = \delta^2 F(x, y)/\delta x \delta y \) (if the partial derivatives exist).
MARGINAL DISTRIBUTIONS

Let us consider the previous sample. Suppose we simply want to know the probability distribution of $X$, say $g(x)$. How can we use the joint p.d.f for $X$ and $Y$ to obtain $g(x)$?

The discrete random variable $X$ can take the values 1 and 0, and

$$g(1) = P(x = 1) = f(1, 1) + f(1, 0) = b/T + d/T = \sum_{y=0}^{1} f(1, y)$$

$$g(0) = P(x = 0) = f(0, 1) + f(0, 0) = a/T + c/T = \sum_{y=0}^{1} f(0, y)$$

Or more compactly, $g(x) = \Sigma_y f(x, y)$, meaning that the p.d.f of $X$, $g(x)$ is obtained by summing the joint p.d.f of $X$ and $Y$ over all values of $Y$ for each value $x$ of $X$.

The p.d.f $g(x)$ is called a marginal distribution because if we obtain the row and column sums of $f(x, y)$, and display them in the margins of the table, we have

<table>
<thead>
<tr>
<th></th>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$b/T$</td>
</tr>
<tr>
<td>0</td>
<td>$d/T$</td>
</tr>
<tr>
<td></td>
<td>$(b + d)/T$</td>
</tr>
</tbody>
</table>

So $g(x)$ is given by the row totals. Similarly the p.d.f of $Y$, $h(y)$, is given by the column totals and in general $h(y) = \Sigma_x f(x, y)$. These probability density functions, obtained from the joint p.d.f, are the MARGINAL PROBABILITY DENSITY FUNCTIONS of $X$ and $Y$.

If $X$ and $Y$ are continuous, the summation signs are replaced by integrals and we get

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad -\infty < x < \infty$$

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad -\infty < y < \infty$$

as the marginal probability density functions for $X$ and $Y$. 

CONDITIONAL DISTRIBUTIONS AND INDEPENDENT RANDOM VARIABLES
Previously we defined the conditional probability of event $A$ given event $B$ as $P(A|B) = P(A \cap B)/P(B)$ if $P(B) > 0$. If $X$ and $Y$ are discrete random variables, then

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

Provided $P(Y = y) = h(y) \neq 0$, where $f(x, y)$ is the joint p.d.f of $X$ and $Y$, $h(y)$ is the p.d.f of $Y$ at $y$, and $f(x \mid y)$ is the probability distribution of $X$ given that $Y$ is fixed at the value $y$. The function $f(x \mid y)$ is called the conditional probability density of $X$ given $Y = y$.

If $X$ and $Y$ are continuous random variables, then

$$f(x \mid y) = \frac{f(x, y)}{h(y)} = f(x|y)$$

is the conditional p.d.f of $X$ given $Y = y$.

Having defined the idea of a conditional p.d.f we can now define the INDEPENDENCE OF RANDOM VARIABLES in the same way as we defined independent events. Namely if $f(x \mid y) = f(x)$.

So that knowing that $Y = y$ does not affect the probability distribution of $X$, then it is reasonable to say that $X$ is independent of $Y$, and vice versa. If $f(x|y) = f(x)$ holds, then

$$f(x, y) = f(x \mid y)h(y) = g(x)h(y)$$

We can then formally define $X$ and $Y$ to be statistically independent if

$$f(x, y) = g(x)h(y)$$

for all values $(x, y)$ of $(X, Y)$ in the sample space.

If $X_1, X_2, \ldots, X_n$ have joint p.d.f. $f(x_1, x_2, \ldots, x_n)$ then the random variables $X_1, \ldots, X_n$ are MUTUALLY STOCHASTICALLY INDEPENDENT if

$$f(x_1, \ldots, x_n) = f(x_1) \times f(x_2) \times \cdots \times f(x_n)$$

Using this definition of stochastic independence, we can define a random sample. Suppose we repeat an experiment $n$ times, and let $X_1, \ldots, X_n$ be random variables representing a certain measurement or values for those $n$ trials. Then $X_1, \ldots, X_n$ constitute a random sample if they are independent and identically distributed.

That is,
f(x_1, \ldots, x_n) = f(x_1) \times \cdots \times f(x_n)

and

\[ f(x) = f(x_2) = \cdots = f(x_n) = f(x) \]

### 0.1.4 Mathematical Expectation and Moments

In order to study random variables and their probability distributions, it is useful to define the concept of mathematical expectation of a random variable and of functions of a random variable; these definitions are given in what follows.

#### A) Expected Value of a Random Variable

Consider the experiment of rolling a single die. Let \( X \) be the value that shows on the die. The probability distribution of \( X \) is

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x) )</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

What would be the average value of \( X \) be if the experiment were repeated an infinite number of times? Intuitively, you would expect \( X = 1 \) or 1/6 of the throws, \( X = 2 \) or 1/6 of the throws and so on. So, on the average, the value of \( X \) is

\[
1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5
\]

If \( X \) is a \textit{discrete random variable} and \( f(x) \) is its p.d.f., then the expected value of \( X \) is \( E[X] = \Sigma x f(x) \). This is a weighted average with the values that \( X \) can take, and weights being the probabilities of each value’s occurrence.

If \( X \) is a \textit{continuous random variable} and \( f(x) \) is its p.d.f., then the expected value of \( X \) is

\[
E[X] = \int_{-\infty}^{\infty} x f(x) dx
\]
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

The expected value of $X$, $E[X]$, is also called mean of $X$ or the mean of the distribution of $X$. It is a measure of the average value of $X$ and a measure of the centre of the probability distribution. In fact, $E[X]$ can be regarded as the centre of gravity of the probability distribution. If a p.d.f. is symmetric with respect to a value $x_0$ or the x-axis, then $E[X] = x_0$ if it exists.

Note: Two other common measures of the centre of the distribution are MEDIAN, which is the value of the random variable below which 50% of the probability mass falls, and the MODE, which is the value of the random variables at which its p.d.f. attains the maximum value.

B) EXPECTATION OF A FUNCTION OF A SINGLE VARIABLE

Mathematical expectation of a function of $X$, say $g(x)$

So,

$E[g(x)] = \sum_x g(x)f(x)$ for discrete random variables, and

$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ for continuous random variables

Some results: If $a$ and $b$ are constants

$E[aX + b] = aE[X] + b$ which implies

$E[aX] = aE[X]$ and $E[b] = b$
C) EXPECTATIONS OF FUNCTION OF SEVERAL RANDOM VARIABLES

Suppose \( Y = g(x_1, \ldots, x_n) \) and let \( f(x_1, \ldots, x_n) \) be the joint p.d.f of the continuous random variables \( X_1, \ldots, X_n \). Then

\[
E[Y] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n) \, dx_1, \ldots, dx_n
\]

In the discrete case \( f \) is replaced by \( \Sigma \).

Some useful results:
(i) Let \( a_0, \ldots, a_n \) be constants, then
\[
E[a_0 + a_1 X_1 + \cdots + a_n X_n] = a_0 + a_1 E[X_1] + \cdots + a_n E[X_n]
\]
(ii) Let \( X_1, \ldots, X_n \) be independent random variables such that \( E[X_i] \) exists, then
\[
E[X_1 X_2 \cdots X_n] = E[X_1] E[X_2] \cdots E[X_n]
\]

D) MOMENTS OF A RANDOM VARIABLE

The \( r \)th moment of the random variable \( X \) about the origin is denoted \( \mu_r \) and is \( E[X^r] \). Note that for \( r = 1 \), \( \mu_1 = E[X] \) is the mean (denoted with \( \mu \) also).

Quotation: “moments” comes from physical interpretation of mathematical expectations such that the fact that \( \mu_r \) is the unit of gravity.

The \( r \)th moment about the mean of \( X \) is \( \mu_r = E[(X - \mu)^r] \)
If \( r = 2 \), \( \mu_2 = E[(X - \mu)^2] \) is called variance of the distribution of \( X \) (or variance of \( X \)), often denoted as \( \sigma^2 \) or \( \text{var}(X) \). The positive square root of \( \sigma^2 \), \( \sigma \) is called the standard deviation of \( X \).

It is useful to note that
\[
\sigma^2 = E[(X - E(X))^2] = E[X^2] - (E[X])^2
\]
where \( \sigma^2 \) is the average of the squared distance between the value of \( X \) (random variable) and \( E[X] \) when the experiment is repeated infinitely many times.

The larger the value of \( \sigma^2 \), the greater the average squared distance between the values of the random variables and \( E[X] \), and thus more likely a value of \( X \) far from its mean is to occur.
0.1. **RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW**

We can graphically represent this part $X$ and $Y$ and let us suppose $\text{var}(x) > \text{var}(y)$

![Graphical representation of random variables X and Y](image)

The probability mass of $X$ is more “spread-out” than that of $Y$. The variance of $X$ is a measure of the dispersion of the probability mass of $X$ about its mean. The greater the value of $\text{var}(X)$, the wider the dispersion of the values of $X$ about its mean.

**Product Moments**

(a) the $s^{th}$ and $r^{th}$ product moment about the origin of the random variables $X$ and $Y$ is given by

$$\mu_{rs} = E[X^sY^r]$$

(b) the $s^{th}$ and $r^{th}$ product moment about the means of the random variables $X$ and $Y$ is given by

$$\mu_{rs} = E[(X - E(X))^r(Y - E(Y))^s]$$

The covariance between two random variables is a measure of the association between them. Let $\mu_{11} = \text{covariance between X and Y}$, also denoted with $\sigma_{xy}$ or $\text{cov}(X,Y)$ i.e.

$$\sigma_{xy} = E[(X - E(X))(Y - E(Y))]$$

$$\sigma_{xy} = E[XY] - E[X]E[Y]$$

Examining $\sigma_{xy}$ and interpreting $E[.]$ as “the average in infinitely many trials”, we can see that $\text{cov}(X,Y) > 0$ if $X$ and $Y$ jointly tend to be greater than
their means, or less than their means - on average. $\text{Cov}(X,Y)$ can also be 0 or negative with obvious meaning.

To see $\text{cov}(X,Y)$ measure linear association, let us define a related measure, the **CORRELATION** between $X$ and $Y$ ($\rho_{xy}$) and defined as

$$\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}$$

- The $\rho_{xy}$ is a positive number and falls between $-1$ and $+1$.
- Let $\text{cov}(X,Y) = 0$ then $\rho_{xy} = 0$
- If $X$ and $Y$ are independent, then $\rho_{xy} = 0$
- If $Y = a + bX$ for some constants $a$ and $b$ ($a,b \neq 0$), the $|\rho_{xy}| = 1$ and $X$ and $Y$ are said to be perfectly correlated. The more the values of $X$ and $Y$ are linearly related, the larger $|\rho_{xy}|$

Some results:

(a) If $a_0 \ldots a_n$ are constants and $X_1 \ldots X_n$ random variables, then

$$\text{var}(a_0 + a_1 X_1 + \ldots + a_n X_n) = \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + 2 \sum_{i<j} a_i a_j \text{cov}(X_iX_j)$$

(b) $\text{var}(a_0 + a_1 X_1) = a_1^2 \text{var}(X_1)$

(c) $\text{var}(X_1 \pm X_2) = \text{var}(X_1) + \text{var}(X_2) \pm 2\text{cov}(X_1X_2)$

Note: if the $X_i$ are statistically independent, then the covariance terms drop out of (a) and (c)

**COEFFICIENT OF VARIATION** = $\sigma/\mu$

**SKEWNESS AND KURTOSIS:**

[A]

If a continuous density $f(x)$ has the property that $f(\mu + a) = f(\mu - a)$ for all $a$ ($\mu$ being the mean of the distribution), then $f(x)$ is said to be **SYMmetric**

© Giovanni Urga, 2009 xxii Adv Stats Induction, 2009-2010
AROUND THE MEAN. If a distribution is not symmetric about the mean, then it is called SKEWED.

A commonly used measure of skewness is

$$E[(X - \mu)^3 / \sigma^3]$$

For a symmetric distribution this is zero.

[B]

The peakness of a distribution is called KURTOSIS. A narrow distribution is called leptokurtic and a flat distribution is called platykurtic.

One measure of kurtosis is
The value

\[ E[(X - \mu)^4/\sigma^4] - 3 \]

is often referred to as EXCESS KURTOSIS.

\[ S \] and \[ K \] will be used in econometrics to characterize forecast errors and non normality of the distributions of such and error.

\[ S \] and \[ K \] are defined on standardized variables.

E) CHEBYSHEV’S THEOREM OR INEQUALITY

It illustrates how the variance and standard deviation are related to the dispersion of the probability mass and very useful for many theoretical proofs in statistics.

If \( X \) is a random variable with mean \( \mu \) and variance \( \sigma^2 \), then

\[ P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2} \]

or

\[ P(|X - \mu| \geq k\sigma) < \frac{1}{k^2} \]

F) EXPECTATIONS INVOLVING MULTIVARIATE R.V.

Let \( \underline{X} \) be a vectors of random variables \( \underline{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \)

where \( E[X_i] = \mu_i \)

\[ \text{var}(X_i) = \sigma_i^2 \]

\[ \text{cov}(X_i, Y_j) = \sigma_{ij} \]

Then

\[ E[\underline{X}] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \mu \]

The expected value of a random vector is the vector of expectations.
We can define covariance matrix in the random vector $X$ as

$$Cov(X) = E[(X - E[X])(X - E[X])']$$

$$Cov(X) = \begin{bmatrix}
(X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_n - \mu_n) \\
(X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_n - \mu_n) \\
\vdots & \vdots & \ddots & \vdots \\
(X_n - \mu_n)(X_1 - \mu_1) & (X_n - \mu_n)(X_2 - \mu_2) & \cdots & (X_n - \mu_n)^2
\end{bmatrix}$$

$$= \begin{bmatrix}
\text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\
\text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n)
\end{bmatrix}$$

$$= \begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\
\sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2
\end{bmatrix}$$

The covariance matrix of $X$, often denoted $\sum_x$, is a **POSITIVE SEMIDEFINITE MATRIX** with variances on the main diagonal and off-diagonal elements that are covariance. If more of the random variables $X_1 \ldots X_n$ are dependent (i.e. if none of them have zero variances), and no exact linear relationships between the $X_i$ exists, the $\sum_x$ is positive definite.

Let $a' = (a_1 \ldots a_n)$ be a vector of constants, then

$$E[a' X] = a' \mu = a_1 \mu_1 + \cdots + a_n \mu_n$$

and

$$\text{Var}(a' X) = a' \sum_x a = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum_{i<j} a_i \sigma_i \sigma_j$$

This result can be expressed:

$$\text{Var}(a' X) = E[|a' X - E[a' X]|^2]$$

$$= E[|a' X - E[a' X]|](a' X - E[a' X])$$

$$= a' E[|X - E[X]|(X - E[X])']a$$

Note: $a' X$ is a $(1 \times 1)$ matrix
Further, let $P$ be a $(m \times n)$ matrix of constants with $m \leq n$. Then $Z = PX$ is a $(m \times 1)$ random vector.

From previous results

$$E[Z] = E[PX] = PE[X] = P\mu$$

and

$$Cov(Z) = cov(PX) = P\sum xP'$$

### 0.1.5 Some Special Distributions

In what follows, I’ll introduce some special distributions that are usually used in econometrics.

**DISCRETE DISTRIBUTIONS:**
- Bernoulli
- Binomial
- Multinomial
- Poisson

**CONTINUOUS DISTRIBUTIONS:**
- Normal
- Bivariate and Multivariate normal
- Gamma
- Chi-square
- t-distribution
- F-distribution

1. Bernoulli Distribution

Consider an experiment in which only two outcomes are possible (H or T). Let us indicate this result as 0 and 1. Then $X$ is said to have a Bernoulli distribution if it can only take the two values 0 and 1 and the probabilities are $P(X = 1) = p$ and $P(X = 0) = 1 - p$ with $0 \leq p \leq 1$.

The p.d.f. of this random variable is

$$f(X|p) = \begin{cases} 
p^x (1 - p)^{1-x} & \text{for } x = 0, 1 \\
0 & \text{otherwise} \end{cases}$$
where the notation \( f(X|p) \) denotes that the p.d.f. of \( X \) depends on the parameter \( p \). It is easy to check that \( E[X] = p \) and \( var(X) = p(1 - p) \)

This random variable arises in econometrics in choice models, where a decision maker must choose between 2 alternatives, like buying a car and not buying a car.

2. Binomial Distribution

If \( X_1, \ldots, X_n \) are independent random variables each having a Bernoulli distribution with parameter \( p \), then \( X = \sum_{i=1}^{n} X_i \) is a discrete random variable that is the number of successes (i.e. Bernoulli experiments with outcome \( X_i = 1 \)) in the \( n \)-trials of the experiment and \( X \) has a binomial distribution.

This random variable has p.d.f.

\[
f(x|n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}
\]

where \( \binom{n}{x} = \frac{n!}{x!(n-x)!} \) is the number of possible combinations of \( n \) things taken \( x \) at a time.

This distribution has two parameters, \( n \) and \( p \), where \( n \) is a positive integer and \( 0 \leq p \leq 1 \). Its mean and variance are

\[
E[X] = \sum_{i=1}^{n} E[X_i] = np
\]

\[
Var(X) = \sum_{i=1}^{n} var(X_i) = np(1 - p)
\]

A related random variable is \( Y = X/n \), which is the proportion of successes in \( n \) trials of an experiment. Its mean and variance are

\[
E[Y] = p
\]

\[
Var(Y) = p(1 - p)/n
\]

3. Normal Distribution

\( \copyright \) Giovanni Urga, 2009 xxvii Adv Stats Induction, 2009-2010
It is the most important distribution in statistics. An important reason in the predominance of the normal distribution is the CENTRAL LIMIT THEOREM, which says that many important functions of observations, like the sample mean, tend to be usually distributed given a large enough sample, no matter what the distribution of the original population. The Normal Distribution has many nice mathematical properties that make it convenient to work with.

A continuous random variable \( X \) has a normal distribution with mean \( \mu \) and variance \( \sigma^2 \) \((-\infty < \mu < +\infty, \sigma^2 > 0)\), often denoted as \( X \sim N(\mu, \sigma^2) \), if \( X \) has p.d.f. of the form

\[
f(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}
\]

with \(-\infty < x < +\infty\).

This p.d.f. is “bell-shaped” as represented in the following graph

The p.d.f. is symmetric about the parameter \( \mu \), which is the mean, median, and mode of the distribution.

\[
E[X] = \mu
\]

\[
Var(X) = \sigma^2
\]

Also, the p.d.f. has points of inflection at \( x = \mu \pm \sigma \)

Note: the p.d.f. of a normal random variable does not have a definite integral that has a closed-form, so probabilities must be computed numerically. The probabilities can be economically presented in one table using a standard form. To see this, we use the fact that a linear function of a normal random variable...
is itself normal. Example: suppose \( X \sim N(\mu, \sigma^2) \) and \( Y = aX + b \) with \( a \) and \( b \) are constants and \( a \neq 0 \). Then \( Y \sim N(a\mu + b, a^2\sigma^2) \)

Consequently

\[
Z = \frac{X - \mu}{\sigma} \sim N(0, 1)
\]

is the Standard Normal random variable. If \( \Phi(z) \) denotes the c.d.f. of \( Z \) evaluated at \( z \), then probability statements about \( X \) can be made in terms of \( X \) any values of \( \Phi(z) \) which are tabulated.

**EXAMPLE:** If \( X \sim N(\mu, \sigma^2) \) then

\[
P(a < x < b) = \Phi \left( \frac{b - \mu}{\sigma} \right) - \Phi \left( \frac{a - \mu}{\sigma} \right)
\]

4. Bivariate Normal Distribution
   (as a link between univariate and multivariate)

Let \( X_1 \) and \( X_2 \) be random variables with joint p.d.f.

\[
f(x_1, x_2 \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = (2\pi)^{-1}\sigma_1\sigma_2\sqrt{1-\rho^2}^{-1/2} \exp \left\{ -\frac{Q}{2} \right\}
\]

where \(-\infty < x_1, x_2 < +\infty\) and

\[
Q = \frac{1}{1-\rho^2} \left[ \frac{(x_1 - \mu_1)}{\sigma_1} \right]^2 - 2\rho \left( \frac{x_1 - \mu_1}{\sigma_1} \right) \left( \frac{x_2 - \mu_2}{\sigma_2} \right) + \left( \frac{x_2 - \mu_2}{\sigma_2} \right)^2
\]

\( f(x_1, x_2) \) is the joint p.d.f. of the bivariate normal \( X_1 \) and \( X_2 \). The distribution of \( X_1 \) and \( X_2 \) are \( N(\mu_1, \sigma_1^2) \) and \( N(\mu_2, \sigma_2^2) \) respectively. The parameter \( \rho = \text{cov}(X_1X_2)/\sigma_1\sigma_2 \)

**Properties:**

- The CONDITIONAL DISTRIBUTION of \( X_2 \) given \( X_1 \) is also a normal, specifically

\[
f(x_2 \mid x_1) \sim N(b, \sigma_2^2(1-\rho^2))
\]

and

\[
b = \mu_2 + \rho(\sigma_2/\sigma_1)(x_1 - \mu_1)
\]

- The conditional mean, \( E[X_2 \mid X_1] \) is a function of \( x \)

© Giovanni Urga, 2009 xxix Adv Stats Induction, 2009-2010
• The bivariate normal has the property that if $\rho = 0$, then $X_1$ and $X_2$ are statistically independent (which is not true of other pair of random variables in general). To see this note that if $\rho = 0$, then

$$f(x_2|x_1) = f(x_2) \sim N(\mu_2, \sigma_2^2)$$

5. Multivariate Normal Distribution

Suppose we have $n$ independent and identically distributed standard normal random variables

$$Z_i \sim N(0, 1) \quad for \ i = 1, \ldots, n$$

Since $Z_i$ are independent, we can obtain their joint p.d.f. as

$$f(z_1, \ldots, z_n) = f(\mathbf{z}) = \prod_{i=1}^{n} f(z_i)$$

$$= \prod (2\pi)^{-1/2} \exp \left\{ -\frac{z_i^2}{2} \right\}$$

$$= (2\pi)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} z_i^2 \right\}$$

$$= (2\pi)^{-\frac{1}{2}} \mathbf{z} \exp \left\{ -\frac{1}{2} \mathbf{z}^T \mathbf{z} \right\}$$

where $\mathbf{z}' = (z_1, \ldots, z_n)$ is a vector of independent random variables and $\mathbf{z}$ is the corresponding vector values.

6. $\chi^2$ (chi-squared), $t$ (student t), and $F$ (Fisher’s F) distributions

A] Relationship between chi-squared ($\chi^2$) and a normal distribution

The square of a standard normal variable has a $\chi^2$ distribution with 1 degree of freedom, that is if $Z \sim N(0, 1)$ then $Z^2 \sim \chi^2(1)$.

Furthermore, if $\chi^2(r_1)$, $\chi^2(r_2)$, $\ldots$, $\chi^2(r_n)$ are mutually stochastically independent $\chi^2$ random variables, then it can be shown that

© Giovanni Urga, 2009 xxx Adv Stats Induction, 2009-2010
0.1. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS: A REVIEW

\[ Y = \chi^2(r_1) + \chi^2(r_2) + \ldots + \chi^2(r_n) \sim \chi^2((r_1 + r_2 + \ldots + r_n)) \]

So, if \( Z_1, \ldots, Z_n \) are independent and standard normal random variables, then

\[ Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi^2(n) \]

B) Student - t statistics

Let \( Z \) be a \( N(0, 1) \) random variable and let \( \chi^2(r) \) denote a \( \chi^2 \) random variable with \( r \) degree of freedom. Furthermore, let \( Z \) and \( \chi^2(r) \) be stochastically independent. Then

\[ T = \frac{Z}{\sqrt{\chi^2(r)/r}} \sim t(r) \]

is a \( t \)-distribution with \( r \) degrees of freedom.

The p.d.f. of \( T \) is symmetric about 0, \( E(T) = 0 \), and is bell shaped but flatter than the \( N(0, 1) \) distribution. The variance of the random variable which is \( T \) is

\[ Var(T) = \frac{r}{r-2} \]

for \( r > 2 \).

As \( r \) approaches \( \infty \), the distribution of the random variable \( T \) can be closely approximated by that of a \( N(0, 1) \) random variable.

C) Fisher’s F statistics

Let \( \chi^2(r_1) \) and \( \chi^2(r_2) \) be independent \( \chi^2 \) random variable with \( r_1 \) and \( r_2 \) degrees of freedom. Then,

\[ F = \frac{\chi^2(r_1)/r_1}{\chi^2(r_2)/r_2} \sim F(r_1, r_2) \]

is a \( F \)-distribution with parameters \( r_1 \) and \( r_2 \) (called numerator and denominator degrees of freedom respectively). The \( F \)-distribution is skewed to the right and non negative since it is generated by two \( \chi^2 \) random variables.
0.2 Sampling Theory

OUTLINE:

• INTRODUCTION
  
  (Statistical model / inference: point vs interval estimates, hypothesis testing)

• METHODS FOR FINDING POINT ESTIMATORS
  
  (Method of moments (MM), maximum likelihood (ML), least squares (LS))

• PROPERTIES OF POINT ESTIMATORS
  
  - SMALL SAMPLE PROPERTIES OF ESTIMATORS
    (Unbiasedness, bias vs precision, efficiency)
  
  - LARGE SAMPLE PROPERTIES OF ESTIMATORS
    (consistency, convergence in distribution, asymptotic efficiency)

• INTERVAL ESTIMATION

• HYPOTHESIS TESTING
0.2. SAMPLING THEORY

0.2.1 Introduction

Statistical Inference is the subject that deals with the problems associated with the estimation of the unknown parameters underlying statistical distributions, measuring the precision, testing hypothesis on them, using them in generating forecasts of random variable, and so forth.

Because it is expensive to measure all the elements of a population (totality of elements about which some information is desired) underlying an experiment, an investigator often resorts to measuring attributes for a small portion of the population (known as a sample) and draws conclusion or makes policy decision based on the data obtained.

The probability distribution of the variables is unknown, and we wish to use the data to learn about the characteristics of that unknown distribution. If the result of an experiment can be described by a single number and the experiment is repeated $T$ times, the sample consists of the $T$ random variables $(Y_1,\ldots,Y_T) = Y'$. The joint probability density function (p.d.f.) of the sample $Y$ is assumed to have a known mathematical form $f(y \mid \vartheta)$ but depends on one or more parameters $(\vartheta_1,\ldots,\vartheta_k) = \vartheta'$ that are known to fall in a set of possible values, $\Omega$, called the parameter space. The investigator is always assumed to know the mathematical form of $f(y \mid \vartheta)$ and $\Omega$, the set of possible values of $\vartheta$.

This constitutes the statistical model for how the values of the random variables $Y$ are obtained and is the basis for classical statistical inference. Given the statistical model, statistical inference consists of using the sample values for $Y$ to specify plausible values for $\vartheta_1,\ldots,\vartheta_k$ (this is the problem of point estimation) or at least to determine a subset of $\Omega$, for which we can assert whether or not $\Omega$ contains some value $\vartheta$ (interval estimation or hypothesis testing).

Example:

We may assume that the amount of expenditure by persons with annual income of $20,000, from a certain population, follow a normal distribution, $N(\vartheta,\sigma^2)$, with unknown value of location parameter $\vartheta$ and the scale parameter $\sigma^2$. Parameter estimation then involves using a sample data on the experiments by persons with $20,000 income to make an inference about $\vartheta$ and $\sigma^2$.

INFEERENCE POINT ESTIMATES
INTERVAL ESTIMATES (sample of values)

HYPOTHESIS-TESTING: problem relates to using the data in a way to provide evidence about a conjecture covering the population.

Example (…continued)

We may entertain the notion that the mean expenditure level by households with $20,000 annual income is $15,000. Using the assumption that the probability distribution of expenditures is normal, we then examine the data, and

© Giovanni Urga, 2009  xxxiii  Adv Stats Induction, 2009-2010
make a judgement as to whether or not the data appear to be consistent with the
conjecture.

0.2.2 Methods for Finding Point Estimators

There are many methods that can be used in order to derive estimators. I will
briefly recall:
1) Method of Moments
2) Method of Maximum Likelihood
3) Method of Least Squares
[ 4) There is a Bayesian approach to estimation as well ]

What essentially is an estimator?

Let $Y$ be a random variable with p.d.f. $f(y|\theta)$, where $\theta$ is a parameter
that we would like to estimate. Let $Y_1,\ldots,Y_T$ be a random sample from this
population, then an estimator of $\theta$ is a function or rule of the form

$$\hat{\theta} = \hat{\theta}(Y_1,\ldots,Y_T)$$

i.e. the estimator $\hat{\theta}$ is a function of $Y_1,\ldots,Y_T$.

$\hat{\theta}$ is a random variable because it is a function of random variables. $\hat{\theta}$ means
that is an estimator, or rule for estimating, the parameter $\theta$. When the values
of the random variables are inserted, an estimate is produced, which is simply
the value of the random variable or estimator $\hat{\theta}$.

Example (..continued):
Expenditure by persons with annual income of $20,000 are approximately
usually distributed with an unknown mean $\theta$ but known variance, say $\sigma^2 =
$5,000$^2$. That is, if $Y$ is the expenditure from a $20,000 income by a person in
this population, then

$$Y \sim N(\theta, \sigma^2 =$5,000$^2)$$

© Giovanni Urga, 2009
A common estimator of \( \vartheta \) is the arithmetic mean

\[
\hat{\vartheta} = \vartheta(Y_1, ..., Y_T) = \frac{1}{T} \sum_{i=1}^{T} Y_i = \bar{Y}
\]

So, where did the idea come from that the arithmetic mean of a sample of values can or should be used to estimate a population mean?

This is the subject of examining methods that may be used to obtain point estimators.

1) METHOD OF MOMENTS

We defined the \( r^{th} \) moment of a random variable \( Y \) about the origin as \( \mu_r = E[Y^r] \).

If the p.d.f. of \( Y \) is \( f(y, \vartheta) \) where \( \vartheta' = (\vartheta_1, \ldots, \vartheta_k) \) is a vector of unknown parameters, then in general \( \mu_r \) will be a known function of \( \vartheta \), say \( \mu_r = \mu_r(\vartheta) \).

The idea of the method of moments is to use a random sample of data, \( Y_1, \ldots, Y_T \), to compute the sample moments

\[
\hat{\mu}_r = \frac{1}{T} \sum_{i=1}^{T} Y_i^r / T = \bar{Y}^r
\]

and then to equate the sample and the true moments \( \hat{\mu}_r = \mu_r(\vartheta) \) and solve the resulting system of \( K \) equations (if possible) for the unknown parameters. The resulting estimator, \( \hat{\vartheta}_{MM} \), is the method of moments estimator (MM).

**Example:**

Let \( Y_1, \ldots, Y_T \) be a random sample from a \( N(\vartheta, \sigma^2) \) population. Recall that \( E(Y) = \vartheta = \mu_1 \) and \( \sigma^2 = E(Y^2) - (E[Y])^2 = \mu_2 - (\mu_1)^2 \). Then, equating the sample to population moments,

\[
\hat{\mu}_1 = \frac{1}{T} \sum_{i=1}^{T} Y_i / T = \bar{Y} = \hat{\vartheta}
\]

\[
\hat{\mu}_2 = \frac{1}{T} \sum_{i=1}^{T} Y_i^2 / T = \hat{\sigma}^2 + \hat{\vartheta}^2
\]

so
\[
\hat{\sigma}^2 = \left( \frac{\sum_{i=1}^{T} Y_i^2}{T} \right) - \bar{Y}^2
\]
\[
= \frac{1}{T} \sum_{i=1}^{T} (Y_i - \bar{Y})^2
\]

**Note:** The MM estimators need not to be unique, and the method can be difficult to apply in with complicated problems. It also depends on the random variable in question having moments, which is not always the case.

In fact, a more powerful and general estimation method is the Maximum Likelihood Estimation Method (ML) that we introduce in the following section.

2) THE METHOD OF MAXIMUM LIKELIHOOD

Let \( Y \) be a Bernoulli random variable with parameter \( p \), which we know, for some reason, can only take the values \( \frac{1}{4} \) or \( \frac{3}{4} \). Suppose we have a random sample of size \( T = 3 \) with values \( y_1 = 1 \), \( y_2 = 1 \), \( y_3 = 0 \).

The question is how to use these data to estimate the unknown parameter \( p \), which for this problem boils down to choosing either \( p = \frac{1}{4} \) or \( p = \frac{3}{4} \).

The method of maximum likelihood (ML) will choose that value of the unknown parameter \( p \) that maximizes the probability (likelihood) of randomly drawing the sample that was actually obtained.

Since \( f(y|p) = p^y(1-p)^{1-y} \) for \( y = 0 \) or 1, we can calculate the probability of our random sample from the point p.d.f. of \( y_1, y_2, y_3 \). It is

\[
f(y_1 = 1, y_2 = 2, y_3 = 0) = f(1, 1, 0) = \prod_{i=1}^{3} p^{y_i}(1-p)^{1-y_i} = pp(1-p)
\]

So we can interpret this function as a function of \( p \) given the sample observations. When we do so, it is called a likelihood function and written \( l(p|y) \).

**Note:** This function is mathematically identical to the joint p.d.f. of the random sample, but it is interpreted as function of the unknown parameters instead of the value of the random variable which are assumed to be known.

To making inferences or decisions about \( p \) after the sample values are observed, all relevant sample information is contained in the likelihood function.

© Giovanni Urga, 2009  xxxvi  Adv Stats Induction, 2009-2010
Then
\[ l\left(\frac{1}{4}/y\right) = 0.046 \]
\[ l\left(\frac{3}{4}/y\right) = 0.141 \]

thus the probability (or likelihood) of obtaining the sample values we actually have is maximized by choosing \( \tilde{p} = \frac{3}{4} \) rather than \( \tilde{p} = \frac{1}{4} \) (these are the only choices), and thus \( \tilde{p} = \frac{3}{4} \) is the maximum likelihood estimate of \( p \) in this problem. It is, in the sense described, the value of \( p \) most likely to have generated our sample.

Alternatively, we can find the value of \( p \) that maximizes \( l(p/y) \) by using calculus techniques.

A local maximum of a continuous function occurs where its slope is zero and its second derivative is negative.

Here, suppose \( l(y/p) = p^2(1 - p) \)

\[
\frac{dl(p/y)}{dp} = 2p - 3p^2
\]
\[
\frac{d^2l(p/y)}{dp^2} = 2 - 6p
\]

Setting the first derivative to zero and solving, we see there are two solutions, \( \tilde{p} = 0 \) and \( \tilde{p} = 2/3 \).

\[ \tilde{p} = 0 \] is ruled out since it clearly does not maximize the likelihood function: if \( \tilde{p} = 0 \) we could not have drawn the sample since \( P(Y = 1) = p \) and thus obtaining a value \( y = 1 \) would be impossible.

© Giovanni Urga, 2009
\( \tilde{p} = 2/3 \) satisfies both the first and second order conditions for a local maximum and is the maximum likelihood estimate of \( p \).

To sum up the maximum likelihood: the likelihood function \( l(\vartheta|y) \) is algebraically identical to the joint p.d.f. \( f(y|\vartheta) \) of the random sample \( Y_1, \ldots, Y_T \), where \( \vartheta \) is a vector of unknown parameters with a true value known to fall in the parameter space \( \Omega \). The difference between \( l(\vartheta|y) \) and \( f(y|\vartheta) \) is that \( l(\vartheta|y) \) is interpreted as a function of \( \vartheta \) given the values of a random sample \( y \); and the joint p.d.f. \( f(y|\vartheta) \) is a function of \( y \) given particular values of \( \vartheta \).

The maximum likelihood of \( \vartheta \) is that value in \( \Omega \), say \( \hat{\vartheta}_{ML} \), that maximizes \( l(\vartheta|y) \). The value \( \hat{\vartheta} \) is in general a function of \( y \), say \( \vartheta = \vartheta(y) \), so the random variable \( \hat{\vartheta} = \hat{\vartheta}(Y) \), where \( Y = (Y_1, \ldots, Y_T) \), is the maximum likelihood estimator of \( \vartheta \).

Practical matter: it is usual practice to maximize the natural logarithm of the likelihood function rather than the likelihood function itself. This is very convenient, since the logarithm likelihood \( L = \ln l(\vartheta|y) \) is composed of sums rather than products and exponential functions simplify nicely.

Since the natural logarithm is a monotonic function, \( \ln l(\vartheta|y) \) and \( l(\vartheta|y) \) attain their maximum at the same value of \( \vartheta \).
3) LEAST SQUARES ESTIMATION

Note: Both MM and ML require a specific assumption about the distribution of the random variable. LS does not require exact specification of the population distribution.

It can be used to estimate central moments of random variable $\mu_r = E[Y^r]$.

Idea: since the mathematical expectations of a random variable is the mean of the random variable, given the values of a random sample of data $Y_1, \ldots, Y_T$, it is reasonable to use the “center” of the data $y_{ri}, i = 1, \ldots, T$ to estimate $\mu_r$.

One way to define the center of a set of data is to find the value $\hat{\mu}_r$ that minimizes

$$S = \sum_{i=1}^{T} (y_{ri} - \mu_r)^2$$

The value $S$ is the sum of squared differences between $y_{ri}$ and the expectation $\mu_r = E[Y^r]$. The value $\hat{\mu}_r$ that minimizes $S$ for a given set of data values of the random variable is called the LEAST SQUARED ESTIMATE of $\mu_r$. If $\hat{\mu}_r$ is considered a function of the random variables $Y_i$ then it is the least squared (LS) estimator.

Example: Let $Y_1, \ldots, Y_T$ be a random sample from a population with mean $\bar{Y}$ and finite variance $\sigma^2$. Then the LS estimator of $\beta$ is obtained by minimizing

$$S = \sum_{i=1}^{T} (y_i - \beta)^2$$

The F.O.C. is

$$\frac{dS}{d\beta} = \sum_{i=1}^{T} -2(y_i - \hat{\beta}) = 0$$

So

$$\hat{\beta} = \frac{1}{T} \sum_{i=1}^{T} y_i$$

Note: We may verify that $\hat{\beta}$ does in fact minimize $S$ by checking the S.O.C.

Graphically the sum of squares $S$ may be viewed as a parabola.
Finding the LS estimator of $\alpha$ amounts to find the value of $\hat{\beta}$ that corresponds to the minimum point on the parabola.

0.2.3 Properties of Point Estimators

Estimators are random variables so evaluating their properties amounts to studying the properties of their probability distributions.

We will consider:
A) “SMALL SAMPLE” properties of estimators (those properties hold for samples of all sizes, even small ones)
B) “LARGE SAMPLE” or “ASYMPTOTIC” properties of estimators (those properties are based on sample whose size is assumed to grow to infinity).

These properties are frequently required in econometrics as estimation rules’ properties such that very little can be said about them without the assumption of very large samples.

Small Sample Properties of Estimators

I. SINGLE PARAMETER CASE

$Y$ is a random variable with p.d.f. $f(y|\theta)$, where $\theta = \text{parameter}$ we would like to estimate and let $Y_1, \ldots, Y_T$ be a sample of observations on $Y$.

Suppose $\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_T)$ is an estimator of $\theta$.

We need to evaluate the estimator $\hat{\theta}$ in a repeated sampling sense, since to do otherwise requires knowledge of the true parameter $\theta$.  

© Giovanni Urga, 2009
0.2. SAMPLING THEORY

Unbiasedness

\( \hat{\vartheta} \) is an unbiased estimator of \( \vartheta \) if

\[
E[\hat{\vartheta}] = \vartheta
\]

If \( E[\hat{\vartheta}] \neq \vartheta \), then \( \hat{\vartheta} \) is a biased estimator and \( E[\hat{\vartheta}] - \vartheta = \delta \) is the amount of the bias.

If the p.d.f. is symmetric and \( \hat{\vartheta} \) is unbiased, the \( \vartheta \) is located at the center of the distribution.

Note: Unbiasedness means that the estimator “on the average” will yield the true parameter value, i.e. if the underlying experiment is repeated infinitely many times by drawing sample of site \( T \), the average value of the estimates \( \hat{\vartheta} \) from all those samples will be \( \vartheta \).

Now, \( \hat{\vartheta} \) is a random variable, so in addition to the mean of the estimator we are also concerned about its variance. Suppose we have two unbiased estimators, \( \tilde{\vartheta} \) and \( \hat{\vartheta} \):
and $\text{Var}(\hat{\vartheta}) > \text{Var}(\tilde{\vartheta})$. Since both estimators are unbiased, we would prefer the estimator $\tilde{\vartheta}$ to $\hat{\vartheta}$ since using $\tilde{\vartheta}$ gives us a higher probability of obtaining an estimate that is close to the true value.

**Bias vs Precision**

We want an estimator that usually yields estimates of the unknown parameter that are “close” to the true parameter value. In order to do this, bias and variance must be evaluated. One way to do this is to consider an estimator’s mean square error (MSE):

$$\text{MSE}(\hat{\vartheta}) = E[(\hat{\vartheta} - \vartheta)^2]$$

It is, within the repeated sampling contest, the average squared distance between $\hat{\vartheta}$ and the true parameter value $\vartheta$.

Furthermore: $(\hat{\vartheta} - \vartheta)^2$ measures the loss form using $\hat{\vartheta}$ as an estimator of $\vartheta$ and taking the mathematical expectation of this loss yields the average loss or risk of using $\hat{\vartheta}$ to estimate $\vartheta$.

This measure encompasses both precision and bias as follows:

\[
\text{MSE}(\hat{\vartheta}) = E[(\hat{\vartheta} - E(\hat{\vartheta})) + E(\hat{\vartheta}) - \vartheta]^2 \\
= E[(\hat{\vartheta} - E(\hat{\vartheta})) + E(\hat{\vartheta}) - \vartheta]^2 \\
= E[(\hat{\vartheta} - E(\hat{\vartheta}))^2 + [E(\hat{\vartheta}) - \vartheta]^2 + 2E(\hat{\vartheta}) - E(\hat{\vartheta})][E(\hat{\vartheta}) - \vartheta]
\]

where

\[
E[(\hat{\vartheta} - E(\hat{\vartheta}))][E(\hat{\vartheta}) - \vartheta] = [E(\hat{\vartheta}) - \vartheta][E(\hat{\vartheta}) - E(\hat{\vartheta})] = [E(\hat{\vartheta}) - \vartheta][E(\hat{\vartheta}) - E(\hat{\vartheta})] = 0
\]
0.2. SAMPLING THEORY

So

\[
MSE(\hat{\vartheta}) = E[\hat{\vartheta} - E(\hat{\vartheta})]^2 + [E(\hat{\vartheta}) - \vartheta]^2 \\
= \text{var}(\hat{\vartheta}) + [\text{bias}(\hat{\vartheta})]^2
\]

The use of MSE allows us to compare estimators such as \( \tilde{\vartheta} \) and \( \hat{\vartheta} \) whose p.d.f. are shown:

\[\begin{align*}
\tilde{\vartheta} \text{ is an estimator unbiased whose variance is larger than that of the biased estimator } \hat{\vartheta}. \text{ For the purpose of using an estimator that will give estimates close to } \vartheta, \text{ } \tilde{\vartheta} \text{ may be preferred even though it is biased, if it yields a smaller mean square error.}
\end{align*}\]

Efficiency

If we compare estimators that are unbiased, then the estimator with the smaller variance would be preferred. [Basis of the estimator efficiency]

An estimator \( \hat{\vartheta} \) is an efficient estimator of \( \vartheta \) if \( E[\hat{\vartheta}] = \vartheta \) and \( \text{var}(\hat{\vartheta}) \leq \text{var}(\tilde{\vartheta}) \), where \( \tilde{\vartheta} \) is any other unbiased estimator of \( \vartheta \).

The CRAMER-RAO INEQUALITY provides a sufficient but not necessary condition for an estimator to be efficient.

Let \( Y \) be a random variable with p.d.f. \( f(y|\vartheta) \). If \( Y_1, \ldots, Y_T \) is a random sample, then the joint p.d.f. of the \( T \) random variables \( Y_i \) is
Recall that if we interpret the joint p.d.f. \( f(y_1, \ldots, y_T | \vartheta) \) as a function of the unknown parameter \( \vartheta \) given the values of the random sample, then the resulting function is known as the likelihood function and written as \( l(\vartheta | y) = l(\vartheta | y_1, \ldots, y_T) \). It is mathematically identical to the joint p.d.f. but has a different interpretation.

Let \( L(\vartheta) \) denote the natural logarithm of the likelihood function, then the Cramer-Rao inequality states that if \( \hat{\vartheta} \) is any unbiased estimator of \( \vartheta \), and certain regularity conditions hold, then

\[
\text{Var}(\hat{\vartheta}) \geq \frac{1}{-E\left[\frac{d^2 L(\vartheta)}{d\vartheta^2}\right]}
\]

(Theil (1971), pp. 384-386 for proof)

It says that if an unbiased estimator \( \hat{\vartheta} \) has a variance equal to \(-\{1/E[d^2 L/d\vartheta^2]\}\), which is known as the Cramer-Rao lower bound (CRLB), then it is efficient since no lower variance is possible for an unbiased estimator.

N.B.: This condition is sufficient but not necessary. It is possible that no unbiased estimator has variance as small as the CRLB. It is often much simpler to find the minimum variance estimator among those that are not only unbiased but also linear function of the sample observations.

Let \( \hat{\vartheta} \) be an estimator of an unknown parameter \( \vartheta \) and be of the form

\[
\hat{\vartheta} = a_1 y_1 + \cdots + a_T y_T
\]

where \( a_i \) are constants.

Then \( \hat{\vartheta} \) is defined to be a linear estimator since it is a linear function of the sample observations. If \( \hat{\vartheta} \) is unbiased and \( \text{var}(\hat{\vartheta}) \leq \text{var}(\vartheta) \), where \( \vartheta \) is any other linear and unbiased estimator of \( \vartheta \), then \( \hat{\vartheta} \) is the BEST LINEAR UNBIASED ESTIMATOR (BLUE) of \( \vartheta \).

Note: It is often possible to find such an estimator without knowledge of the underlying p.d.f. of \( Y \).

II. THE CASE OF SEVERAL PARAMETERS

© Giovanni Urga, 2009  xliiv  Adv Stats Induction, 2009-2010
0.2. SAMPLING THEORY

This is the case of statistical model involving unknown parameters: $\vartheta_1, \ldots, \vartheta_k$.

Let $Y$ be a random variable with p.d.f. $f(y|\vartheta)$ with $\vartheta = (\vartheta_1, \ldots, \vartheta_k)'$ is a $(k \times 1)$ vector of unknown parameters that we wish to estimate. Let $\hat{\vartheta} = (\hat{\vartheta}_1, \ldots, \hat{\vartheta}_r)'$ be an $(k \times 1)$ vector. $\hat{\vartheta}$ is the estimator of the element unknown $\vartheta_i$ of $\vartheta$.

The three properties will be introduced following the structure below:

Unbiasedness

Consider $\vartheta$ $(k \times 1)$ such that $E[\hat{\vartheta}] = \vartheta$, that is, the vector $\hat{\vartheta}$ is an unbiased estimator for $\vartheta$:

$$E(\hat{\vartheta}) = \begin{bmatrix} E(\hat{\vartheta}_1) \\ E(\hat{\vartheta}_2) \\ \vdots \\ E(\hat{\vartheta}_k) \end{bmatrix} = \begin{bmatrix} \vartheta_1 \\ \vartheta_2 \\ \vdots \\ \vartheta_k \end{bmatrix} = \vartheta$$

If $E[\hat{\vartheta}] \neq \vartheta$, then $\hat{\vartheta}$ is biased and $\delta = E[\hat{\vartheta}] - \vartheta$ is the bias vector. The precision of $\hat{\vartheta}$ is measured by its variance-covariance matrix (a $k \times k$ positive definite and symmetric), which is defined as

$$Cov(\hat{\vartheta}) = E[(\hat{\vartheta} - E(\hat{\vartheta}))(\hat{\vartheta} - E(\hat{\vartheta}))']$$

$$= \begin{bmatrix} \text{var}(\hat{\vartheta}_1) & \text{cov}(\hat{\vartheta}_1, \hat{\vartheta}_2) & \cdots & \text{cov}(\hat{\vartheta}_1, \hat{\vartheta}_k) \\ \text{cov}(\hat{\vartheta}_2, \hat{\vartheta}_1) & \text{var}(\hat{\vartheta}_2) & \cdots & \text{cov}(\hat{\vartheta}_2, \hat{\vartheta}_2) \\ \vdots & \vdots & & \vdots \\ \text{cov}(\hat{\vartheta}_k, \hat{\vartheta}_1) & \text{cov}(\hat{\vartheta}_k, \hat{\vartheta}_2) & \cdots & \text{var}(\hat{\vartheta}_k) \end{bmatrix}$$

Note: An obvious function arises as to how the precision of two competing estimators of $\vartheta$ can be compared. That is, if $\hat{\vartheta}$ and $\tilde{\vartheta}$ are both estimators of $\vartheta$, how do we compare $\text{cov}(\hat{\vartheta})$ and $\text{cov}(\tilde{\vartheta})$ since they are matrices?

We will say that $\text{cov}(\hat{\vartheta})$ is “smaller than” $\text{cov}(\tilde{\vartheta})$, in a matrix sense, if $CC = \text{cov}(\tilde{\vartheta}) - \text{cov}(\hat{\vartheta})$ is a positive semi-definite matrix. So, if $CC$ is a positive semi-definite matrix then

(i) $\text{var}(\tilde{\vartheta}_i) \geq \text{var}(\hat{\vartheta}_i), \quad i = 1, \ldots, k$

(ii) $|\text{cov}(\tilde{\vartheta})| \geq |\text{cov}(\hat{\vartheta})|$

(iii) If $c$ is any $(k \times 1)$ vector of constants, then $\text{var}(c'\tilde{\vartheta}) = c'(\text{cov}(\tilde{\vartheta})) c \geq \text{var}(c'\hat{\vartheta}) = c'(\text{cov}(\hat{\vartheta})) c$

(see Goldberger, 1991)

© Giovanni Urga, 2009 xlv Adv Stats Induction, 2009-2010
Bias vs Precision

The mean squared error matrix of an estimator $\hat{\vartheta}$ for $\vartheta$

$$MSE(\hat{\vartheta}) = E[(\hat{\vartheta} - \vartheta)(\hat{\vartheta} - \vartheta)']$$

thus

$$MSE(\hat{\vartheta}) = E[\hat{\vartheta} - \vartheta][\hat{\vartheta} - \vartheta] + [E(\hat{\vartheta}) - \vartheta][E(\hat{\vartheta}) - \vartheta]$$

A common scalar measure of the mean squared error of an estimator $\hat{\vartheta}$ is

$$tr\{MSE(\hat{\vartheta})\} = \sum_{i=1}^{K}\{\text{var}(\hat{\vartheta}_i) + [\text{bias}(\hat{\vartheta}_i)]^2\}$$

And is often simply called the mean squared error of $\hat{\vartheta}$ or the average loss or risk incurred when using $\hat{\vartheta}$ to estimate $\vartheta$.

Efficiency

An estimator $\hat{\vartheta}$ is efficient for $\vartheta$ if $\hat{\vartheta}$ is unbiased and $\text{cov}(\tilde{\vartheta}) - \text{cov}(\hat{\vartheta})$ is positive semidefinite where $\tilde{\vartheta}$ is any other unbiased estimator for $\vartheta$.

We can find a multivariate version of the Cramer-Rao inequality

If we consider the likelihood function $l(\vartheta|y)$, which is mathematically equivalent to the joint p.d.f. of the random sample but is a function of $\vartheta$ instead of $y$.

If $L(\vartheta)$ denote the logarithm of the likelihood function then the matrix

$$I(\vartheta) = -E\left[\frac{\partial^2 L}{\partial \vartheta \partial \vartheta'}\right]$$

where $I(\vartheta)$ is the INFORMATION MATRIX and

$$[I(\vartheta)]^{-1}$$

is the Cramer-Rao lower bound matrix.
Large Sample Properties of Estimators

There are estimations when the small sample properties of estimators cannot be obtained easily, so in order to compare estimators we must study their asymptotic properties, i.e. their approximate behavior when the sample size $T$ is large and approaches infinity. There are three important definitions:

a.) CONSISTENCY

b.) CONVERGENCE IN DISTRIBUTION

c.) ASYMPTOTIC EFFICIENCY

CONSISTENCY

This property assumes that the estimator rule will produce an estimate that is close to the true parameter value with high probability if the sample size is large enough. More precisely, let $\hat{\vartheta}_T$ be an estimator of $\vartheta$ based on a sample of size $T$. Then $\hat{\vartheta}_T$ is a consistent estimator of $\vartheta$ if

$$\lim_\limits{T \to \infty} P\left( |\hat{\vartheta}_T - \vartheta| < \epsilon \right) = 1$$

where $\epsilon$ is an arbitrary small positive number.

This means that the probability that the value of $\hat{\vartheta}_T$ falls in the interval $[\vartheta - \epsilon, \vartheta + \epsilon]$ can be made arbitrary close to 1 given a large enough sample size, no matter how small $\epsilon$ is.

If $\lim_\limits{T \to \infty} P\left( |\hat{\vartheta}_T - \vartheta| < \epsilon \right) = 1$ is true, then the sequence of random variables $\hat{\vartheta}_T$ is said to converge in probability to the constant $\vartheta$ and $\vartheta$ is said to be the probability limit of the sequence $\hat{\vartheta}_T$. This is usually abbreviated as $\text{plim} \, \hat{\vartheta}_T = \vartheta$.

Thus the estimator $\hat{\vartheta}_T$ is consistent for $\vartheta$ if $\lim_\limits{T \to \infty} P\left( |\hat{\vartheta}_T - \vartheta| < \epsilon \right) = 1$ is true or, equivalently, when $\text{plim} \, \hat{\vartheta}_T = \vartheta$.

Example:

Let $Y_1, \ldots, Y_T$ be a random sample from a $N(\beta, \sigma^2)$ population, and consider the estimator of $\beta$,

$$\hat{\beta}_T = \sum_{i=1}^{T} y_i / T$$

which is the sample mean.
We know that

\[ \hat{\beta}_T \sim N(\beta, \sigma^2/T) \]

Suppose we have three sample size \( T_3 > T_2 > T_1 \), the sampling distributions are

We see that the probability mass in the interval \([\hat{\beta} - \epsilon, \hat{\beta} + \epsilon]\) is getting large as the sample size increases from \( T_1 \) to \( T_2 \) to \( T_3 \), and the sampling distribution of \( \hat{\beta}_T \) is collapsing about the true parameter \( \hat{\beta} \). In fact, since

\[ \lim_{T \to \infty} \text{var}(\hat{\beta}_T) = \lim_{T \to \infty} \frac{\sigma^2}{T} = 0 \]

The sampling distribution of \( \hat{\beta}_T \) becomes degenerate at the true parameter value in the limit, so that all its probability mass occurs at \( \beta \).

So we can conclude that, sufficient, but not necessary, conditions for an estimator \( \hat{\theta}_T \) to be consistent for \( \theta \) are

- \( \lim_{T \to \infty} [\hat{\theta}_T] = \theta \quad \text{(asymptotically unbiased)} \)
- \( \lim_{T \to \infty} \text{var}(\hat{\theta}_T) = 0 \)

CONVERGENCE IN DISTRIBUTION

© Giovanni Urga, 2009 xlviii Adv Stats Induction, 2009-2010
In this case we study the probability distribution of an estimator as the sample size becomes increasingly large. It is remarkable that estimator whose distributions are unknown in small samples can sometimes be shown to have a particular, known distribution in large samples.

A well-known example of this is the **CENTRAL LIMIT THEOREM**:

Let $Y_1, \ldots, Y_T$ be independently and identically distributed random variables with $E[Y_i] = \beta$ and $Var[Y_i] = \sigma^2$.

Also let $\hat{\beta}_T = \frac{1}{T} \sum_{i=1}^{T} y_i$. Then,

$$\sqrt{T} \left( \hat{\beta}_T - \beta \right) \xrightarrow{d} N(0,\sigma^2)$$

This remarkable result says that given a random sample of observation from an infinite population with any distribution as long as it has finite mean and finite variance then a simple function of the sample mean

$$\sqrt{T} \left( \hat{\beta}_T - \beta \right)$$

has a limiting distribution that is $N(0,\sigma^2)$

**Comments**

1. There are situations when we have to make probability statements about an estimator when the exact small sample distribution of the estimator cannot be determined. If a limiting distribution exists, then we can consider that limiting distribution as an approximation to the true but unknown distribution if the sample we have at our disposal is sufficiently large for us to believe the approximation will be a reasonable one.

Suppose that the situations stated by CLT hold, the exact p.d.f. of $\hat{\beta}_T$ is unknown, then if the sample size is large, we may take $\sqrt{T} \left( \hat{\beta}_T - \beta \right)$ to be approximately normally distributed $(0,\sigma^2)$. And if $\sqrt{T} \left( \hat{\beta}_T - \beta \right) \sim N(0,\sigma^2)$ then $\hat{\beta}_T \sim N(\beta, \sigma^2/T)$ where $\sim$ means “approximately distributed”. Many times this result is written as $\hat{\beta}_T \xrightarrow{asy} N(\beta, \sigma^2/T)$ and $N(\beta, \sigma^2/T)$ is called the **ASYMPTOTIC DISTRIBUTION** of $\hat{\beta}_T$.

2. Why the CLT is stated in terms of $\sqrt{T} \left( \hat{\beta}_T - \beta \right)$?
We known that when the conditions of CLT hold $\hat{\beta}_T \sim N(\beta, \sigma^2/T)$ for any sample size $T$. So, as $T \to \infty$, $\text{var}(\hat{\beta}_T) \to 0$, and the distribution of $\hat{\beta}_T$ becomes degenerate, a definitely non-normal distribution! However, it can be shown that the sequence of random variables $\sqrt{T}(\hat{\beta}_T - \beta)$, as $T \to \infty$, has a sequence of c.d.f.'s $F_T$ that converge, pointwise, to the c.d.f. $F$ of a $N(0, \sigma^2)$ random variable. That is, $\lim_{T \to \infty} F_T \to F$ at all continuity points of $F$.

This is in fact a definition of the convergence in distribution of the sequence of random variables $Z_T = \sqrt{T}(\hat{\beta}_T - \beta)$, and $Z_T$ is said to have the limiting distribution $F$. ⇒ All what has been said can be applied to the case of vectors of estimators if $\hat{\vartheta}_T$ is a consistent estimator for $\vartheta$ and $\sqrt{T}(\hat{\vartheta}_T - \vartheta)$ converges in distribution to $N(0, \Sigma)$, then $\hat{\vartheta}_T$ is said to have the asymptotic distribution $N(\vartheta, \Sigma/T)$.

ASYMPTOTIC EFFICIENCY

One-parameter case
Suppose we have two estimators $\hat{\vartheta}$ and $\tilde{\vartheta}$ of a parameter $\vartheta$ such that

$$\sqrt{T} \left( \hat{\vartheta} - \vartheta \right) \overset{d}{\rightarrow} N(0, \sigma_1^2)$$

and

$$\sqrt{T} \left( \tilde{\vartheta} - \vartheta \right) \overset{d}{\rightarrow} N(0, \sigma_2^2)$$

If $\sigma_2^2 \geq \sigma_1^2$, then $\hat{\vartheta}$ is asymptotically efficient relative to $\tilde{\vartheta}$.

Scalar case
If $\vartheta$ is a vector of parameters and $\hat{\vartheta}$ and $\tilde{\vartheta}$ are consistent estimators such that

$$\sqrt{T} \left( \hat{\vartheta} - \vartheta \right) \overset{d}{\rightarrow} N(0, \Sigma)$$

and

$$\sqrt{T} \left( \tilde{\vartheta} - \vartheta \right) \overset{d}{\rightarrow} N(0, \Omega)$$

then $\hat{\vartheta}$ is asymptotically efficient relative to $\tilde{\vartheta}$ if $(\Omega - \Sigma)$ is positive semidefinite matrix.

© Giovanni Urga, 2009
0.2. SAMPLING THEORY

As in the finite sampling efficiency, showing that an estimator is asymptotically efficient relative to any other consistent estimator, requires knowledge of the parent distribution, so that the information matrix and the asymptotic Cramer-Rao lower bound can be established. Let \( \hat{\vartheta} \) be a consistent estimator of \( \vartheta \) such that

\[
\sqrt{T} \left( \hat{\vartheta} - \vartheta \right) \overset{d}{\to} N \left( 0, \Sigma \right)
\]

Then the estimator \( \hat{\vartheta} \) is asymptotically efficient if

\[
\Sigma = \lim_{T \to \infty} \left[ \frac{1}{T} I(\vartheta) \right]^{-1}
\]

where \( I(\vartheta) \) is the information matrix.

So both small sample and asymptotic efficiency are established using CRLB as a reference point.

**Note:** A property of the "method of maximum likelihood".

Under some fairly general conditions maximum likelihood estimators are consistent, asymptotically normal, asymptotically unbiased, and asymptotically efficient.

That is, if \( \hat{\vartheta} \) is the maximum likelihood estimator of the vector of parameter \( \vartheta \) then

\[
\sqrt{T} \left( \hat{\vartheta} - \vartheta \right) \overset{d}{\to} N \left( 0, \lim_{T \to \infty} \left[ \frac{1}{T} I(\vartheta) \right]^{-1} \right)
\]

Consequently, in a problem like the one introduced earlier, where we obtain the MLE of the parameter \( p \) of a Bernoulli population, even though we do not know the small sample distribution of the estimator, we can rely on the fact that it is asymptotically normal to test hypotheses or make confidence interval statements.
0.2.4 Interval estimation and hypothesis testing

Interval estimation

Once we get point estimates, we are concerned about the “reliability” of the estimate. Knowledge of the variance of the estimator (or an estimate of it) plus knowledge of the finite or asymptotic sampling distribution conveys this information.

It is sometimes useful, however, when reporting results to give a range of possible values that incorporates both a point estimate and a measure of precision. An INTERVAL ESTIMATE does just that. Interval estimates are usually based on the sampling distribution of an estimator, although Chebychev’s inequality could also be used. We will illustrate using three common examples.

Example:
Let \( Y_1, \ldots, Y_T \) be the known sample from a \( N(\beta, \sigma^2) \) population, where \( \sigma^2 \) is known. The MLE. of \( \beta \) is

\[
\hat{\beta} = \frac{1}{T} \sum_{i=1}^{T} y_i \sim N(\beta, \sigma^2/T)
\]

Then we can use the sampling distribution of \( \hat{\beta} \) to make probability statements. Since

\[
z = \frac{\hat{\beta} - \beta}{\sigma/\sqrt{T}} \sim N(0,1)
\]

it follows that

\[
P[-Z_{(\alpha/2)} \leq Z \leq Z_{(\alpha/2)}] = 1 - \alpha
\]

where \( Z_{(\alpha/2)} \) is the upper \(-\alpha/2\) percentile of the \( N(0,1) \) distribution, as illustrated in the figure below and the values are tabulated.
0.2. SAMPLING THEORY

Substituting for $Z$ and rearranging

$$P[-Z_{(\alpha/2)} \leq Z \leq Z_{(\alpha/2)}] = 1 - \alpha$$

$$P[-Z_{(\alpha/2)} \leq \hat{\beta} - \beta / \sigma / \sqrt{T} \leq Z_{(\alpha/2)}] = 1 - \alpha$$

$$P[\hat{\beta} - Z_{(\alpha/2)} (\sigma / \sqrt{T}) \leq \hat{a} \leq \hat{\beta} + Z_{(\alpha/2)} (\sigma / \sqrt{T})] = 1 - \alpha$$

The end points of the interval containing $\beta$ are random since $\hat{\beta}$ is random. The random interval $[\hat{\beta} - Z_{(\alpha/2)} (\sigma / \sqrt{T}), \hat{\beta} + Z_{(\alpha/2)} (\sigma / \sqrt{T})]$ is an interval estimator of $\beta$ and contains the unknown parameter $\beta$ with $(1 - \alpha)$ confidence interval, and $\alpha$ is the significant level for the interval estimator. An interval estimate is obtained when the estimator $\hat{\beta}$ is replaced by an estimate based on a particular sample of values.

**Example**: referring back to the previous example, an interval estimate of the mean consumption level with confidence level $1 - \alpha = 0.95$ is

$$£13,000 \pm (1.96)(500 / \sqrt{3})$$

or

$$[£7,341.97, 18,658.03]$$

So, we can state that $[£7,341.97, 18,658.03]$ is a 95% confidence interval for $\beta$, that is that it is one realisation of the interval estimator

$$P[\hat{\beta} - Z_{(\alpha/2)} (\sigma / \sqrt{T}) \leq \hat{a} \leq \hat{\beta} + Z_{(\alpha/2)} (\sigma / \sqrt{T})] = 1 - \alpha$$

© Giovanni Urga, 2009
Hypothesis Testing

This is the second branch of statistics inference. We’ll look at various properties of statistical tests and then at procedures or principles by which test statistics can be obtained.

Elements of a Statistical Test

A statistical test is a decision problem involving an unknown parameter \( \theta \) that must lie in a certain parameter space \( \Omega \). However, \( \Omega \) can be divided into two disjoint subsets \( \Omega_0 \) and \( \Omega_1 \), and then we must decide, using a sample of data, whether \( \theta \) lies in \( \Omega_0 \) or \( \Omega_1 \).

Let \( H_0 \) denote the null hypothesis that \( \theta \in \Omega_0 \).
Let \( H_1 \) denote the alternative hypothesis that \( \theta \in \Omega_1 \).

Since \( \Omega_0 \) and \( \Omega_1 \) are disjoint, only one of the hypotheses \( H_0 \) and \( H_1 \) is true, and by accepting one we are at the same time rejecting the other.

Testing a hypothesis then involves accepting either \( H_0 \) or \( H_1 \), in light of the costs of making an incorrect decision and using whatever data are available as efficiently as possible.

Suppose we observe a random sample \( Y = (Y_1, \ldots, Y_T)' \) that has joint p.d.f. \( f(y|\theta) \). The set of all possible value of \( \theta \) is the sample space of the experiment.

A test procedure is defined by dividing the sample space into two parts, one containing the subset of values of \( Y \) that will lead to \( H_0 \) being accepted, and the other being the subset of values of \( Y \) where \( H_1 \) will be accepted (and \( H_0 \) rejected). The latter subset where \( H_0 \) will be rejected is called the critical region or the rejection region of the test.

In practice, a test is carried out using a test statistic. A test statistic is one that has known distribution under the null hypothesis (i.e. ensuring \( H_0 \) is true) and has some other distribution when \( H_1 \) is true. Thus the critical region of a hypothesis test is that set of values of the test statistic for which the null hypothesis will be rejected.

Thus the basic elements of every statistical test are:

© Giovanni Urga, 2009  liv  Adv Stats Induction, 2009-2010
0.2. SAMPLING THEORY

- a null hypothesis ($H_0$) that will be maintained until evidence to the contrary is known,
- an alternative hypothesis ($H_1$) that will be adopted if the null hypothesis is rejected, a test statistic and a region of rejection or critical region.

Example:
We wish to test a simple null hypothesis about the unknown mean $\beta$ of a normal population with variance that is known to be $\sigma^2 = 10$. Specifically, let the hypothesis be

$H_0 = \beta = 1$

$H_1 = \beta \neq 1$

Given a random sample of size $T = 10$, $Y_1, \ldots, Y_{10}$ we know that

$$\hat{\beta} = \sum_{i=1}^{T} Y_i / T \sim N(\beta, \sigma^2 / T) = N(\beta, 1)$$

Under a null hypothesis (i.e. ensuring that it is true) $\hat{\beta} \sim N(1, 1)$ is depicted in the figure below:

To define a critical region, we choose values of $\hat{\beta}$ that will lead us to reject $H_0$, and accept $H_1$. An intuitive rule is to select the rejection region to be values
of \( \hat{\beta} \) that are unlikely to occur if the null hypothesis is true. Then, if one of the unlikely values does occur, that provides evidence against the null hypothesis and would lead us to reject it.

In the previous figure we choose a and b so that.

\[
P[\hat{\beta} \leq a] = P[\hat{\beta} \geq b] = \frac{\alpha}{2}
\]

we can define an unbiased (since equal amounts of probability are in the two tails) test with size \( \alpha \) (which represents the probability of rejecting \( H_0 \) when it is true) as follows

Reject \( H_0 \) if \( \hat{\beta} \leq a \) or \( \hat{\beta} \geq b \)

Accept \( H_0 \) (at least do not reject it) if \( a < \hat{\beta} < b \).

For the problem at hand, the test is usually carried out using the test statistic

\[
z = \left( \frac{\hat{\beta} - 1}{\sigma / \sqrt{T}} \right)
\]

which is distributed as \( N(0,1) \) if \( H_0 \) is true. If \( H_0 \) is not true, say \( \beta = \hat{\beta}_0 \neq 1 \), then \( z \sim N(\beta_0 - 1, 1) \).

The rejection region for the test is \( z \geq z(\alpha/2) \) or \( z \leq -z(\alpha/2) \) or \( |z| \geq z(\alpha/2) \) as shown in figure below:

---

© Giovanni Urga, 2009

lvi

Adv Stats Induction, 2009-2010
If $\alpha = 0.05$, then $z(\alpha/2) = 1.96$ using tables of standard normal. In previous figure this means that $a = -0.96$ and $b = 2.96$

In terms of the test statistics $z$, the reasoning of the statistical test is as follows. If a sample value of $\hat{\beta}$ is obtained that leads to a calculated value of $|z| \geq z(\alpha/2)$, then we reject $H_0$ since

(i) If the hypothesis is true, the chances of obtaining a value of $z$ in the rejection region are “only” $\alpha$.

(ii) This is an unlikely event, so we conclude that the test statistic is unlikely to have a $N(0,1)$ distribution.

(iii) If the null hypothesis is true, the test statistic $z$ does have a $N(0,1)$ distribution.

We conclude that the hypothesis is unlikely to be true since the sample evidence does not support it.

Formulation of null and alternative hypothesis

Example:

In a court of law an accused person is “innocent until proven guilty”. The null hypothesis is that the person is innocent, and the alternative is that the person is guilty.

Null hypothesis in economic applications usually take the form of asserting that some parameter, linear combination of parameters, or set linear combinations of parameters take specified values. The alternative is simply that they do not.

Then, if $\vartheta = (\vartheta_1, \ldots, \vartheta_k)'$ is a set of parameters, the following pairs of null or alternative hypothesis are frequently seen.

(i) $H_o = \vartheta_i = 0$

$H_1 = \vartheta_i \neq 0$

(ii) $H_o = \vartheta_i + \vartheta_j = 1$

$H_1 = \vartheta_i + \vartheta_j \neq 1$

(iii) $H_o = \vartheta_1 = 0$ and $\vartheta_2 = 0$ and $\ldots$ and $\vartheta_k = 0$

$H_1$: at least one of the equalities in $H_0$ is false
Case (i) and Case (ii): the null hypothesis represents one condition on the parameter vector \( \vartheta \) and thus constitutes a single hypothesis. Also, the single hypothesis is said to be a simple hypothesis since it specifies a single value for the parameter and linear combination of parameters. The alternative hypothesis in cases (i) and (ii) is simply the negative of \( H_0 \) and a specific value is not given. Such a hypothesis is said to be composite as it represents more than one specific value for the parameter or linear combination of parameters.

Case (iii): the null hypothesis is joint since values are specified for several parameters and the function is do all these hypotheses hold simultaneously? The alternative in this case is that any one of the individual hypotheses is false.

In addition to these cases, inequality hypotheses are often stated such as

(iv) \( H_0 = \vartheta_i \geq 0 \)
    \( H_1 = \vartheta_i < 0 \)

In this case both the null and alternative hypotheses are composite.

Power of a Test

Since a hypothesis test involves making a choice, there is a chance that the decision made is an incorrect one.

Suppose
\( H_0 : \vartheta \in \Omega_0 \)
\( H_1 : \vartheta \in \Omega_1 \)

Once a test statistic and critical region have been defined, then the probability of rejecting \( H_0 \) can be determined for every \( \vartheta \in \Omega \).

Let \( \Pi(\vartheta) = P[\text{rejecting } H_0 | \vartheta] \) where \( H_0 \) is rejected when the test falls in the critical region. \( \Pi(\vartheta) \) is called power function of the test.

Ideally \( \Pi(\vartheta) = 0 \) for every value of \( \vartheta \in \Omega_0 \) and \( \Pi(\vartheta) = 1 \) for every \( \vartheta \in \Omega_1 \). This would imply that the test procedure always leads to the correct decision.

Unfortunately, such perfect test does not usually exist. Consequently, there is usually a nonzero probability of rejecting \( H_0 \) even when it is true (called Type 1 error).

When testing hypotheses, it is customary to only consider test procedures such that this probability is bounded by a constant \( \alpha \) called the size of the test or level of significance of the test. The size of the test \( \alpha \) is the largest value of \( \Pi(\vartheta) \) for any value of \( \vartheta \) that makes \( H_0 \) true, that is \( \vartheta \in \Omega_0 \).
Therefore
\[ P[\text{Type I error}] = P[\text{rejecting } H_0 \mid H_0 \text{ true}] \leq \alpha \]
If \( H_0 \) is “simple” or in cases (i) and (ii) then only the equality holds

The second possible error (called \textit{Type II error}) when testing a hypothesis is to accept a false hypothesis. The probability of a Type II error (often denoted \( \beta \), not to confused with a parameter \( \beta \)) is
\[ \beta = P[\text{Type II error}] = P[\text{accepting } H_0 \mid H_1 \text{ true}] = 1 - \Pi(\vartheta) \text{ for } \vartheta \in \Omega_1. \]

It is unfortunately true that no test procedure exist for a given sample size, that allow both \textit{Type I} and \textit{Type II errors} to be made arbitrarily small.

It is generally true that reducing the size of the test \( \alpha \) increases the probability of a Type II error and vice versa. The magnitude of the Type I error is conventionally fixed and usually at a small value, the reason being that it is the probability of rejecting the null hypothesis incorrectly, which is the error we want to avoid the most.

We do not want to reject the null hypothesis, as in the court room example, unless convincing evidence to the contrary has been provided.

Let us illustrate this point with the continuation of the previous example.

\textbf{Example:}

When \( H_0 \) is true, the power of the test is less than or equal to the size of the test \( \alpha \). In our example, the power of the test equals 0.05 when \( \beta = 1 \) and the hypothesis is true. To investigate the power when \( \beta \neq 1 \), we will begin by assuming that, in fact, \( \beta = 2 \). Then the true distribution of \( \hat{\beta} \) is \( \hat{\beta} \sim N(2,1) \) and the true distribution of the test statistic \( z \) is \( N(1,1) \).

In the next figure we compare the true distribution of \( \hat{\beta} \sim [N(2,1)] \) with the distribution under \( H_0 \) \( [N(1,1)] \).

Having defined the critical region of the test, we can compute the power of the test when \( a=2 \) as

\[
\Pi(2) = P[\hat{\beta} \leq -0.96 \text{ or } \hat{\beta} \geq 2.96 \mid \beta = 2] \\
= 1 - P \left[ \frac{-0.96 - 2}{1} \leq \frac{\hat{\beta} - 2}{1} \leq \frac{2.96 - 2}{1} \right] \\
= 1 - P[-2.96 \leq \hat{\beta} - 2 \leq 0.96] \\
= 0.17
\]
Relationship between confidence intervals and hypothesis tests

Suppose we are considering a population that is known to be \( N(\beta, 10) \) and we draw a random sample of size \( T = 10 \), say \( Y_1, \ldots, Y_{10} \). Then the maximum likelihood estimator of \( \beta \) is \( \hat{\beta} = \Sigma Y_i / 10 \sim N(\beta, 1) \). A 95% confidence interval for \( \beta \) is

\[
\hat{\beta} \pm z_{(\alpha/2)} \sigma / \sqrt{T} = \hat{\beta} \pm (1.96)
\]

Now consider testing the null hypothesis

\( H_0 = \beta = \beta_0 \) against

\( H_1 = \beta \neq \beta_0 \)

The test statistic is \( z = \sqrt{T}(\hat{\beta} - \beta_0) / \sigma = \hat{\beta} - \beta_0 \). The hypothesis is rejected if \( |z| \geq 1.96 \) and not rejected if \( |z| < 1.96 \).
Examining the latter statement, the hypothesis is not rejected if
\[-1.96 < \hat{\beta} - \beta_0 < 1.96\]
onlyear{or}
\[\hat{\beta} - 1.96 < \beta_0 < \beta + 1.96\]

That is, it is clear in the hypothesis-testing problem that any value \(\beta_0\) that falls “inside” the 95% confidence interval will lead to acceptance of the hypothesis at 0.05 level of significance. Any value of \(\beta_0\) outside the 95% confidence interval will lead to \(H_0\) being rejected at the level of significance 0.05.

Stated in another way, if a \((1 - \alpha) \times 100\%\) confidence interval covers the hypothesized value then the associated hypothesis test will be accepted.

The notes are based on the first four chapters of

© Giovanni Urga, 2009 lxi Adv Stats Induction, 2009-2010
Bibliography


Additional references are
Bibliography


