Notes on Probability Theory and Statistics

Antonis Demos

(Athens University of Economics and Business)

October 2002

Part I

Probability Theory

Chapter 1

INTRODUCTION

1.1 Set Theory Digression

A **set** is defined as any collection of objects, which are called **points** or **elements**. The biggest possible collection of points under consideration is called the **space**, **universe**, or **universal set**. For Probability Theory the space is called the **sample space**.

A set A is called a **subset** of B (we write $A \subseteq B$ or $B \supseteq A$) if every element of A is also an element of B. A is called a **proper subset** of B (we write $A \subset B$ or $B \supset A$) if every element of A is also an element of B and there is at least one element of B which does not belong to A.

Two sets A and B are called **equivalent sets** or **equal sets** (we write $A = B$) if $A \subseteq B$ and $B \subseteq A$.

If a set has no points, it will be called the **empty** or **null** set and denoted by ϕ.

The **complement** of a set A with respect to the space Ω, denoted by $\bar{A}$, A^c, or $\Omega - A$, is the set of all points that are in but not in A.

The **intersection** of two sets A and B is a set that consists of the common elements of the two sets and it is denoted by $A \cap B$ or AB.

The **union** of two sets A and B is a set that consists of all points that are in A or B or both (but only once) and it is denoted by $A \cup B$.

The **set difference** of two sets A and B is a set that consists of all points in

A that are not in B and it is denoted by $A - B$.

Properties of Set Operations

Commutative: $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

Associative: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$.

Distributive: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

$(A^c)^c = \overline{(\bar{A})} = A$ i.e. the complement of the A-complement is A.

If A subset of Ω (the space) then: $A \cap \Omega = A$, $A \cup \Omega = \Omega$, $A \cap \phi = \phi$, $A \cup \phi = A$, $A \cap \overline{A} = \phi$, $A \cup \overline{A} = \Omega$, $A \cap A = A$, and $A \cup A = A$.

De Morgan Law: $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$, and $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$.

Disjoint or mutually exclusive sets are the sets that their intersection is the empty set, i.e. A and B are mutually exclusive if $A \cap B = \phi$. Subsets $A_1, A_2, ...$ are mutually exclusive if $A_i \cap A_j = \phi$ for any $i \neq j$.

Uncertainty or variability are prevalent in many situations and it is the purpose of the probability theory to understand and quantify this notion. The basic situation is an experiment whose outcome is unknown before it takes place e.g., a) coin tossing, b) throwing a die, c) choosing at random a number from $\mathbb{N}$, d) choosing at random a number from $(0, 1)$.

The **sample space** is the collection or totality of all possible outcomes of a conceptual experiment. An **event** is a subset of the sample space. The class of all events associated with a given experiment is defined to be the **event space**.

Let us describe the sample space S, i.e. the set of all possible relevant outcomes of the above experiments, e.g., $S = \{H, T\}, S = \{1, 2, 3, 4, 5, 6\}$. In both of these examples we have a finite sample space. In example c) the sample space is a countable infinity whereas in d) it is an uncountable infinity.

Classical or a priori Probability: If a random experiment can result in N mutually exclusive and equally likely outcomes and if $N(A)$ of these outcomes have an attribute A, then the **probability** of A is the fraction $N(A)/N$ i.e. $P(A) = N(A)/N$,

where $N = N(A) + N(\overline{A})$.

EXAMPLE: Consider the drawing an ace (event A) from a deck of 52 cards. What is $P(A)$?

We have that $N(A) = 4$ and $N(\overline{A}) = 48$. Then $N = N(A) + N(\overline{A}) = 4 + 48 = 52$ and $P(A) = \frac{N(A)}{N} = \frac{4}{52}$

Frequency or a posteriori Probability: Is the ratio of the number α that an event A has occurred out of n trials, i.e. $P(A) = \alpha/n$.

EXAMPLE: Assume that we flip a coin 1000 times and we observe 450 heads. Then the a posteriori probability is $P(A) = \alpha/n = 450/1000 = 0.45$ (this is also the relative frequency). Notice that the a priori probability is in this case 0.5.

Subjective Probability: This is based on intuition or judgment.

We shall be concerned with a priori probabilities. These probabilities involve, many times, the counting of possible outcomes.

1.1.1 Some Counting Problems

Some more sophisticated discrete problems require counting techniques. For example:

a) What is the probability of getting four of a kind in a five card poker?

b) What is the probability that two people in a classroom have the same birthday?

The sample space in both cases, although discrete, can be quite large and it not feasible to write out all possible outcomes.

1. Duplication is permissible and Order is important (Multiple Choice Arrangement), i.e. the element AA is permitted and AB is a different element from BA. In this case where we want to arrange n objects in x places the possible outcomes is given from: $M_x^n = n^x$.

EXAMPLE: Find all possible combinations of the letters A, B, C, and D when duplication is allowed and order is important.

The result according to the formula is: $n = 4$, and $x = 2$, consequently the

possible number of combinations is $M_2^4 = 4^2 = 16$. To find the result we can also use a tree diagram.

2. Duplication is not permissible and Order is important (Permutation Arrangement), i.e. the element AA is **not** permitted and AB is a different element from BA. In this case where we want to permute n objects in x places the possible outcomes is given from:

$$P_x^n \quad or \quad P(n,x) = n \times (n-1) \times ..(n-x+1) = \frac{n!}{(n-x)!}.$$

EXAMPLE: Find all possible permutations of the letters A, B, C, and D when duplication is not allowed and order is important.

The result according to the formula is: $n = 4$, and $x = 2$, consequently the possible number of combinations is $P_2^4 = \frac{4!}{(4-2)!} = \frac{2*3*4}{2} = 12$.

3. Duplication is not permissible and Order is not important (Combination Arrangement), i.e. the element AA is **not** permitted and AB is **not** a different element from BA. In this case where we want the combinations of n objects in x places the possible outcomes is given from:

$$C_x^n \quad or \quad C(n,x) = \frac{P(n,x)}{x!} = \frac{n!}{(n-x)!x!} = \binom{n}{x}$$

EXAMPLE: Find all possible combinations of the letters A, B, C, and D when duplication is not allowed and order is not important.

The result according to the formula is: $n = 4$, and $x = 2$, consequently the possible number of combinations is $C_2^4 = \frac{4!}{2!*(4-2)!} = \frac{2*3*4}{2*2} = 6$.

Let us now define probability rigorously.

1.1.2 Definition of Probability

Consider a collection of sets A_α with index $\alpha \in \Gamma$, which is denoted by $\{A_\alpha : \alpha \in \Gamma\}$. We can define for an index Γ of arbitrary cardinality (the cardinal number of a set is the number of elements of this set):

$$\bigcup_{\alpha\in\Gamma} A_\alpha = \{x \in S : x \in A_\alpha \text{ for some } \alpha \in \Gamma\}$$

$\underset{\alpha\in\Gamma}{\cap} A_\alpha = \{x \in S : x \in A_\alpha \text{ for all } \alpha \in \Gamma\}$

A collection is exhaustive if $\cup_{\alpha\in\Gamma} A_\alpha = S$ (partition), and is pairwise exclusive or disjoint if $A_\alpha \cap A_\beta = \phi$, $\alpha \neq \beta$.

To define probabilities we need some further structure. This is because in uncountable cases we can not just define probability for all subsets of S, as there are some sets on the real line whose probability can not be determined, i.e., they are unmeasurable. We shall define probability on a family of subsets of S, of which we require the following structure.

Definition 1 *Let be $\mathcal{A}$ a non-empty class of subsets of S. $\mathcal{A}$ is an algebra if*

1. $A^c \in \mathcal{A}$, whenever $A \in \mathcal{A}$

2. $A_1 \cup A_2 \in \mathcal{A}$, whenever $A_1, A_2 \in \mathcal{A}$.

$\mathcal{A}$ is a σ-algebra if also

2′. $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$, whenever $A_n \in \mathcal{A}$, n=1,2,3,...

Note that since $\mathcal{A}$ is non-empty, (1) and (2) $\Rightarrow \phi \in \mathcal{A}$ and $S \in \mathcal{A}$. Note also that $\cap_{n=1}^{\infty} A_n \in \mathbf{A}$. The largest σ-algebra is the set of all subsets of S, denoted by $\mathcal{P}(S)$, and the smallest is $\{\phi, S\}$. We can generate a σ-algebra from any collection of subsets by adding to the set the complements and the unions of its elements. For example let $S = \mathbb{R}$, and

$$\mathcal{B} = \{[a,b], (a,b], [a,b), (a,b),\ a,b \in \mathbb{R}\},$$

and let $\mathcal{A} = \sigma(\mathcal{B})$ consists of all intervals and countable unions of intervals and complements thereof. This is called the Borel σ-algebra and is the usual σ-algebra we work when $S = \mathbb{R}$. The σ-algebra $\mathcal{A} \subset \mathcal{P}(\mathbb{R})$, i.e., there are sets in $\mathcal{P}(\mathbb{R})$ not in $\mathcal{A}$. These are some pretty nasty ones like the Cantor set. We can alternatively construct the Borel σ-algebra by considering $\mathcal{J}$ the set of all intervals of the form $(-\infty, x]$, $x \in \mathbb{R}$. We can prove that $\sigma(\mathcal{J}) = \sigma(\mathcal{B})$. We can now give the definition of probability measure which is due to Kolmogorov.

Definition 2 *Given a sample space S and a σ-algebra $(S, \mathcal{A})$, a probability measure is a mapping from $\mathcal{A} \longrightarrow \mathbb{R}$ such that*

1. $P(A) \geq 0$ *for all* $A \in \mathcal{A}$
2. $P(S) = 1$
3. *if* $A_1, A_2, ...$ *are pairwise disjoint, i.e.,* $A_i \cap A_j = \phi$ *for all* $i \neq j$*, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P\left(A_i\right)$$

In such a way we have a probability space $(S, \mathcal{A}, P)$. When S is discrete we usually take $\mathcal{A} = \mathcal{P}(S)$. When $S = \mathbb{R}$ or some subinterval thereof, we take $\mathcal{A} = \sigma(\mathcal{B})$.

P is a matter of choice and will depend on the problem. In many discrete cases, the problem can usually be written such that outcomes are equally likely.

$$P(\{x\}) = 1/n, \quad n = \sharp(S).$$

In continuous cases, P is usually like Lebesgue measure, i.e.,

$$P((a, b)) \propto b - a.$$

Properties of P

1. $P(\phi) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$
4. $P(B \cap A^c) = P(B) - P(B \cap A)$
5. If $A \subset B \Rightarrow P(A) \leq P(B)$
6. $P(B \cup A) = +P(A) + P(B) - P(A \cap B)$ More generally, for events $A_1, A_2, ...A_n \in \mathcal{A}$ we have:

$$P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} P[A_i] - \sum\sum_{i<j} P[A_i A_j] + \sum\sum\sum_{i<j<k} P[A_i A_j A_k] - .. + (-1)^{n+1} P[A_1..A_n].$$

For $n = 3$ the above formula is:

$$P\left[A_1 \bigcup A_2 \bigcup A_3\right] = P[A_1] + P[A_2] + P[A_3] - P[A_1 A_2] - P[A_1 A_3] - P[A_2 A_3] + P[A_1 A_2 A_2].$$

7. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$

Proofs involve manipulating sets to obtain disjoint sets and then apply the axioms.

1.2 Conditional Probability and Independence

In many statistical applications we have variables X and Y (or events A and B) and want to explain or predict Y or A from X or B, we are interested not only in marginal probabilities but in conditional ones as well, i.e., we want to incorporate some information in our predictions. Let A and B be two events in $\mathcal{A}$ and a probability function $P(.)$. The **conditional probability** of A given event B, is denoted by $P[A|B]$ and is defined as follows:

Definition 3 *The probability of an event A given an event B, denoted by $P(A|B)$, is given by*

$$P([A|B) = \frac{P(A \cap B)}{P(B)} \quad if \quad P(B) > 0$$

and is left undefined if $P(B) = 0$.

From the above formula is evident $P[AB] = P[A|B]P[B] = P[B|A]P[A]$ if both $P[A]$ and $P[B]$ are nonzero. Notice that when speaking of conditional probabilities we are conditioning on some given event B; that is, we are assuming that the experiment has resulted in some outcome in B. B, in effect then becomes our "new" sample space. All probability properties of the previous section apply to conditional probabilities as well, i.e. $P(\cdot|B)$ is a probability measure. In particular:

1. $P(A|B) \geq 0$
2. $P(S|B) = 1$
3. $P(\cup_{i=1}^{\infty} A_i|B) = \sum_{i=1}^{\infty} P(A_i|B)$ for any pairwise disjoint events$\{A_i\}_{i=1}^{\infty}$.

Note that if A and B are mutually exclusive events, $P(A|B) = 0$. When $A \subseteq B$, $P(A|B) = \frac{P(A)}{P(B)} \geq P(A)$ with strict inequality unless $P(B) = 1$. When $B \subseteq A$, $P(A|B) = 1$.

However, there is an additional property (Law) called the **Law of Total Probabilities** which states that:

LAW OF TOTAL PROBABILITY:

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

For a given probability space $(\Omega, \mathcal{A}, P[.])$, if $B_1, B_2, ..., B_n$ is a collection of mutually exclusive events in $\mathcal{A}$ satisfying $\bigcup_{i=1}^{n} B_i = \Omega$ and $P[B_i] > 0$ for $i = 1, 2, ..., n$ then for every $A \in \mathcal{A}$,

$$P[A] = \sum_{i=1}^{n} P[A|B_i]P[B_i]$$

Another important theorem in probability is the so called **Bayes' Theorem** which states:

BAYES RULE: Given a probability space $(\Omega, \mathcal{A}, P[.])$, if $B_1, B_2, ..., B_n$ is a collection of mutually exclusive events in $\mathcal{A}$ satisfying $\bigcup_{i=1}^{n} B_i = \Omega$ and $P[B_i] > 0$ for $i = 1, 2, ..., n$ then for every $A \in \mathcal{A}$ for which $P[A] > 0$ we have:

$$P[B_j|A] = \frac{P[A|B_j]P[B_j]}{\sum_{i=1}^{n} P[A|B_i]P[B_i]}$$

Notice that for events A and $B \in \mathcal{A}$ which satisfy $P[A] > 0$ and $P[B] > 0$ we have:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

This follows from the definition of conditional independence and the law of total probability. The probability $P(B)$ is a prior probability and $P(A|B)$ frequently is a likelihood, while $P(B|A)$ is the posterior.

Finally the **Multiplication Rule** states:

Given a probability space $(\Omega, \mathcal{A}, P[.])$, if $A_1, A_2, ..., A_n$ are events in $\mathcal{A}$ for which $P[A_1 A_2 .....A_{n-1}] > 0$ then:

$$P[A_1A_2.....A_n] = P[A_1]P[A_2|A_1]P[A_3|A_1A_2].....P[A_n|A_1A_2....A_{n-1}]$$

EXAMPLE: A plant has two machines. Machine A produces 60% of the total output with a fraction defective of 0.02. Machine B the rest output with a fraction defective of 0.04. If a single unit of output is observed to be defective, what is the probability that this unit was produced by machine A?

If A is the event that the unit was produced by machine A, B the event that the unit was produced by machine B and D the event that the unit is defective. Then we ask what is $P[A|D]$. But $P[A|D] = \frac{P[AD]}{P[D]}$. Now $P[AD] = P[D|A]P[A] = 0.02 * 0.6 = 0.012$. Also $P[D] = P[D|A]P[A] + P[D|B]P[B] = 0.012 + 0.04 * 0.4 = 0.028$. Consequently, $P[A|D] = 0.571$. Notice that $P[B|D] = 1 - P[A|D] = 0.429$. We can also use a tree diagram to evaluate $P[AD]$ and $P[BD]$.

EXAMPLE: A marketing manager believes the market demand potential of a new product to be high with a probability of 0.30, or average with probability of 0.50, or to be low with a probability of 0.20. From a sample of 20 employees, 14 indicated a very favorable reception to the new product. In the past such an employee response (14 out of 20 favorable) has occurred with the following probabilities: if the actual demand is high, the probability of favorable reception is 0.80; if the actual demand is average, the probability of favorable reception is 0.55; and if the actual demand is low, the probability of the favorable reception is 0.30. Thus given a favorable reception, what is the probability of actual high demand?

Again what we ask is $P[H|F] = \frac{P[HF]}{P[F]}$. Now $P[F] = P[H]P[F|H]+P[A]P[F|A]+P[L]P[F|L] = 0.24+0.275+0.06 = 0.575$. Also $P[HF] = P[F|H]P[H] = 0.24$. Hence $P[H|F] = \frac{0.24}{0.575} = 0.4174$

EXAMPLE: There are five boxes and they are numbered 1 to 5. Each box contains 10 balls. Box i has i defective balls and $10-i$ non-defective balls, $i = 1, 2, .., 5$. Consider the following random experiment: First a box is selected at random, and then a ball is selected at random from the selected box. 1) What is the probability

that a defective ball will be selected? 2) If we have already selected the ball and noted that is defective, what is the probability that it came from the box 5?

Let A denote the event that a defective ball is selected and B_i the event that box i is selected, $i = 1, 2, .., 5$. Note that $P[B_i] = 1/5$, for $i = 1, 2, .., 5$, and $P[A|B_i] = i/10$. Question 1) is what is $P[A]$? Using the theorem of total probabilities we have:

$P[A] = \sum_{i=1}^{5} P[A|B_i]P[B_i] = \sum_{i=1}^{5} \frac{i}{5}\frac{1}{5} = 3/10$. Notice that the total number of defective balls is 15 out of 50. Hence in this case we can say that $P[A] = \frac{15}{50} = 3/10$. This is true as the probabilities of choosing each of the 5 boxes is the same. Question 2) asks what is $P[B_5|A]$. Since box 5 contains more defective balls than box 4, which contains more defective balls than box 3 and so on, we expect to find that $P[B_5|A] > P[B_4|A] > P[B_3|A] > P[B_2|A] > P[B_1|A]$. We apply Bayes' theorem:

$$P[B_5|A] = \frac{P[A|B_5]P[B_5]}{\sum_{i=1}^{5} P[A|B_i]P[B_i]} = \frac{\frac{1}{2}\frac{1}{5}}{\frac{3}{10}} = \frac{1}{3}$$

Similarly $P[B_j|A] = \frac{P[A|B_j]P[B_j]}{\sum_{i=1}^{5} P[A|B_i]P[B_i]} = \frac{\frac{j}{10}\frac{1}{5}}{\frac{3}{10}} = \frac{j}{15}$ for $j = 1, 2, ..., 5$. Notice that unconditionally all $B_i's$ were equally likely.

Let A and B be two events in $\mathcal{A}$ and a probability function $P(.)$. Events A and B are defined **independent** if and only if one of the following conditions is satisfied:

(i) $P[AB] = P[A]P[B]$.

(ii) $P[A|B] = P[A]$ if $P[B] > 0$.

(iii) $P[B|A] = P[B]$ if $P[A] > 0$.

These are equivalent definitions except that (i) does not really require $P(A)$, $P(B) > 0$. **Notice** that the property of two events A and B and the property that A and B are mutually exclusive are distinct, though related properties. We know that if A and B are mutually exclusive then $P[AB] = 0$. Now if these events are

also independent then $P[AB] = P[A]P[B]$, and consequently $P[A]P[B] = 0$, which means that either $P[A] = 0$ or $P[B] = 0$. Hence two mutually exclusive events are independent if $P[A] = 0$ or $P[B] = 0$. On the other hand if $P[A] \neq 0$ and $P[B] \neq 0$, then if A and B are independent can not be mutually exclusive and oppositely if they are mutually exclusive can not be independent. Also notice that independence is not transitive, i.e., A independent of B and B independent of C does not imply that A is independent of C.

EXAMPLE: Consider tossing two dice. Let A denote the event of an odd total, B the event of an ace on the first die, and C the event of a total of seven. We ask the following:

(i) Are A and B independent?

(ii) Are A and C independent?

(iii) Are B and C independent?

(i) $P[A|B] = 1/2$, $P[A] = 1/2$ hence $P[A|B] = P[A]$ and consequently A and B are independent.

(ii) $P[A|C] = 1 \neq P[A] = 1/2$ hence A and C are not independent.

(iii) $P[C|B] = 1/6 = P[C] = 1/6$ hence B and C are independent.

Notice that although A and B are independent and C and B are independent A and C are not independent.

Let us extend the independence of two events to several ones:

For a given probability space $(\Omega, \mathcal{A}, P[.])$, let $A_1, A_2, ..., A_n$ be n events in $\mathcal{A}$. Events $A_1, A_2, ..., A_n$ are defined to be **independent** if and only if:

$P[A_iA_j] = P[A_i]P[A_j]$ for $i \neq j$

$P[A_iA_jA_k] = P[A_i]P[A_j]P[A_k]$ for $i \neq j, i \neq k, k \neq j$

and so on

$P[\bigcap_{i=1}^{n} A_i] = \prod_{i=1}^{n} P[A_i]$

Notice that pairwise independence does not imply independence, as the following example shows.

EXAMPLE: Consider tossing two dice. Let A_1 denote the event of an odd face in the first die, A_2 the event of an odd face in the second die, and A_3 the event of an odd total. Then we have: $P[A_1]P[A_2] = \frac{1}{2}\frac{1}{2} = P[A_1A_2]$, $P[A_1]P[A_3] = \frac{1}{2}\frac{1}{2} = P[A_3|A_1]P[A_1] = P[A_1A_3]$, and $P[A_2A_3] = \frac{1}{4} = P[A_2]P[A_3]$ hence A_1, A_2, A_3 are pairwise independent. However notice that $P[A_1A_2A_3] = 0 \neq \frac{1}{8} = P[A_1]P[A_2]P[A_3]$. Hence A_1, A_2, A_3 are **not** independent.

Chapter 2

RANDOM VARIABLES, DISTRIBUTION FUNCTIONS, AND DENSITIES

The probability space $(S, \mathcal{A}, P)$ is not particularly easy to work with. In practice, we often need to work with spaces with some structure (metric spaces). It is convenient therefore to work with a cardinalization of S by using the notion of random variable.

Formally, a random variable X is just a mapping from the sample space to the real line, i.e.,

$$X : S \longrightarrow \mathbb{R},$$

with a certain property, it is a measurable mapping, i.e.

$$\mathcal{A}_X = \{A \subset S : X(A) \in \mathcal{B}\} = \left\{X^{-1}(B) : B \in \mathcal{B}\right\} \subseteq \mathcal{A},$$

where $\mathcal{B}$ is a sigma-algebra on $\mathbb{R}$, for any B in $\mathcal{B}$ the inverse image belongs to $\mathcal{A}$. The probability measure P_X can then be defined by

$$P_X\left(X \in B\right) = P\left(X^{-1}\left(B\right)\right).$$

It is straightforward to show that $\mathcal{A}_X$ is a σ-algebra whenever $\mathcal{B}$ is. Therefore, P_X is a probability measure obeying Kolmogorov's axioms. Hence we have transferred $(S, \mathcal{A}, P) \longrightarrow (\mathbb{R}, \mathcal{B}, P_X)$, where $\mathcal{B}$ is the Borel σ-algebra when $X(S) = \mathbb{R}$ or any uncountable set, and $\mathcal{B}$ is $\mathcal{P}\left(X\left(S\right)\right)$ when $X\left(S\right)$ is finite. The function $X(.)$ must be such that the set A_r, defined by $A_r = \{\omega : X(\omega) \leq r\}$, belongs to $\mathcal{A}$ for every real number r, as elements of $\mathcal{B}$ are left-closed intervals of $\mathbb{R}$.

The important part of the definition is that in terms of a random experiment, S is the totality of outcomes of that random experiment, and the function, or random variable, $X(.)$ with domain S makes some real number correspond to each outcome of the experiment. The fact that we also require the collection of $\omega's$ for which $X(\omega) \leq r$ to be an event (i.e. an element of $\mathcal{A}$) for each real number r is not much of a restriction since the use of random variables is, in our case, to describe only events.

EXAMPLE: Consider the experiment of tossing a single coin. Let the random variable X denote the number of heads. In this case $S = \{head, tail\}$, and $X(\omega) = 1$ if $\omega = head$, and $X(\omega) = 0$ if $\omega = tail$. So the random variable X associates a real number with each outcome of the experiment. To show that X satisfies the definition we should show that $\{\omega : X(\omega) \leq r\}$, belongs to $\mathcal{A}$ for every real number r. $\mathcal{A} = \{\phi, \{head\}, \{tail\}, S\}$. Now if $r < 0$, $\{\omega : X(\omega) \leq r\} = \phi$, if $0 \leq r < 1$ then $\{\omega : X(\omega) \leq r\} = \{tail\}$, and if $r \geq 1$ then $\{\omega : X(\omega) \leq r\} = \{head, tail\} = S$. Hence, for each r the set $\{\omega : X(\omega) \leq r\}$ belongs to $\mathcal{A}$ and consequently $X(.)$ is a random variable.

In the above example the random variable is described in terms of the random experiment as opposed to its functional form, which is the usual case.

We can now work with $(\mathbb{R}, \mathcal{B}, P_X)$, which has metric structure and algebra. For example, we toss two die in which case the sample space is

$$S = \{(1,1), (1,2), ..., (6,6)\}.$$

We can define two random variables: the Sum and Product:

$$\begin{aligned} X(S) &= \{2,3,4,5,6,7,8,9,10,11,12\} \\ X(S) &= \{1,2,3,4,5,6,8,9,10,...,36\} \end{aligned}$$

The simplest form of random variables are the indicators I_A

$$I_A(s) = \begin{cases} 1 & if \quad s \in A \\ 0 & if \quad s \notin A \end{cases}$$

This has associated sigma algebra in S

$$\{\phi, S, A, A^c\}$$

Finally, we give formal definition of a continuous real-valued random variable.

Definition 4 *A random variable is continuous if its probability measure P_X is absolutely continuous with respect to Lebesgue measure, i.e., $P_X(A) = 0$ whenever $\lambda(A) = 0$.*

2.0.1 Distribution Functions

Associated with each random variable there is the distribution function

$$F_X(x) = P_X(X \leq x)$$

defined for all $x \in \mathbb{R}$. This function effectively replaces P_X. Note that we can reconstruct P_X from F_X.

EXAMPLE. $S = \{H, T\}$, $X(H) = 1$, $X(T) = 0$, $(p = 1/2)$.

If $x < 0$, $F_X(x) = 0$

If $0 \leq x < 1$, $F_X(x) = 1/2$

If $x \geq 1$, $F_X(x) = 1$.

EXAMPLE. The logit c.d.f. is

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

It is continuous everywhere and asymptotes to 0 and 1 at $\pm\infty$ respectively. Strictly increasing.

Note that the distribution function $F_X(x)$ of a continuous random variable is a continuous function. The distribution function of a discrete random variable is a step function.

Theorem 5 *A function $F(\cdot)$ is a c.d.f. of a random variable X if and only if the following three conditions hold*

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

2. F *is a nondecreasing function in* x

3. F *is right-continuous, i.e., for all* x_0, $\lim_{x\to x_0^+} F(x) = F(x_0)$

4. F *is continuous except at a set of points of Lebesgue measure zero.*

2.0.2 Discrete Random Variables.

As we have already said, a random variable X will be defined to be **discrete** if the range of X is countable. If a random variable X is discrete, then its corresponding cumulative distribution function $F_X(.)$ will be defined to be **discrete**, i.e. a step function.

By the range of X being countable we mean that there exists a finite or denumerable set of real numbers, say $x_1, x_2, ...x_n, ...$, such that X takes on values only in that set. If X is discrete with distinct values $x_1, x_2, ...x_n, ...$, then $S = \bigcup_n \{\omega : X(\omega) = x_n\}$, and $\{X = x_i\} \cap \{X = x_j\} = \phi$ for $i \neq j$. Hence $1 = P[S] = \sum_n P[X = x_n]$ by the third axiom of probability.

If X is a discrete random variable with distinct values $x_1, x_2, ...x_n, ...$, then the function, denoted by $f_X(.)$ and defined by

$$f_X(x) = \begin{cases} P[X = x] & if \quad x = x_j, \quad j = 1, 2, ..., n, ... \\ 0 & if \quad x \neq x_j \end{cases}$$

is defined to be the **discrete density function** of X.

Notice that the discrete density function tell us how likely or probable each of the values of a discrete random variable is. It also enables one to calculate the probability of events described in terms of the discrete random variable. Also notice that for any discrete random variable X, $F_X(.)$ can be obtained from $f_X(.)$, and vice versa

EXAMPLE: Consider the experiment of tossing a single die. Let X denote the number of spots on the upper face. Then for this case we have:

X takes any value from the set $\{1, 2, 3, 4, 5, 6\}$. So X is a discrete random variable. The density function of X is: $f_X(x) = P[X = x] = 1/6$ for any

$x \in \{1,2,3,4,5,6\}$ and 0 otherwise. The cumulative distribution function of X is: $F_X(x) = P[X \leq x] = \sum_{n=1}^{[x]} P[X = n]$ where $[x]$ denotes the integer part of x.. Notice that x can be any real number. However, the points of interest are the elements of $\{1,2,3,4,5,6\}$. Notice also that in this case $\Omega = \{1,2,3,4,5,6\}$ as well, and we do not need any reference to $\mathcal{A}$. ■

EXAMPLE: Consider the experiment of tossing two dice. Let X denote the total of the upturned faces. Then for this case we have:

$\Omega = \{(1,1),(1,2),...(1,6),(2,1),(2,2),....(2,6),(3,1),.....,(6,6)\}$ a total of (using the Multiplication rule) $36 = 6^2$ elements. X takes values from the set $\{2,3,4,5,6,7,8,9,10,11,12$ The density function is:

$$f_X(x) = P[X = x] = \begin{cases} 1/36 & for \quad x = 2 \quad or \quad x = 12 \\ 2/36 & for \quad x = 3 \quad or \quad x = 11 \\ 3/36 & for \quad x = 4 \quad or \quad x = 10 \\ 4/36 & for \quad x = 5 \quad or \quad x = 9 \\ 5/36 & for \quad x = 6 \quad or \quad x = 8 \\ 1/36 & for \quad x = 7 \\ 0 & for \quad any \quad other \quad x \end{cases}$$

The cumulative distribution function is:

$$F_X(x) = P[X \leq x] = \sum_{n=1}^{[x]} P[X = n] = \begin{cases} 0 & for \quad x < 2 \\ \frac{1}{36} & for \quad 2 \leq x < 3 \\ \frac{3}{36} & for \quad 3 \leq x < 4 \\ \frac{6}{36} & for \quad 4 \leq x < 5 \\ \frac{10}{36} & for \quad 5 \leq x < 6 \\ & ..... \\ \frac{35}{36} & for \quad 11 \leq x < 12 \\ 1 & for \quad 12 \leq x \end{cases}$$

Notice that, again, we do not need any reference to $\mathcal{A}$. ■

In fact we can speak of discrete density functions without reference to some random variable at all.

Any function $f(.)$ with domain the real line and counterdomain $[0,1]$ is defined to be a **discrete density function** if for some countable set $x_1, x_2, ...x_n, ...$ has the following properties:

i) $f(x_j) > 0$ for $j = 1, 2, ...$

ii) $f(x) = 0$ for $x \neq x_j$; $j = 1, 2, ...$

iii) $\sum f(x_j) = 1$, where the summation is over the points $x_1, x_2, ...x_n, ...$

2.0.3 Continuous Random Variables

A random variable X is called **continuous** if there exist a function $f_X(.)$ such that $F_X(x) = \int_{-\infty}^{x} f_X(u)du$ for every real number x. In such a case $F_X(x)$ is the **cumulative distribution** and the function $f_X(.)$ is the **density function**.

Notice that according to the above definition the density function is not uniquely determined. The idea is that if the a function change value if a few points its integral is unchanged. Furthermore, notice that $f_X(x) = dF_X(x)/dx$.

The notations for discrete and continuous density functions are the same, yet they have different interpretations. We know that for discrete random variables $f_X(x) = P[X = x]$, which is not true for continuous random variables. Furthermore, for discrete random variables $f_X(.)$ is a function with domain the real line and counterdomain the interval $[0,1]$, whereas, for continuous random variables $f_X(.)$ is a function with domain the real line and counterdomain the interval $[0, \infty)$. Note that for a continuous r.v.

$$P(X = x) \leq P(x - \varepsilon \leq X \leq x) = F_X(x) - F_X(x - \varepsilon) \to 0$$

as $\varepsilon \to 0$, by the continuity of $F_X(x)$. The set $\{X = x\}$ is an example of a set of measure (in this case the measure is P or P_X) zero. In fact, any countable set is of measure zero under a distribution which is absolutely continuous with respect to Lebesgue measure. Because the probability of a singleton is zero

$$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$$

for any a, b.

EXAMPLE: Let X be the random variable representing the length of a telephone conversation. One could model this experiment by assuming that the distribution of X is given by $F_X(x) = (1 - e^{-\lambda x})$ where λ is some positive number and the random variable can take values only from the interval $[0, \infty)$. The density function is $dF_X(x)/dx = f_X(x) = \lambda e^{-\lambda x}$. If we assume that telephone conversations are measured in minutes, $P[5 < X \leq 10] = \int_5^{10} f_X(x)dx = \int_5^{10} \lambda e^{-\lambda x}dx = e^{-5\lambda} - e^{-10\lambda}$, and for $\lambda = 1/5$ we have that $P[5 < X \leq 10] = e^{-1} - e^{-2} = 0.23$. ■

The example above indicates that the density functions of continuous random variables are used to calculate probabilities of events defined in terms of the corresponding continuous random variable X i.e. $P[a < X \leq b] = \int_a^b f_X(x)dx$. Again we can give the definition of the density function without any reference to the random variable i.e. any function $f(.)$ with domain the real line and counterdomain $[0, \infty)$ is defined to be a **probability density function** iff

(i) $f(x) \geq 0$ for all x

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$.

In practice when we refer to the certain distribution of a random variable, we state its density or cumulative distribution function. However, notice that not all random variables are either discrete or continuous.

Chapter 3

EXPECTATIONS AND MOMENTS OF RANDOM VARIABLES

An extremely useful concept in problems involving random variables or distributions is that of expectation.

3.0.4 Mean or Expectation

Let X be a random variable. The **mean** or the **expected value** of X, denoted by $E[X]$ or μ_X, is defined by:

(i) $E[X] = \sum x_j P[X = x_j] = \sum x_j f_X(x_j)$

if X is a discrete random variable with counterdomain the countable set $\{x_1, ..., x_j, ..\}$

(ii) $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$

if X is a continuous random variable with density function $f_X(x)$ and if either $\left|\int_0^{\infty} x f_X(x) dx\right| < \infty$ or $\left|\int_{-\infty}^{0} x f_X(x) dx\right| < \infty$ or both.

(iii) $E[X] = \int_0^{\infty} [1 - F_X(x)] dx - \int_{-\infty}^{0} F_X(x) dx$

for an arbitrary random variable X.

(i) and (ii) are used in practice to find the mean for discrete and continuous random variables, respectively. (iii) is used for the mean of a random variable that is neither discrete nor continuous.

Notice that in the above definition we assume that the sum and the integrals exist. Also that the summation in (i) runs over the possible values of j and the j^{th} term is the value of the random variable multiplied by the probability that the random variable takes this value. Hence $E[X]$ is an average of the values that the

random variable takes on, where each value is weighted by the probability that the random variable takes this value. Values that are more probable receive more weight. The same is true in the integral form in (ii). There the value x is multiplied by the approximate probability that X equals the value x, i.e. $f_X(x)dx$, and then integrated over all values.

Notice that in the definition of a mean of a random variable, only density functions or cumulative distributions were used. Hence we have really defined the mean for these functions without reference to random variables. We then call the defined mean the mean of the cumulative distribution or the appropriate density function. Hence, we can speak of the mean of a distribution or density function as well as the mean of a random variable.

Notice that $E[X]$ is the center of gravity (or centroid) of the unit mass that is determined by the density function of X. So the mean of X is a measure of where the values of the random variable are centered or located i.e. is a measure of central location.

EXAMPLE: Consider the experiment of tossing two dice. Let X denote the total of the upturned faces. Then for this case we have:

$$E[X] = \sum_{i=2}^{12} i f_X(i) = 7$$

EXAMPLE: Consider a X that can take only to possible values, 1 and -1, each with probability 0.5. Then the mean of X is:

$$E[X] = 1 * 0.5 + (-1) * 0.5 = 0$$

Notice that the mean in this case is not one of the possible values of X.

EXAMPLE: Consider a continuous random variable X with density function $f_X(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$. Then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{0}^{\infty} x \lambda e^{-\lambda x} dx = 1/\lambda$$

EXAMPLE: Consider a continuous random variable X with density function $f_X(x) = x^{-2}$ for $x \in [1, \infty)$. Then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_{1}^{\infty} x x^{-2} dx = \lim_{b \to \infty} \log b = \infty$$

so we say that the mean does not exist, or that it is infinite.

Median of X: When F_X is continuous and strictly increasing, we can define the median of X, denoted $m(X)$, as being the unique solution to

$$F_X(m) = \frac{1}{2}.$$

Since in this case, $F_X^{-1}(\cdot)$ exists, we can alternatively write $m = F_X^{-1}(\frac{1}{2})$. For discrete r.v., there may be many m that satisfy this or may none. Suppose

$$X = \begin{cases} 0 & 1/3 \\ 1 & 1/3 \\ 2 & 1/3 \end{cases},$$

then there does not exist an m with $F_X(m) = \frac{1}{2}$. Also, if

$$X = \begin{cases} 0 & 1/4 \\ 1 & 1/4 \\ 2 & 1/4 \\ 3 & 1/4 \end{cases},$$

then any $1 \leq m \leq 2$ is an adequate median.

Note that if $E(X^n)$ exists, then so does $E(X^{n-1})$ but not vice versa ($n > 0$). Also when the support is infinite, the expectation does not necessarily exist.

If $\int_0^\infty x f_X(x)dx = \infty$ but $\int_{-\infty}^0 x f_X(x)dx > -\infty$, then $E(X) = \infty$

If $\int_0^\infty x f_X(x)dx = \infty$ and $\int_{-\infty}^0 x f_X(x)dx = -\infty$, then $E(X)$is not defined.

EXAMPLE: [Cauchy] $f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. This density function is symmetric about zero, and one is temted to say that $E(X) = 0$. But $\int_0^\infty x f_X(x)dx = \infty$ and $\int_{-\infty}^0 x f_X(x)dx = -\infty$, so $E(X)$ does not exist according to the above definition.

Now consider $Y = g(X)$, where g is a (piecewise) monotonic continuous function. Then

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{-\infty}^{\infty} g(x) f_X(x)dx = E(g(x))$$

Theorem 6 *Expectation has the following properties:*

1. [Linearity] $E\left(a_1 g_1(X) + a_2 g_2(X) + a_3\right) = a_1 E\left(g_1(X)\right) + a_2 E\left(g_2(X)\right) + a_3$

2. [Monotonicity] If $g_1(x) \geq g_2(x) \Rightarrow E\left(g_1(X)\right) \geq E\left(g_2(X)\right)$

3. Jensen's inequality. If $g(x)$ *is a weakly convex function, i.e.,* $g\left(\lambda x + (1-\lambda) y\right) \leq \lambda g\left(x\right) + (1-\lambda) g\left(y\right)$ *for all* x, y, *and all with* $0 \leq \lambda \leq 1$, *then*

$$E\left(g(X)\right) \geq g\left(E\left(X\right)\right).$$

An Interpretation of Expectation

We claim that $E\left(X\right)$ is the unique minimizer of $E\left(X-\theta\right)^2$ with respect to θ, assuming that the second moment of X is finite.

Theorem 7 *Suppose that* $E\left(X^2\right)$ *exists and is finite, then* $E\left(X\right)$ *is the unique minimizer of* $E\left(X-\theta\right)^2$ *with respect to* θ.

This Theorem says that the Expectation is the closest quantity to θ, in mean square error.

3.0.5 Variance

Let X be a random variable and μ_X be $E[X]$. The **variance** of X, denoted by σ_X^2 or $var[X]$, is defined by:

(i) $var[X] = \sum (x_j - \mu_X)^2 P[X = x_j] = \sum (x_j - \mu_X)^2 f_X(x_j)$

if X is a discrete random variable with counterdomain the countable set $\{x_1, ..., x_j, ..\}$

(ii) $var[X] = \int_{-\infty}^{\infty} (x_j - \mu_X)^2 f_X(x) dx$

if X is a continuous random variable with density function $f_X(x)$.

(iii) $var[X] = \int_0^{\infty} 2x[1 - F_X(x) + F_X(-x)]dx - \mu_X^2$

for an arbitrary random variable X.

The variances are defined only if the series in (i) is convergent or if the integrals in (ii) or (iii) exist. Again, the variance of a random variable is defined in terms of

the density function or cumulative distribution function of the random variable and consequently, variance can be defined in terms of these functions without reference to a random variable.

Notice that variance is a measure of spread since if the values of the random variable X tend to be far from their mean, the variance of X will be larger than the variance of a comparable random variable whose values tend to be near their mean. It is clear from (i), (ii) and (iii) that the variance is a nonnegative number.

If X is a random variable with variance σ_X^2, then the **standard deviation** of X, denoted by σ_X, is defined as $\sqrt{var(X)}$

The standard deviation of a random variable, like the variance, is a measure of spread or dispersion of the values of a random variable. In many applications it is preferable to the variance since it will have the same measurement units as the random variable itself.

EXAMPLE: Consider the experiment of tossing two dice. Let X denote the total of the upturned faces. Then for this case we have ($\mu_X = 7$):

$$var[X] = \sum_{i=2}^{12}(i - \mu_X)^2 f_X(i) = 210/36$$

EXAMPLE: Consider a X that can take only to possible values, 1 and -1, each with probability 0.5. Then the variance of X is ($\mu_X = 0$):

$$var[X] = 0.5 * 1^2 + 0.5 * (-1)^2 = 1$$

EXAMPLE: Consider a X that can take only to possible values, 10 and -10, each with probability 0.5. Then we have:

$$\mu_X = E[X] = 10 * 0.5 + (-10) * 0.5 = 0$$

$$var[X] = 0.5 * 10^2 + 0.5 * (-10)^2 = 100$$

Notice that in examples 2 and 3 the two random variables have the same mean but different variance, larger being the variance of the random variable with values further away from the mean.

EXAMPLE: Consider a continuous random variable X with density function $f_X(x) = \lambda e^{-\lambda x}$ for $x \in [0, \infty)$. Then ($\mu_X = 1/\lambda$):

$$var[X] = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)dx = \int_{0}^{\infty} (x-1/\lambda)^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

EXAMPLE: Consider a continuous random variable X with density function $f_X(x) = x^{-2}$ for $x \in [1, \infty)$. Then we know that the mean of X does not exist. Consequently, we can not define the variance.

Notice that

$$Var\left(X\right) = E\left[\left(X - E(X)\right)^2\right] = E\left(X^2\right) - E^2\left(X\right)$$

and that

$$Var\left(aX+b\right) = a^2 Var\left(X\right), \quad SD = \sqrt{Var}, \quad SD\left(aX+b\right) = |a|SD(X),$$

i.e., $SD(X)$ changes proportionally. Variance/standard deviation measures dispersion, higher variance more spread out. Interquartile range: $F_X^{-1}(3/4) - F_X^{-1}(1/4)$, the range of middle half always exists and is an alternative measure of dispersion.

3.0.6 Higher Moments of a Random Variable

If X is a random variable, the **$\mathbf{r^{th}}$ raw moment** of X, denoted by $\mu_r^{/}$, is defined as:

$$\mu_r^{/} = E[X^r]$$

if this expectation exists. Notice that $\mu_r^{/} = E[X] = \mu_1^{/} = \mu_X$, the mean of X.

If X is a random variable, the **$\mathbf{r^{th}}$ central moment** of X **about** α is defined as $E[(X-\alpha)^r]$. If $\alpha = \mu_X$, we have the **$\mathbf{r^{th}}$ central moment** of X about μ_X, denoted by μ_r, which is:

$$\mu_r = E[(X-\mu_X)^r]$$

We have measures defined in terms of quantiles to describe some of the characteristics of random variables or density functions. The **$\mathbf{q^{th}}$ quantile** of a random variable X or of its corresponding distribution is denoted by ξ_q and is defined as the smallest number ξ satisfying $F_X(\xi) \geq q$. If X is a continuous random variable, then the **$\mathbf{q^{th}}$ quantile** of X is given as the smallest number ξ satisfying $F_X(\xi) \geq q$.

The **median** of a random variable X, denoted by med_X or $med(X)$, or ξ_q, is the 0.5^{th} quantile. Notice that if X a continuous random variable the median of X satisfies:

$$\int_{-\infty}^{med(X)} f_X(x)dx = \frac{1}{2} = \int_{med(X)}^{\infty} f_X(x)dx$$

so the median of X is any number that has half the mass of X to its right and the other half to its left. The median and the mean are measures of central location.

The third moment about the mean $\mu_3 = E\left(X - E\left(X\right)\right)^3$ is called a measure of asymmetry, or **skewness**. Symmetrical distributions can be shown to have $\mu_3 = 0$. Distributions can be skewed to the left or to the right. However, knowledge of the third moment gives no clue as to the shape of the distribution, i.e. it could be the case that $\mu_3 = 0$ but the distribution to be far from symmetrical. The ratio $\frac{\mu_3}{\sigma^3}$ is unitless and is call the **coefficient of skewness**. An alternative measure of skewness is provided by the ratio: (mean-median)/(standard deviation)

The fourth moment about the mean $\mu_4 = E\left(X - E\left(X\right)\right)^4$ is used as a measure of **kurtosis**, which is a degree of flatness of a density near the center. The **coefficient of kurtosis** is defined as $\frac{\mu_4}{\sigma^4} - 3$ and positive values are sometimes used to indicate that a density function is more peaked around its center than the normal (leptokurtic distributions). A positive value of the coefficient of kurtosis is indicative for a distribution which is flatter around its center than the standard normal (platykurtic distributions). This measure suffers from the same failing as the measure of skewness i.e. it does not always measure what it supposed to.

While a particular moment or a few of the moments may give little information about a distribution the entire set of moments will determine the distribution exactly. In applied statistics the first two moments are of great importance, but the third and forth are also useful.

3.0.7 Moment Generating Functions

Finally we turn to the moment generating function (mgf) and characteristic Function (cf). The mgf is defined as

$$M_X(t) = E\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$$

for any real t, provided this integral exists in some neighborhood of 0. It is the Laplace transform of the function $f_X(\cdot)$ with argument $-t$. We have the useful inversion formula

$$f_X(x) = \int_{-\infty}^{\infty} M_X(t)\, e^{-tx} dt$$

The mgf is of limited use, since it does not exist for many r.v. the cf is applicable more generally, since it always exists:

$$\varphi_X(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx = \int_{-\infty}^{\infty} \cos(tx)\, f_X(x)dx + i\int_{-\infty}^{\infty} \sin(tx)\, f_X(x)dx$$

This essentially is the Fourier transform of the function $f_X(\cdot)$ and there is a well defined inversion formula

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t)\, dt$$

If X is symmetric about zero, the complex part of cf is zero. Also,

$$\frac{d^r}{dt^r}\varphi_X(0) = E\left(i^r X^r e^{itX}\right) \downarrow_{t=0} = i^r E\left(X^r\right), \quad r = 1, 2, 3, ..$$

Thus the moments of X are related to the derivative of the cf at the origin.

If

$$c(t) = \int_{-\infty}^{\infty} \exp(itx)\, dF(x)$$

notice that

$$\frac{d^r c(t)}{dt^r} = \int_{-\infty}^{\infty} (ix)^r \exp(itx)\, dF(x)$$

and

$$\left.\frac{d^r c(t)}{dt^r}\right|_{t=0} = \int_{-\infty}^{\infty} (ix)^r\, dF(x) = (i)^r \mu_r^{/} \Rightarrow \mu_r^{/} = (-i)^r \left.\frac{d^r c(t)}{dt^r}\right|_{t=0}$$

the r^{th} uncenterd moment. Now expanding $c(t)$ in powers of t we get

$$c(t) = c(0) + \left.\frac{d^r c(t)}{dt^r}\right|_{t=0} t + ... + \left.\frac{d^r c(t)}{dt^r}\right|_{t=0} \frac{(t)^r}{r!} + ... = 1 + \mu_1^{/}(it) + ... + \mu_r^{/}\frac{(it)^r}{r!} + ...$$

The cummulants are defined as the coefficients κ_1, $\kappa_2, ..., \kappa_r$ of the identity in it

$$\begin{aligned} \exp\left(\kappa_1(it) + \kappa_2\frac{(it)^2}{2!} + ... + \kappa_r\frac{(it)^r}{r!} + ...\right) &= 1 + \mu_1^{/}(it) + ... + \mu_r^{/}\frac{(it)^r}{r!} + ... \\ &= c(t) = \int_{-\infty}^{\infty} \exp(itx)\, dF(x) \end{aligned}$$

The cumulant-moment connection:

Suppose X is a random variable with n moments $a_1, ...a_n$. Then X has n cumulants $k_1, ...k_n$ and

$$a_{r+1} = \sum_{j=0}^{r} \binom{r}{j} a_j k_{r+1-j} \text{ for } r = 0, ..., n-1.$$

Writing out for $r = 0, ...3$ produces:

$$\begin{aligned} a_1 &= k_1 \\ a_2 &= k_2 + a_1 k_1 \\ a_3 &= k_3 + 2a_1 k_2 + a_2 k_1 \\ a_4 &= k_4 + 3a_1 k_3 + 3a_2 k_2 + a_3 k_1. \end{aligned}$$

These recursive formulas can be used to calculate the $a's$ efficiently from the $k's$, and vice versa. When X has mean 0, that is, when $a_1 = 0 = k_1$, a_j becomes

$$\mu_j = E((X - E(X))^j),$$

so the above formulas simplify to:

$$\begin{aligned} \mu_2 &= k_2 \\ \mu_3 &= k_3 \\ \mu_4 &= k_4 + 3k_2^2. \end{aligned}$$

3.0.8 Expectations of Functions of Random Variablers

Product and Quotient

Let $f(X,Y) = \frac{X}{Y}$, $E(X) = \mu_X$ and $E(Y) = \mu_Y$. Then, expanding $f(X,Y) = \frac{X}{Y}$ around (μ_X, μ_Y), we have

$$f(X,Y) = \frac{\mu_X}{\mu_Y} + \frac{1}{\mu_Y}(X-\mu_X) - \frac{\mu_X}{(\mu_Y)^2}(Y-\mu_Y) + \frac{\mu_X}{(\mu_Y)^3}(Y-\mu_Y)^2 - \frac{1}{(\mu_Y)^2}(X-\mu_X)(Y-\mu_Y)$$

as $\frac{\partial f}{\partial X} = \frac{1}{Y}$, $\frac{\partial f}{\partial Y} = -\frac{X}{Y^2}$, $\frac{\partial^2 f}{\partial X^2} = 0$, $\frac{\partial^2 f}{\partial X \partial Y} = \frac{\partial^2 f}{\partial Y \partial X} = -\frac{1}{Y^2}$, and $\frac{\partial^2 f}{\partial Y^2} = 2\frac{X}{Y^3}$. Taking expectations we have

$$E\left(\frac{X}{Y}\right) = \frac{\mu_X}{\mu_Y} + \frac{\mu_X}{(\mu_Y)^3} Var(Y) - \frac{1}{(\mu_Y)^2} Cov(X,Y).$$

For the variance, take again the variance of the Taylor expansion and keeping only terms up to order 2 we have:

$$Var\left(\frac{X}{Y}\right) = \frac{(\mu_X)^2}{(\mu_Y)^2}\left[\frac{Var(X)}{(\mu_X)^2} + \frac{Var(Y)}{(\mu_Y)^2} - 2\frac{Cov(X,Y)}{\mu_X \mu_Y}\right].$$

Chapter 4

EXAMPLES OF PARAMETRIC UNIVARIATE DISTRIBUTIONS

A parametric family of density functions is a collection of density functions that are indexed by a quantity called parameter, e.g. let $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x > 0$ and some $\lambda > 0$. λ is the parameter, and as λ ranges over the positive numbers, the collection $\{f(.;\lambda) : \lambda > 0\}$ is a parametric family of density functions.

4.0.9 Discrete Distributions

UNIFORM:

Suppose that for $j = 1, 2, 3, ..., n$

$$P(X = x_j|\mathcal{X}) = \frac{1}{n}$$

where $\{x_1, x_2, ...x_n\} = \mathcal{X}$ is the support. Then

$$E(X) = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad Var(X) = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2.$$

The c.d.f. here is

$$P(X \leq x) = \frac{1}{n}\sum_{j=1}^{n} 1(x_j \leq x)$$

Bernoulli

A random variable whose outcome have been classified into two categories, called "success" and "failure", represented by the letters s and f, respectively, is called a Bernoulli trial. If a random variable X is defined as 1 if a Bernoulli trial results in

success and 0 if the same Bernoulli trial results in failure, then X has a Bernoulli distribution with parameter $p = P[success]$. The definition of this distribution is:

A random variable X has a Bernoulli distribution if the discrete density of X is given by:

$$f_X(x) = f_X(x;p) = \begin{cases} p^x(1-p)^{1-x} & for \quad x = 0,1 \\ 0 & otherwise \end{cases}$$

where $p = P[X = 1]$. For the above defined random variable X we have that:

$$E[X] = p \qquad and \qquad var[X] = p(1-p)$$

BINOMIAL:

Consider a random experiment consisting of n repeated independent Bernoulli trials with p the probability of success at each individual trial. Let the random variable X represent the number of successes in the n repeated trials. Then X follows a Binomial distribution. The definition of this distribution is:

A random variable X has a **binomial** distribution, $X \sim Binomial(n,p)$, if the discrete density of X is given by:

$$f_X(x) = f_X(x;n,p) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} p^x(1-p)^{n-x} & for \quad x = 0,1,...,n \\ 0 & otherwise \end{cases}$$

where $p = P[X = 1]$ i.e. the probability of success in each independent Bernoulli trial and n is the total number of trials. For the above defined random variable X we have that:

$$E[X] = np \qquad and \qquad var[X] = np(1-p)$$

Mgf

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

EXAMPLE: Consider a stock with value $S = 50$. Each period the stock moves up or down, independently, in discrete steps of 5. The probability of going up is

$p = 0.7$ and down $1 - p = 0.3$. What is the expected value and the variance of the value of the stock after 3 period?

If we call X the random variable which is a success if the stock moves up and failure if the stock moves down. Then $P[X = success] = P[X = 1] = 0.7$, and $X \tilde{} Binomial(3, p)$. Now X can take the values $0, 1, 2, 3$ i.e. no success, 1 success and 2 failures, etc.. The value of the stock in each case and the probabilities are:

$$S = 35, \text{ and } f_X(0) = \begin{pmatrix} 3 \\ 0 \end{pmatrix} p^0(1-p)^{3-0} = 1 * 0.3^3 = 0.027,$$

$$S = 45, \text{ and } f_X(1) = \begin{pmatrix} 3 \\ 1 \end{pmatrix} p^1(1-p)^{3-1} = 3 * 0.7 * 0.3^2 = 0.189,$$

$$S = 55, \text{ and } f_X(2) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} p^2(1-p)^{3-2} = 3 * 0.7^2 * 0.3 = 0.441,$$

$$S = 65 \text{ and } f_X(3) = \begin{pmatrix} 3 \\ 3 \end{pmatrix} p^3(1-p)^{3-3} = 1 * 0.7^3 = 0.343.$$

Hence the expected stock value is:

$E[S] = 35 * 0.027 + 45 * 0.189 + 55 * 0.441 + 65 * 0.343 = 56$, and $var[S] = (35 - 56)^2 * 0.027 + (-11)^2 * 0.189 + (-1)^2 * 0.441 + (9)^2 * 0.343$.

Hypergeometric

Let X denote the number of defective balls in a sample of size n when sampling is done **without** replacement from a box containing M balls out of which K are defective. The X has a hypergeometric distribution. The definition of this distribution is:

A random variable X has a **hypergeometric** distribution if the discrete den-

sity of X is given by:

$$f_X(x) = f_X(x; M, K, n) = \begin{cases} \dfrac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where M is a positive integer, K is a nonnegative that is at most M, and n is a positive integer that is at most M. For this distribution we have that:

$$E[X] = n\frac{K}{M} \qquad and \qquad var[X] = n\frac{K}{M}\frac{M-K}{M}\frac{M-n}{M-1}$$

Notice the difference of the binomial and the hypergeometric i.e. for the binomial distribution we have Bernoulli trials i.e. independent trials with fixed probability of success or failure, whereas in the hypergeometric in each trial the probability of success or failure changes depending on the result.

Geometric

Consider a sequence of independent Bernoulli trials with p equal the probability of success on an individual trial. Let the random variable X represent the number of trials required before the first success. Then X has a geometric distribution. The definition of this distribution is: A random variable X has a **geometric** distribution, $X \sim geometric(p)$, if the discrete density of X is given by:

$$f_X(x) = f_X(x; p) = \begin{cases} p(1-p)^x & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where p is the probability of success in each Bernoulli trial. For this distribution we have that:

$$E[X] = \frac{1-p}{p} \qquad and \qquad var[X] = \frac{1-p}{p^2}$$

It is worth noticing that the Binomial distribution $Binomial(n,p)$ can be approximated by a $Poisson(np)$ (see below). The approximation is more valid as $n \to \infty$, $p \to 0$, in such a way so that $np = constant$.

POISSON:

A random variable X has a **Poisson** distribution, $X \sim Poisson(\lambda)$, if the discrete density of X is given by:

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, 3, ...$$

In calculations with the Poisson distribution we may use the fact that

$$e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!} \quad for \quad any \quad t.$$

Employing the above we can prove that

$$E(X) = \lambda, \quad E(X(X-1)) = \lambda^2, \quad Var(X) = \lambda.$$

The Poisson distribution provides a realistic model for many random phenomena. Since the values of a Poisson random variable are nonnegative integers, any random phenomenon for which a count of some sort is of interest is a candidate for modeling in assuming a Poisson distribution. Such a count might be the number of fatal traffic accidents per week in a given place, the number of telephone calls per hour, arriving in a switchboard of a company, the number of pieces of information arriving per hour, etc.

EXAMPLE: It is known that the average number of daily changes in excess of 1%, for a specific stock Index, occurring in each six-month period is 5. What is the probability of having one such a change within the next 6 months? What is the probability of at least 3 changes within the same period?

We model the number of in excess of 1% changes, X, within the next 6 months as a Poisson process. We know that $E[X] = \lambda = 5$. Hence $f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}5^x}{x!}$,

for $x = 0, 1, 2, , ...$ Then $P[X = 1] = f_X(1) = \frac{e^{-5}5^1}{1!} = 0.0337$. Also $P[X \geq 3] = 1 - P[X < 3] =$

$$= 1 - P[X = 0] - P[X = 1] - P[X = 2] =$$
$$= 1 - \frac{e^{-5}5^0}{0!} - \frac{e^{-5}5^1}{1!} - \frac{e^{-5}5^2}{2!} = 0.875.$$

We can approximate the Binomial with Poisson. The approximation is better the smaller the p and the larger the n.

4.0.10 Continuous Distributions

UNIFORM ON $[a, b]$.

A very simple distribution for a continuous random variable is the uniform distribution. Its density function is:

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & if \quad x \in [a,b] \\ 0 & otherwise \end{cases},$$

and

$$F(x|a,b) = \int_a^x f(z|a,b)\, dz = \frac{x-a}{b-a},$$

where $-\infty < a < b < \infty$. Then the random variable X is defined to be **uniformly** distributed over the interval $[a, b]$. Now if X is uniformly distributed over $[a, b]$ then

$$E(X) = \frac{a+b}{2}, \quad median = \frac{a+b}{2}, \quad Var(X) = \frac{(b-a)^2}{12}.$$

If $X \backsim U[a,b] \Longrightarrow X - a \backsim U[0, b-a] \Longrightarrow \frac{X-a}{b-a} \backsim U[0, b-a]$. Notice that if a random variable is uniformly distributed over one of the following intervals $[a, b)$, $(a, b]$, (a, b) the density function, expected value and variance does not change.

Exponential Distribution

If a random variable X has a density function given by:

$$f_X(x) = f_X(x;\lambda) = \lambda e^{-\lambda x} \qquad for \quad 0 \leq x < \infty$$

where $\lambda > 0$ then X is defined to have an (negative) exponential distribution. Now this random variable X we have

$$E[X] = \frac{1}{\lambda} \qquad and \qquad var[X] = \frac{1}{\lambda^2}$$

Pareto-Levy or Stable Distributions

The stable distributions are a natural generalization of the normal in that, as their name suggests, they are stable under addition, i.e. a sum of stable random variables is also a random variable of the same type. However, nonnormal stable distributions have more probability mass in the tail areas than the normal. In fact, the nonnormal stable distributions are so fat-tailed that their variance and all higher moments are infinite.

Closed form expressions for the density functions of stable random variables are available for only the cases of normal and Cauchy.

If a random variable X has a density function given by:

$$f_X(x) = f_X(x; \gamma, \delta) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - \delta)^2} \qquad for \quad -\infty < x < \infty$$

where $-\infty < \delta < \infty$ and $0 < \gamma < \infty$, then X is defined to have a **Cauchy** distribution. Notice that for this random variable even the mean is infinite.

Normal or Gaussian:

We say that $X \backsim N[\mu, \sigma^2]$ then

$$f\left(x|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$
$$E(X) = \mu, \quad Var(X) = \sigma^2.$$

The distribution is symmetric about μ, it is also unimodal and positive everywhere. Notice

$$\frac{X - \mu}{\sigma} = Z \backsim N[0, 1]$$

is the standard normal distribution.

Lognormal Distribution

Let X be a positive random variable, and let a new random variable Y be defined as $Y = \log X$. If Y has a normal distribution, then X is said to have a lognormal distribution. The density function of a lognormal distribution is given by

$$f_X(x;\mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}} \qquad for \quad 0 < x < \infty$$

where μ and σ^2 are parameters such that $-\infty < \mu < \infty$ and $\sigma^2 > 0$. We haven

$$E[X] = e^{\mu+\frac{1}{2}\sigma^2} \qquad and \qquad var[X] = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$$

Notice that if X is lognormally distributed then

$$E[\log X] = \mu \qquad and \qquad var[\log X] = \sigma^2$$

Gamma-χ^2

$$f\left(x|\alpha,\beta\right) = \frac{1}{\Gamma\left(\alpha\right)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, \quad 0 < x < \infty, \quad \alpha,\beta > 0$$

α is shape parameter, β is a scale parameter. Here $\Gamma\left(\alpha\right) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt$ is the Gamma function, $\Gamma\left(n\right) = n!$. The χ^2_k is when $\alpha = k$, and $\beta = 1$.

Notice that we can approximate the Poisson and Binomial functions by the normal, in the sense that if a random variable X is distributed as Poisson with parameter λ, then $\frac{X-\lambda}{\sqrt{\lambda}}$ is distributed approximately as standard normal. On the other hand if $Y \sim Binomial(n,p)$ then $\frac{Y-np}{\sqrt{np(1-p)}} \sim N(0,1)$.

The standard normal is an important distribution for another reason, as well. Assume that we have a sample of n independent random variables, $x_1, x_2, ..., x_n$, which are coming from the same distribution with mean m and variance s^2, then we have the following:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{x_i - m}{s} \sim N(0,1)$$

This is the well known **Central Limit Theorem** for independent observations.

4.1 Multivariate Random Variables

We now consider the extension to multiple r.v., i.e.,

$$X = (X_1, X_2, .., X_k) \in \mathbb{R}^k$$

The joint pmf, $f_X(x)$, is a function with

$$P(X \in A) = \sum_{x \in A} f_X(x)$$

The joint pdf, $f_X(x)$, is a function with

$$P(X \in A) = \int_{x \in A} f_X(x)dx$$

This is a multivariate integral, and in general difficult to compute. If A is a rectangle $A = [a_1, b_1] \times ... \times [a_k, b_k]$, then

$$\int_{x \in A} f_X(x)dx = \int_{a_k}^{b_k} ... \int_{a_1}^{b_1} f_X(x)dx_1..dx_k$$

The joint c.d.f. is defined similarly

$$F_X(x) = \sum_{z_1 \leq x_1, ..., z_k \leq x_k} f_X(z_1, z_2, ..., z_k)$$

$$F_X(x) = P(X_1 \leq x_1, ..., X_k \leq x_k) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_X(z_1, z_2, ..., z_k)dz_1..dz_k$$

The multivariate c.d.f. has similar coordinate-wise properties to a univariate c.d.f. For continuously differentiable c.d.f.'s

$$f_X(x) = \frac{\partial^k F_X(x)}{\partial x_1 \partial x_2..\partial x_k}$$

4.1.1 Conditional Distributions and Independence

We defined conditional probability $P(A|B) = P(A \cap B)/P(B)$ for events with $P(B) \neq 0$. We now want to define conditional distributions of $Y|X$. In the discrete case there is no problem

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

when the event $\{X = x\}$ has nonzero probability. Likewise we can define

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \frac{\sum_{Y \leq y} f(y,x)}{f_X(x)}$$

Note that $f_{Y|X}(y|x)$ is a density function and $F_{Y|X}(y|x)$ is a c.d.f.

1) $f_{Y|X}(y|x) \geq 0$ for all y

2) $\sum_y f_{Y|X}(y|x) = \frac{\sum_y f(y,x)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1$

In the continuous case, it appears a bit anomalous to talk about the $P(y \in A|X = x)$, since $\{X = x\}$ itself has zero probability of occurring. Still, we define the conditional density function

$$f_{Y|X}(y|x) = \frac{f(y,x)}{f_X(x)}$$

in terms of the joint and marginal densities. It turns out that $f_{Y|X}(y|x)$ has the properties of p.d.f.

1) $f_{Y|X}(y|x) \geq 0$

2) $\int_{-\infty}^{\infty} f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} f(y,x)dy}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$

We can define Expectations within the conditional distribution

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} y f(y,x)dy}{\int_{-\infty}^{\infty} f(y,x)dy}$$

and higher moments of the conditional distribution

4.1.2 Independence

We say that Y and X are independent (denoted by $\perp\!\!\!\perp$) if

$$P(Y \in A, X \in B) = P(Y \in A)P(X \in B)$$

for all events A, B, in the relevant sigma-algebras. This is equivalent to the cdf's version which is simpler to state and apply.

$$F_{YX}(y,x) = F(y)F(x)$$

In fact, we also work with the equivalent density version

$$\begin{aligned} f(y,x) &= f(y)f(x) \quad for \quad all \quad y,x \\ f_{Y|X}(y|x) &= f(y) \quad for \quad all \quad y \\ f_{X|Y}(x|y) &= f(x) \quad for \quad all \quad x \end{aligned}$$

If $Y \perp\!\!\!\perp X$, then $g(X) \perp\!\!\!\perp h(Y)$ for any measurable functions g, and h.

We can generalise the notion of independence to multiple random variables. Thus Y, X, and Z are mutually independent if:

$$\begin{aligned} f(y,x,z) &= f(y)f(x)f(z) \\ f(y,x) &= f(y)f(x) \quad for \quad all \quad y,x \\ f(x,z) &= f(x)f(z) \quad for \quad all \quad x,z \\ f(y,z) &= f(y)f(z) \quad for \quad all \quad y,z \end{aligned}$$

for all y, x, z.

4.1.3 Examples of Multivariate Distributions

Multivariate Normal

We say that $X\left(X_1, X_2, ..., X_k\right) \backsim MVN_k\left(\mu, \Sigma\right)$, when

$$f_X\left(x|\mu,\Sigma\right) = \frac{1}{\left(2\pi\right)^{k/2}\left[\det\left(\Sigma\right)\right]^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^{/}\Sigma^{-1}\left(x-\mu\right)\right)$$

where Σ is a $k \times k$ covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1k} \\ & \ddots & & \vdots \\ & & \ddots & \vdots \\ & & & \sigma_{kk} \end{pmatrix}$$

and $\det\left(\Sigma\right)$ is the determinant of Σ.

Theorem 8 *(a) If* $X \backsim MVN_k(\mu, \Sigma)$ *then* $X_i \backsim N(\mu_i, \sigma_{ii})$ *(this is shown by integration of the joint density with respect to the other variables).*
(b) The conditional distributions $X = (X_1, X_2)$ *are Normal too*

$$f_{X_1|X_2}(x_1|x_2) \backsim N\left(\mu_{X_1|X_2}, \Sigma_{X_1|X_2}\right)$$

where

$$\mu_{X_1|X_2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Sigma_{X_1|X_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

(c) Iff Σ *diagonal then* $X_1, X_2, .., X_k$ *are mutually independent. In this case*

$$\begin{aligned} \det(\Sigma) &= \sigma_{11}\sigma_{22}..\sigma_{kk} \\ -\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu) &= -\frac{1}{2}\sum_{j=1}^{k}\frac{(x_j-\mu_j)^2}{\sigma_{jj}} \end{aligned}$$

so that

$$f_X(x|\mu, \Sigma) = \prod_{j=1}^{k}\frac{1}{\sqrt{2\pi\sigma_{jj}}}\exp\left(-\frac{1}{2}\frac{(x_j-\mu_j)^2}{\sigma_{jj}}\right)$$

4.1.4 More on Conditional Distributions

We now consider the relationship between two, or more, r.v. when they are not independent. In this case, conditional density $f_{Y|X}$ and c.d.f. $F_{Y|X}$ is in general varying with the conditioning point x. Likewise for conditional mean $E(Y|X)$, conditional median $M(Y|X)$, conditional variance $V(Y|X)$, conditional cf $E\left(e^{itY}|X\right)$, and other functionals, all of which characterize the relationship between Y and X. Note that this is a directional concept, unlike covariance, and so for example $E(Y|X)$ can be very different from $E(X|Y)$.

Regression Models:

We start with random variable (Y, X). We can write for any such random variable

$$Y = \overbrace{\underbrace{E(Y|X)}_{systematic\ \ part}}^{m(X)} + \overbrace{\underbrace{Y - E(Y|X)}_{\text{rand}\,om\ \ part}}^{\varepsilon}$$

By construction ε satisfies $E(\varepsilon|X) = 0$, but ε is not necessarily independent of X. For example, $Var(\varepsilon|X) = Var(Y - E(Y|X)|X) = Var(Y|X) = \sigma^2(X)$ can be expected to vary with X as much as $m(X) = E(Y|X)$. A convenient and popular simplification is to assume that

$$\begin{aligned} E(Y|X) &= \alpha + \beta X \\ Var(Y|X) &= \sigma^2 \end{aligned}$$

For example, in the bivariate normal distribution $Y|X$ has

$$\begin{aligned} E(Y|X) &= \mu_Y + \rho_{YX}\frac{\sigma_Y}{\sigma_X}(X - \mu_X) \\ Var(Y|X) &= \sigma_Y^2\left(1 - \rho_{YX}^2\right) \end{aligned}$$

and in fact $\varepsilon \perp\!\!\!\perp X$.

We have the following result about conditional expectations

Theorem 9 *(1)* $E(Y) = E[E(Y|X)]$
(2) $E(Y|X)$ *minimizes* $E\left[(Y - g(X))^2\right]$ *over all measurable functions* $g(\cdot)$
(3) $Var(Y) = E[Var(Y|X)] + Var[E(Y|X)]$

Proof. (1) Write $f_{YX}(y,x) = f_{Y|X}(y|x) f_X(x)$ then we have $E(Y) = \int y f_Y(y)dy = \int y\left(\int f_{YX}(y,x)dx\right)dy = \int y\left(\int f_{Y|X}(y|x) f_X(x)\,dx\right)dy =$

$= \int\left(\int y f_{Y|X}(y|x)\,dy\right) f_X(x)\,dx = \int [E(Y|X = x] f_X(x)\,dx = E(E(Y|X))$

(2) $E\left[(Y - g(X))^2\right] = E\left[[Y - E(Y|X) + E(Y|X) - g(X)]^2\right]$

$= E[Y - E(Y|X)]^2 + 2E[[Y - E(Y|X)][E(Y|X) - g(X)]] + E[E(Y|X) - g(X)]^2$

as now $E(YE(Y|X)) = E\left[(E(Y|X))^2\right]$, and $E(Yg(X)) = E(E(Y|X)g(X))$ we get that $E\left[(Y - g(X))^2\right] = E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \geq E[Y - E(Y|X)]^2$.

(3) $Var(Y) = E[Y - E(Y)]^2 = E[Y - E(Y|X)]^2 + E[E(Y|X) - E(Y)]^2$

$+2E[[Y - E(Y|X)][E(Y|X) - E(Y)]]$

The first term is $E[Y - E(Y|X)]^2 = E\{E\left[[Y - E(Y|X)]^2|X\right]\} = E[Var(Y|X)]$

The second term is $E[E(Y|X) - E(Y)]^2 = Var[E(Y|X)]$

The third term is zero as $\varepsilon = Y - E(Y|X)$ is such that $E(\varepsilon|X) = 0$, and $E(Y|X) - E(Y)$ is measurable with respect to X. ■

Covariance

$$Cov\left(X,Y\right)=E\left[X-E\left(X\right)\right]E\left[Y-E\left(Y\right)\right]=E\left(XY\right)-E\left(X\right)E\left(Y\right)$$

Note that if X or Y is a constant then $Cov\left(X,Y\right)=0$. Also

$$Cov\left(aX+b,cY+d\right)=acCov\left(X,Y\right)$$

An alternative measure of association is given by the **correlation coefficient**

$$\rho_{XY}=\frac{Cov\left(X,Y\right)}{\sigma_X\sigma_Y}$$

Note that

$$\rho_{aX+b,cY+d}=sign\left(a\right)\times sign\left(c\right)\times\rho_{XY}$$

If $E\left(Y|X\right)=a=E\left(Y\right)$ almost surely, then $Cov\left(X,Y\right)=0$. Also if X and Y are independent r.v. then $Cov\left(X,Y\right)=0$.

Both the covariance and the correlation of random variables X and Y are measures of a linear relationship of X and Y in the following sense. $cov[X,Y]$ will be positive when $(X-\mu_X)$ and $(Y-\mu_Y)$ tend to have the same sign with high probability, and $cov[X,Y]$ will be negative when $(X-\mu_X)$ and $(Y-\mu_Y)$ tend to have opposite signs with high probability. The actual magnitude of the $cov[X,Y]$ does not much meaning of how strong the linear relationship between X and Y is. This is because the variability of X and Y is also important. The correlation coefficient does not have this problem, as we divide the covariance by the product of the standard deviations. Furthermore, the correlation is unitless and $-1\leq\rho\leq1$.

The properties are very useful for evaluating the **expected return** and **standard deviation** of a **portfolio**. Assume r_a and r_b are the returns on assets A and B, and their variances are σ_a^2 and σ_b^2, respectively. Assume that we form a portfolio of the two assets with weights w_a and w_b, respectively. If the correlation of the returns of these assets is ρ, find the expected return and standard deviation of the portfolio.

If R_p is the return of the portfolio then $R_p = w_a r_a + w_b r_b$. The expected portfolio return is $E[R_p] = w_a E[r_a] + w_b E[r_b]$. The variance of the portfolio is $var[R_p] = var[w_a r_a + w_b r_b] = E[(w_a r_a + w_b r_b)^2] - (E[w_a r_a + w_b r_b])^2 =$

$= w_a^2 E[r_a^2] + w_b^2 E[r_b^2] + 2w_a w_b E[r_a r_b]$

$-w_a^2 (E[r_a])^2 - w_b^2 (E[r_b])^2 - 2w_a w_b E[r_a]E[r_b] =$

$= w_a^2 \{E[r_a^2] - (E[r_a])^2\} + w_b^2 \{E[r_b^2] - (E[r_b])^2\} + 2w_a w_b \{E[r_a r_b] - E[r_a]E[r_b]\}$

$= w_a^2 var[r_a] + w_b^2 var[r_b] + 2w_a w_b cov[r_a, r_b]$ or $= w_a^2 \sigma_a^2 + w_b^2 \sigma_b^2 + 2w_a w_b \rho \sigma_a \sigma_b$

In a vector format we have:

$$E[R_p] = \begin{pmatrix} w_a & w_b \end{pmatrix} \begin{pmatrix} E[r_a] \\ E[r_b] \end{pmatrix} \text{ and}$$

$$var[R_p] = \begin{pmatrix} w_a & w_b \end{pmatrix} \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \begin{pmatrix} w_a \\ w_b \end{pmatrix}$$

From the above example we can see that $var[aX+bY] = a^2 var[X] + b^2 var[Y] + 2abcov[X,Y]$ for random variables X and Y and constants a and b. In fact we can generalize the formula above for several random variables $X_1, X_2, ..., X_n$ and constants $a_1, a_2, a_3, ..., a_n$ i.e. $var[a_1 X_1 + a_2 X_2 + ...a_n X_n] = \sum_{i=1}^{n} a_i^2 var[X_i] + 2\sum_{i<j}^{n} a_i a_j cov[X_i, X_j]$

4.2 Inequalities

This section gives some inequalities that are useful in establishing a variety of probabilistic results.

4.2.1 Markov

Let Y be a random variable and consider a function $g(.)$ such that $g(y) \geq 0$ for all $y \in \mathbb{R}$. Assume that $E[g(Y)]$ exists. Then

$$P[g(Y) \geq c] \leq c^{-1} E[g(Y)], \quad for \quad all \quad c > 0.$$

PROOF:

Assume that Y is continuous random variable (the discrete case follows anal-

ogously) with p.d.f. $f(.)$. Define $A_1 = \{y|g(y) \geq c\}$ and $A_2 = \{y|g(y) < c\}$. Then

$$\begin{aligned} E[g(Y)] &= \int_{A_1} g(y) f(y) dy + \int_{A_2} g(y) f(y) dy \\ &\geq \int_{A_1} g(y) f(y) dy \geq \int_{A_1} cf(y) dy = cP[g(Y) \geq c]. \end{aligned}$$

■

4.2.2 Chebychev's Inequality

$$P[|X - E(X)| \geq \eta] \leq \frac{Var(X)}{\eta^2}$$

or alternatively

$$P\left[|X - E(X)| \geq r\sqrt{Var(X)}\right] \leq \frac{1}{r^2}$$

PROOF:

To prove the above, assume that $E(X) = 0$ and compare $1(|X| \geq \eta)$ with $\frac{X^2}{\eta^2}$. Clearly $1(|X| \geq \eta) \leq \frac{X^2}{\eta^2}$ and it follows that $E[1(|X| \geq \eta)] \leq \frac{E(X^2)}{\eta^2} \Rightarrow P[|X| \geq \eta] \leq \frac{Var(X)}{\eta^2}$. Alternatively, apply Markov's inequality by setting $g(y) = [x - E(X)]^2$ and $c = r^2 Var(X)$.■

4.2.3 Minkowski

Let Y and Z be random variables such that $[E(|Y|^\alpha)] < \infty$ and $[E(|Z|^\alpha)] < \infty$ for some $1 \leq \alpha < \infty$. Then

$$[E(|Y + Z|^\alpha)]^{1/\alpha} \leq [E(|Y|^\alpha)]^{1/\alpha} + [E(|Z|^\alpha)]^{1/\alpha}$$

For $\alpha = 1$ we have the triangular inequality

4.2.4 Triangle

$$E|X + Y| \leq E|X| + E|Y|.$$

4.2.5 Cauchy-Schwarz

$$\begin{aligned} E^2(XY) &\leq E(X)^2 E(Y)^2 \\ (\sum a_j b_j)^2 &\leq \left(\sum a_j^2\right)\left(\sum b_j^2\right) \end{aligned}$$

PROOF:

Let $0 \leq h(t) = E\left[(tX - Y)^2\right] = t^2 E(X^2) + E(Y^2) - 2tE(XY)$. Then the function $h(t)$ is a quadratic function in t which is increasing as $t \to \pm\infty$. It has a unique minimum at $h^{/}(t) = 0 \Rightarrow 2tE(X^2) - 2E(XY) = 0 \Rightarrow t = \frac{E(XY)}{E(X^2)}$. Hence $0 \leq h\left(\frac{E(XY)}{E(X^2)}\right) \Rightarrow E^2(XY) \leq E(X)^2 E(Y)^2$.■

4.2.6 Hölder's Inequality

For any p, q satisfying $\frac{1}{p} + \frac{1}{q} = 1$ we have

$$E|XY| \leq (E|X|^p)^{\frac{1}{p}} (E|Y|^q)^{\frac{1}{q}}$$

In fact the Cauchy-Schwarz inequality corresponds for $p = q = 2$.

4.2.7 Jensen Inequality

Let X be a random variable with mean $E[X]$, and let $g(.)$ be a convex function. Then

$$E[g(X)] \geq g(E[X]).$$

Now a continuous function $g(.)$ with domain and counterdomain the real line is called **convex** if for any x_0 on the real line, there exist a line which goes through the point $(x_0, g(x_0))$ and lies on or under the graph of the function $g(.)$. Also if $g^{//}(x_0) \geq 0$ then $g(.)$ is convex.

Part II

Statistical Inference

Chapter 5

SAMPLING THEORY

To proceed we shall recall the following definitions.

Let $X_1, X_2, ..., X_k$ be k random variables all defined on the same probability space $(\Omega, \mathcal{A}, P[.])$. The **joint cumulative distribution function** of $X_1, X_2, ..., X_k$, denoted by $F_{X_1,X_2,...X_n}(\bullet, \bullet, ..., \bullet)$, is defined as

$$F_{X_1,X_2,...X_k}(x_1, x_2, ..., x_k) = P[X_1 \leq x_1; X_2 \leq x_2; ...; X_k \leq x_k]$$

for all $(x_1, x_2, ..., x_k)$.

Let $X_1, X_2, ..., X_k$ be k discrete random variables, then the **joint discrete density function** of these, denoted by $f_{X_1,X_2,...X_k}(\bullet, \bullet, ..., \bullet)$, is defined to be

$$f_{X_1,X_2,...X_k}(x_1, x_2, ..., x_k) = P[X_1 = x_1; X_2 = x_2; ...; X_k = x_k]$$

for $(x_1, x_2, ..., x_k)$, a value of $(X_1, X_2, ..., X_k)$ and is 0 otherwise.

Let $X_1, X_2, ..., X_k$ be k continuous random variables, then the **joint continuous density function** of these, denoted by $f_{X_1,X_2,...X_k}(\bullet, \bullet, ..., \bullet)$, is defined to be a function such that

$$F_{X_1,X_2,...X_k}(x_1, x_2, ..., x_k) = \int_{-\infty}^{x_k} ... \int_{-\infty}^{x_1} f_{X_1,X_2,...X_k}(u_1, u_2, ..., u_k) du_1 .. du_k$$

for all $(x_1, x_2, ..., x_k)$.

The totality of elements which are under discussion and about which information is desired will be called the **target population**. The statistical problem is

to find out something about a certain target population. It is generally impossible or impractical to examine the entire population, but one may examine a part of it (a sample from it) and, on the basis of this limited investigation, make inferences regarding the entire target population.

The problem immediately arises as to how the sample of the population should be selected. Of practical importance is the case of a simple random sample, usually called a random sample, which can be defined as follows:

Let the random variables $X_1, X_2, ..., X_n$ have a joint density $f_{X_1,X_2,...X_n}(x_1, x_2, ..., x_n)$ that factors as follows:

$$f_{X_1,X_2,...X_n}(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n)$$

where $f(.)$ is the common density of each X_i. Then $X_1, X_2, ..., X_n$ is defined to be a **random sample** of size n from a population with density $f(.)$. Note that identical distribution can be weakened - could have different population for each j - reflecting heterogeneous individuals. Also, in time series we might want to allow dependence, i.e., X_j and X_k are dependent. When we are dealing with finite population, sampling without replacement causes some heterogeneity since if $X_1 = x_1$, then the distribution of X_2 must be affected.

5.1 Sample Statistics

A **sample statistic** is a function of observable random variables, which is itself an observable random variable, which does not contain any unknown parameters, i.e. a sample statistic is any quantity we can write as a measurable function, $T(X_1, ..., X_n)$. For example, let $X_1, X_2, ..., X_k$ be a random sample from the density $f(.)$. Then the r^{th} **sample moment**, denoted by $M_r^{/}$, is defined as:

$$M_r^{/} = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

In particular, if $r = 1$, we get the sample mean, which is usually denoted by $\overline{X}$ or$\overline{X}_n$; that is:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Also the r^{th} **sample central moment (about** $\overline{X}_n$**)**, denoted by M_r, is defined as:

$$M_r = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^r.$$

In particular, if $r = 2$, we get the sample variance, and the sample standard deviation

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad s = \sqrt{s^2}$$

or maybe another sample statistic for the variance,

$$s_*^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

We can also get the sample Median,

$$\overline{M} = median\left\{X_1, ..., X_n\right\} = \begin{cases} X_{(r)} & if\ n = 2r-1 \\ \frac{1}{2}\left[X_{(r)} + X_{(r+1)}\right] & if\ n = 2r \end{cases}$$

the empirical cumulative distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \leq x\right)$$

$$\varphi_n(t) = \frac{1}{n}\sum_{i=1}^{n} e^{itX_i} = \frac{1}{n}\sum_{i=1}^{n}\sin\left(tX_i\right) + i\frac{1}{n}\sum_{i=1}^{n}\cos\left(tX_i\right)$$

These are analogous of corresponding population characteristics and will be shown to be similar to them when n is large. We calculate the properties of these variables: (1) Exact properties; (2) Asymptotic.

5.2 Means and Variances

We can prove the following theorems:

Theorem 10 *Let $X_1, X_2, ..., X_k$ be a random sample from the density $f(.)$. The expected value of the r^{th} sample moment is equal to the r^{th} population moment, i.e. the r^{th} sample moment is an unbiased estimator of the r^{th} population moment (Proof omitted).*

Theorem 11 *Let $X_1, X_2, ..., X_n$ be a random sample from a density $f(.)$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then*

$$E[\overline{X}_n] = \mu \qquad and \qquad var[\overline{X}_n] = \frac{1}{n}\sigma^2$$

where μ and σ^2 are the mean and variance of $f(.)$, respectively. Notice that this is true for any distribution $f(.)$, provided that is not infinite.

PROOF

$E[\overline{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$. Also

$var[\overline{X}_n] = var[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} var[X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2$ ■

Theorem 12 *Let $X_1, X_2, ..., X_n$ be a random sample from a density $f(.)$, and let s_*^2 defined as above. Then*

$$E[s_*^2] = \sigma^2 \qquad and \qquad var[s_*^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right)$$

where σ^2 and μ_4 are the variance and the 4^{th} central moment of $f(.)$, respectively. Notice that this is true for any distribution $f(.)$, provided that μ_4 is not infinite.

PROOF

We shall prove first the following identity, which will be used latter:

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2 + n\left(\overline{X}_n - \mu\right)^2$$

$$\sum (X_i - \mu)^2 = \sum \left(X_i - \overline{X}_n + \overline{X}_n - \mu\right)^2 = \sum \left[\left(X_i - \overline{X}_n\right) + \left(\overline{X}_n - \mu\right)\right]^2 =$$

$$= \sum \left[\left(X_i - \overline{X}_n\right)^2 + 2\left(X_i - \overline{X}_n\right)\left(\overline{X}_n - \mu\right) + \left(\overline{X}_n - \mu\right)^2\right] =$$

$$= \sum \left(X_i - \overline{X}_n\right)^2 + 2\left(\overline{X}_n - \mu\right)\sum \left(X_i - \overline{X}_n\right) + n\left(\overline{X}_n - \mu\right)^2 =$$

$= \sum \left(X_i - \overline{X}_n\right)^2 + n\left(\overline{X}_n - \mu\right)^2$

Using the above identity we obtain:

$E[S_n^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n\left(\overline{X}_n - \mu\right)^2\right] =$

$= \frac{1}{n-1}\left[\sum_{i=1}^{n} E(X_i-\mu)^2 - nE\left(\overline{X}_n - \mu\right)^2\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma^2 - nvar\left(\overline{X}_n\right)\right] =$

$= \frac{1}{n-1}\left[n\sigma^2 - n\frac{1}{n}\sigma^2\right] = \sigma^2$

The derivation of the variance of S_n^2 is omitted. ■

Theorem 13 *Let $X_1, ..., X_n$ be a random sample from a population with mean μ, variance σ^2, skewness κ_3, and kurtosis κ_4. Then,*

(3) $E(F_n(x)) = F(x)$, and $Var(F_n(x)) = F(x)(1-F(x))/n$

(4) Characteristic Function of $\overline{X}$, $\varphi_{\overline{X}}(t) = [\varphi_X(t/n)]^n$.

PROOF

$E(F_n(x)) = E\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x)\right) = E(1(X_i \le x)) = F(x)$. Also $Var(F_n(x)) =$

$E[F_n(x) - F(x)]^2 = E\left\{\frac{1}{n}\sum_{i=1}^{n} 1[X_i \le x] - F(x)\right\}^2 =$

$= E\left\{\frac{1}{n^2}\sum_{i=1}^{n}\{1[X_i \le x] - F(x)\}^2 + \frac{1}{n^2}\sum_{i\neq j}^{n}\{1[X_i \le x] - F(x)\}\{1[X_j \le x] - F(x)\}\right\}$

$= \frac{1}{n}E\left\{\{1[X_i \le x] - F(x)\}^2\right\} = \frac{1}{n}\left\{E\{1[X_i \le x] - F^2(x)\}\right\} = \frac{1}{n}F^2(x)[1-F(x)]$

■

5.3 Sampling from the Normal Distribution

Theorem 14 *Let denote $\overline{X}_n$ the sample mean of a random sample of size n from a normal distribution with mean μ and variance σ^2. Then (1) $\overline{X} \backsim N(\mu, \sigma^2/n)$.*

(2) $\overline{X}$ and s^2 are independent.

(3) $\frac{(n-1)s_*^2}{\sigma^2} \backsim \chi^2_{n-1}$

(4) $\frac{\overline{X}-\mu}{s_*/\sqrt{n}} \backsim t_{n-1}$.

PROOF

(1) From a Theorem above we have that

$$\varphi_{\overline{X}}(t) = [\varphi_X(t/n)]^n .$$

Now $\varphi_X(t/n) = \exp\left(i\mu t - \frac{1}{2}\sigma^2 t^2\right)$. Hence $\varphi_{\overline{X}}(t) = \left[\exp\left(i\mu\frac{t}{n} - \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right)\right]^n = \exp\left(i\mu t - \frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2\right)$, which is the cf of a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

(2) For $n = 2$ we have that if $X_1 \backsim N(0,1)$ and $X_2 \backsim N(0,1)$ then $\overline{X} = \frac{X_1+X_2}{2}$ and $s^2 = \frac{(X_1-X_2)^2}{4}$. Define $Z_1 = \frac{X_1+X_2}{2}$ and $Z_2 = \frac{X_1-X_2}{2}$. Then Z_1 and Z_2 are uncorrelated and by normality independent. ■

5.3.1 The Gamma Function

The **gamma function** is defined as:

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \qquad for \quad t > 0$$

Notice that $\Gamma(t+1) = t\Gamma(t)$,as

$$\Gamma(t+1) = \int_0^\infty x^t e^{-x}dx = -\int_0^\infty x^t de^{-x} = -\left.x^t e^{-x}\right|_0^\infty + t\int_0^\infty x^{t-1}de^{-x} = t\Gamma(t)$$

and if t is an integer then $\Gamma(t+1) = t!$. Also if t is again an integer then $\Gamma(t+\frac{1}{2}) = \frac{1*3*5*...(2t-1)}{2^t}\sqrt{\pi}$. Finally $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Recall that if X is a random variable with density

$$f_X(x) = \frac{1}{\Gamma(k/2)}\left(\frac{1}{2}\right)^{k/2} x^{\frac{k}{2}-1}e^{-\frac{1}{2}x} \qquad for \quad 0 < x < \infty$$

where $\Gamma(.)$ is the gamma function, then X is defined to have a **chi-square distribution with k degrees of freedom**.

Notice that X is distributed as above then:

$$E[X] = k \qquad and \qquad var[X] = 2k$$

We can prove the following theorem

Theorem 15 *If the random variables X_i, $i = 1, 2, .., k$ are normally and independently distributed with means μ_i and variances σ_i^2 then*

$$U = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

has a chi-square distribution with k degrees of freedom. Proof omitted.

Furthermore,

Theorem 16 *If the random variables X_i, $i = 1, 2, .., k$ are normally and independently distributed with mean μ and variance σ^2, and let $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$ then*

$$U = \frac{(n-1)S^2}{\sigma^2} \backsim \chi^2_{n-1}$$

where χ^2_{n-1} is the chi-square distribution with $n-1$ degrees of freedom. Proof omitted.

5.3.2 The F Distribution

If X is a random variable with density

$$f_X(x) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} \frac{x^{\frac{m}{2}-1}}{[1 + (m/n)x]^{(m+n)/2}} \qquad for \quad 0 < x < \infty$$

where $\Gamma(.)$ is the gamma function, then X is defined to have a **F distribution with m and n degrees of freedom**.

Notice that if X is distributed as above then:

$$E[X] = \frac{n}{n-2} \qquad and \qquad var[X] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

Theorem 17 *If the random variables U and V are independently distributed as chi-square with m and n degrees of freedom, respectively i.e. $U \backsim \chi^2_m$ and $V \backsim \chi^2_n$ independently, then*

$$\frac{U/m}{V/n} = X \backsim F_{m,n}$$

where $F_{m,n}$ is the F distribution with m, n degrees of freedom. Proof omitted.

5.3.3 The Student-t Distribution

If X is a random variable with density

$$f_X(x) = \frac{\Gamma[(k+1)/2]}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{[1+x^2/k]^{(k+1)/2}} \qquad for \quad -\infty < x < \infty$$

where $\Gamma(.)$ is the gamma function, then X is defined to have a **t distribution with k degrees of freedom**.

Notice that if X is distributed as above then:

$$E[X] = 0 \qquad and \qquad var[X] = \frac{k}{k-2}$$

Theorem 18 *If the random variables Z and V are independently distributed as standard normal and chi-square with k, respectively i.e. $Z \backsim (N(0,1)$ and $V \backsim \chi^2_k$ independently, then*

$$\frac{Z}{\sqrt{V/k}} = X \backsim t_k$$

where t_k is the t distribution with k degrees of freedom. Proof omitted.

The above Theorems are very useful especially to get the distribution of various tests and construct confidence intervals.

Chapter 6

POINT AND INTERVAL ESTIMATION

The problem of estimation is defined as follows. Assume that some characteristic of the elements in a population can be represented by a random variable X whose density is $f_X(.;\theta) = f(.;\theta)$, where the form of the density is assumed known except that it contains an unknown parameter θ (if θ were known, the density function would be completely specified, and there would be no need to make inferences about it. Further assume that the values $x_1, x_2, ..., x_n$ of a random sample $X_1, X_2, ..., X_n$ from $f(.;\theta)$ can be observed. On the basis of the observed sample values $x_1, x_2, ..., x_n$ it is desired to estimate the value of the unknown parameter θ or the value of some function, say $\tau(\theta)$, of the unknown parameter. The estimation can be made in two ways. The first, called **point estimation**, is to let the value of some statistic, say $t(X_1, X_2, ..., X_n)$, represent or estimate, the unknown $\tau(\theta)$. Such a statistic is called the **point estimator**. The second, called **interval estimation**, is to define two statistics, say $t_1(X_1, X_2, ..., X_n)$ and $t_2(X_1, X_2, ..., X_n)$, where $t_1(X_1, X_2, ..., X_n) < t_2(X_1, X_2, ..., X_n)$, so that $(t_1(X_1, X_2, ..., X_n), t_2(X_1, X_2, ..., X_n))$ constitutes an **interval** for which the probability can be determined that it contains the unknown $\tau(\theta)$.

6.1 Parametric Point Estimation

The point estimation admits two problems. The first is to devise some means of obtaining a statistic to use as an estimator. The second, to select criteria and techniques

to define and find a "best" estimator among many possible estimators.

6.1.1 ***Methods of Finding Estimators***

Any statistic (known function of observable random variables that is itself a random variable) whose values are used to estimate $\tau(\theta)$, where $\tau(.)$ is some function of the parameter θ, is defined to be an **estimator** of $\tau(\theta)$.

Notice that for specific values of the realized random sample the estimator takes a specific value called **estimate**.

6.1.2 ***Method of Moments***

Let $f(.;\theta_1,\theta_2,...,\theta_k)$ be a density of a random variable X which has k parameters $\theta_1,\theta_2,...,\theta_k$. As before let $\mu_r^{/}$ denote the r^{th} moment i.e. $= E[X^r]$. In general $\mu_r^{/}$ will be a known function of the k parameters $\theta_1,\theta_2,...,\theta_k$. Denote this by writing $\mu_r^{/} = \mu_r^{/}(\theta_1,\theta_2,...,\theta_k)$. Let $X_1,X_2,...,X_n$ be a random sample from the density $f(.;\theta_1,\theta_2,...,\theta_k)$, and, as before, let $M_j^{/}$ be the j^{th} sample moment, i.e. $M_j^{/} = \frac{1}{n}\sum_{i=1}^{n} X_i^j$. Then equating sample moments to population ones we get k equations with k unknowns, i.e.

$$M_j^{/} = \mu_j^{/}(\theta_1,\theta_2,...,\theta_k) \qquad for \quad j = 1,2,...,k$$

Let the solution to these equations be $\widehat{\theta}_1,\widehat{\theta}_2,...,\widehat{\theta}_k$. We say that these k estimators are the estimators of $\theta_1,\theta_2,...,\theta_k$ obtained by the **method of moments**.

EXAMPLE: Let $X_1,X_2,...,X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Let $(\theta_1,\theta_2) = (\mu,\sigma^2)$. Estimate the parameters μ and σ by the method of moments.. Recall that $\sigma^2 = \mu_2^{/} - (\mu_1^{/})^2$ and $\mu = \mu_1^{/}$. The method of moment equations become:

$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = M_1^{/} = \mu_1^{/} = \mu_1^{/}(\mu,\sigma^2) = \mu$

$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = M_2^{/} = \mu_2^{/} = \mu_2^{/}(\mu,\sigma^2) = \sigma^2 + \mu^2$

Solving the two equations for μ and σ we get:

$\widehat{\mu} = \overline{X}$, and $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})}$ which are the M-M estimators of μ and σ.

EXAMPLE: Let $X_1, X_2, ..., X_n$ be a random sample from a Poisson distribution with parameter λ. There is only one parameter, hence only one equation, which is:

$\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = M_1^{/} = \mu_1^{/} = \mu_1^{/}(\lambda) = \lambda$

Hence the M-M estimator of λ is $\widehat{\lambda} = \overline{X}$.

6.1.3 ***Maximum Likelihood***

Consider the following estimation problem. Suppose that a box contains a number of black and a number of white balls, and suppose that it is known that the ratio of the number is 3/1 but it is not known whether the black or the white are more numerous, i.e. the number of drawing a black ball is either 1/4 or 3/4. If n balls are drawn with replacement from the box, the distribution of X, the number of black balls, is given by the binomial distribution

$$f(x;p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x} \qquad for \qquad x = 0, 1, 2, ..., n$$

where p is the probability of drawing a black ball. Here $p = 1/4$ or $p = 3/4$. We shall draw a sample of three balls, i.e. $n = 3$, with replacement and attempt to estimate the unknown parameter p of the distribution. the estimation is simple in this case as we have to choose only between the two numbers $1/4 = 0.25$ and $3/4 = 0.75$. The possible outcomes and their probabilities are given below:

$outcome : x$	0	1	2	3
$f(x; 0.75)$	1/64	9/64	27/64	27/64
$f(x; 0.25)$	27/64	27/64	9/64	1/64

In the present example, if we found $x = 0$ in a sample of 3, the estimate 0.25 for p would be preferred over 0.75 because the probability 27/64 is greater than 1/64, i.e.

And in general we should estimate p by 0.25 when $x = 0$ or 1 and by 0.75 when $x = 2$ or 3. The estimator may be defined as

$$\widehat{p} = \widehat{p}(x) = \begin{cases} 0.25 & for \quad x = 0, 1 \\ 0.75 & for \quad x = 2, 3 \end{cases}$$

The estimator thus selects fro every possible x the value of p, say $\widehat{p}$, such that

$$f(x; \widehat{p}) > f(x; p^{/})$$

where $p^{/}$ is the other value of p.

More generally, if several values of p were possible, we might reasonably proceed in the same manner. Thus if we found $x = 2$ in a sample of 3 from a binomial population, we should substitute all possible values of p in the expression

$$f(2; p) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} p^2(1-p) \qquad for \quad 0 \leq p \leq 1$$

and choose as our estimate that value of p which maximizes $f(2; p)$. The position of the maximum of the function above is found by setting equal to zero the first derivative with respect to p, i.e. $\frac{d}{dp} f(2; p) = 6p - 9p^2 = 3p(2 - 3p) = 0 \Rightarrow p = 0$ or $p = 2/3$. The second derivative is: $\frac{d^2}{dp^2} f(2; p) = 6 - 18p$. Hence, $\frac{d^2}{dp^2} f(2; 0) = 6$ and the value of $p = 0$ represents a minimum, whereas $\frac{d^2}{dp^2} f(2; \frac{2}{3}) = -6$ and consequently $p = \frac{2}{3}$ represents the maximum. Hence $\widehat{p} = \frac{2}{3}$ is our estimate which has the property

$$f(x; \widehat{p}) > f(x; p^{/})$$

where $p^{/}$ is any other value in the interval $0 \leq p \leq 1$.

The **likelihood function** of n random variables $X_1, X_2, ..., X_n$ is defined to be the joint density of the n random variables, say $f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n; \theta)$, which is considered to be a function of θ. In particular, if $X_1, X_2, ..., X_n$ is a random sample from the density $f(x; \theta)$, then the likelihood function is $f(x_1; \theta) f(x_2; \theta) ..... f(x_n; \theta)$. To think of the likelihood function as a function of θ, we shall use the notation $L(\theta; x_1, x_2, ..., x_n)$ or $L(\bullet; x_1, x_2, ..., x_n)$ for the likelihood function in general.

The likelihood is a value of a density function. Consequently, for discrete random variables it is a probability. Suppose for the moment that θ is known, denoted by θ_0. The particular value of the random variables which is "most likely to occur" is that value $x_1^{/}, x_2^{/}, ..., x_n^{/}$ such that $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n; \theta_0)$ is a maximum. for example, for simplicity let us assume that $n = 1$ and X_1 has the normal density with mean 0 and variance 1. Then the value of the random variable which is most likely to occur is $X_1 = 0$. By "most likely to occur" we mean the value $x_1^{/}$ of X_1 such that $\phi_{0,1}(x_1^{/}) > \phi_{0,1}(x_1)$. Now let us suppose that the joint density of n random variables is $f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n; \theta)$, where θ is known. Let the particular values which are observed be represented by $x_1^{/}, x_2^{/}, ..., x_n^{/}$. We want to know from which density is this particular set of values most likely to have come. We want to know from which density (what value of θ) is the likelihood largest that the set $x_1^{/}, x_2^{/}, ..., x_n^{/}$ was obtained. in other words, we want to find the value of θ in the admissible set, denoted by $\widehat{\theta}$, which maximizes the likelihood function $L(\theta; x_1^{/}, x_2^{/}, ..., x_n^{/})$. The value $\widehat{\theta}$ which maximizes the likelihood function is, in general, a function of $x_1, x_2, ..., x_n$, say $\widehat{\theta} = \widehat{\theta}(x_1, x_2, ..., x_n)$. Hence we have the following definition:

Let $L(\theta) = L(\theta; x_1, x_2, ..., x_n)$ be the likelihood function for the random variables $X_1, X_2, ..., X_n$. If $\widehat{\theta}$ [where $\widehat{\theta} = \widehat{\theta}(x_1, x_2, ..., x_n)$ is a function of the observations $x_1, x_2, ..., x_n$] is the value of θ in the admissible range which maximizes $L(\theta)$, then $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$ is the **maximum likelihood estimator** of θ. $\widehat{\theta} = \widehat{\theta}(x_1, x_2, ..., x_n)$ is the **maximum likelihood estimate** of θ for the sample $x_1, x_2, ..., x_n$.

The most important cases which we shall consider are those in which $X_1, X_2, ..., X_n$ is a random sample from some density function $f(x; \theta)$, so that the likelihood function is

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) ..... f(x_n; \theta)$$

Many likelihood functions satisfy regularity conditions so the maximum likelihood estimator is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0$$

Also $L(\theta)$ and $\log L(\theta)$ have their maxima at the same value of θ, and it is sometimes easier to find the maximum of the logarithm of the likelihood. Notice also that if the likelihood function contains k parameters then we find the estimator from the solution of the k first order conditions.

EXAMPLE: Let a random sample of size n is drawn from the Bernoulli distribution

$$f(x;p) = p^x(1-p)^{1-x}$$

where $0 \leq p \leq 1$. The sample values $x_1, x_2, ..., x_n$ will be a sequence of 0s and 1s, and the likelihood function is

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Let $y = \sum x_i$ we obtain that

$$logL(p) = y \log p + (n-y)\log(1-p)$$

and

$$\frac{d \log L(p)}{dp} = \frac{y}{p} - \frac{n-y}{1-p}$$

Setting this expression equal to zero we get

$$\widehat{p} = \frac{y}{n} = \frac{1}{n}\sum x_i = \overline{x}$$

which is intuitively what the estimate for this parameter should be.

EXAMPLE: Let a random sample of size n is drawn from the normal distribution with density

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The likelihood function is

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

the logarithm of the likelihood function is

$$\log L(\mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To find the maximum with respect to μ and σ^2 we compute

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

and putting these derivatives equal to 0 and solving the resulting equations we find the estimates

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

and

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

which turn out to be the sample moments corresponding to μ and σ^2.

6.1.4 ***Properties of Point Estimators***

One needs to define criteria so that various estimators can be compared. One of these is the unbiasedness. An estimator $T = t(X_1, X_2, ..., X_n)$ is defined to be an **unbiased** estimator of $\tau(\theta)$ if and only if

$$E_\theta[T] = E_\theta[t(X_1, X_2, ..., X_n)] = \tau(\theta)$$

for all θ in the admissible space.

Other criteria are consistency, mean square error etc.

6.2 Interval Estimation

In practice estimates are often given in the form of the estimate plus or minus a certain amount e.g. the cost per volume of a book could be 83 ± 4.5 per cent which

means that the actual cost will lie somewhere between 78.5% and 87.5% with high probability. Let us consider a particular example. Suppose that a random sample $(1.2, 3.4, .6, 5.6)$ of four observations is drawn from a normal population with unknown mean μ and a known variance 9. The maximum likelihood estimate of μ is the sample mean of the observations:

$\overline{x} = 2.7$

We wish to determine upper and lower limits which are rather certain to contain the true unknown parameter value between them. We know that the sample mean, $\overline{x}$, is distributed as normal with mean μ and variance $9/n$ i.e. $\overline{x} \backsim N(\mu, \sigma^2/n)$. Hence we have

$$Z = \frac{\overline{X} - \mu}{\frac{3}{2}} \backsim N(0,1)$$

Hence Z is standard normal. Consequently we can find the probability that Z will be between two arbitrary values. For example we have that

$$P[-1.96 < Z < 1.96] = \int_{-1.96}^{1.96} \phi(z)dz = 0.95$$

Hence we get that μ must be in the interval

$$\overline{X} + \frac{3}{2}1.96 > \mu > \overline{X} - \frac{3}{2}1.96$$

and for the specific value of the sample mean we have that $5.64 > \mu > -.24$ i.e. $P[5.64 > \mu > -.24] = .95$. This leads us to the following definition for the confidence interval.

Let $X_1, X_2, ..., X_n$ be a random sample from the density $f(\bullet; \theta)$. Let $T_1 = t_1(X_1, X_2, ..., X_n)$ and $T_2 = t_2(X_1, X_2, ..., X_n)$ be two statistics satisfying $T_1 \leq T_2$ for which $P_\theta[T_1 < \tau(\theta) < T_2] = \gamma$, where γ does not depend on θ. Then the random interval (T_1, T_2) is called a 100γ **percent confidence interval** for $\tau(\theta)$. γ is called the **confidence coefficient**. T_1 and T_2 are called the **lower** and **upper** confidence limits, respectively. A value (t_1, t_2) of the random interval (T_1, T_2) is **also** called a 100γ **percent confidence interval** for $\tau(\theta)$.

Let $X_1, X_2, ..., X_n$ be a random sample from the density $f(\bullet;\theta)$. Let $T_1 = t_1(X_1, X_2, ..., X_n)$ be a statistic for which $P_\theta[T_1 < \tau(\theta)] = \gamma$. Then T_1 is called a **one-sided lower confidence interval** for $\tau(\theta)$. Similarly, let $T_2 = t_2(X_1, X_2, ..., X_n)$ be a statistic for which $P_\theta[\tau(\theta) < T_2] = \gamma$. Then T_2 is called a **one-sided upper confidence interval** for $\tau(\theta)$.

EXAMPLE: Let $X_1, X_2, ..., X_n$ be a random sample from the density $f(x;\theta) = \phi_{\theta,9}(x)$. Set $T_1 = t_1(X_1, X_2, ..., X_n) = \overline{X} - 6/\sqrt{n}$ and $T_2 = t_2(X_1, X_2, ..., X_n) = \overline{X} + 6/\sqrt{n}$. Then (T_1, T_2) constitutes a random interval and is a confidence interval for $\tau(\theta) = \theta$, with confidence coefficient $\gamma = P[\overline{X} - 6/\sqrt{n} < \theta < \overline{X} + 6/\sqrt{n}] =$

$= P[-2 < \frac{\overline{X}-\theta}{\frac{3}{\sqrt{n}}} < 2] = \Phi(2) - \Phi(-2) = 0.9772 - 0.0228 = 0.9544.$ hence if the random sample of 25 observations has a sample mean of, say, 17.5, then the interval $(17.5 - 6/\sqrt{25}, 17.5 + 6/\sqrt{25})$ is also called a confidence interval of θ.

6.2.1 ***Sampling from the Normal Distribution***

Let $X_1, X_2, ..., X_n$ be a random sample from the normal distribution with mean μ and variance σ^2. If σ^2 is unknown then $\theta = (\mu, \sigma^2)$, the unknown parameters and $\tau(\theta) = \mu$, the parameter we want to estimate by interval estimation. We know that

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \backsim N(0,1)$$

However, the problem with this statistic is that we have two parameters. Consequently we can not an interval. hence we look for a statistic that involves only the parameter we want to estimate, i.e. μ. Notice that

$$\frac{(\overline{X} - \mu)/\frac{\sigma}{\sqrt{n}}}{\sqrt{\sum(X_i - \overline{X})^2/(n-1)\sigma^2}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \backsim t_{n-1}$$

This statistic involves only the parameter we want to estimate. Hence we have

$$\left\{ q_1 < \frac{\overline{X} - \mu}{S/\sqrt{n}} < q_2 \right\} \Leftrightarrow \left\{ \overline{X} - q_2(S/\sqrt{n}) < \mu < \overline{X} - q_1(S/\sqrt{n}) \right\}$$

where q_1, q_2 are such that

$$P\left[q_1 < \frac{\overline{X} - \mu}{S/\sqrt{n}} < q_2 \right] = \gamma$$

Hence the interval $\left(\overline{X} - q_2(S/\sqrt{n}), \overline{X} - q_1(S/\sqrt{n})\right)$ is the 100γ percent confidence interval for μ. It can be proved that if q_1, q_2 are symmetrical around 0, then the length is the interval is minimized.

Alternatively, if we want to find a confidence interval for σ^2, when μ is unknown, then we use the statistic

$$\frac{\sum(X_i - \overline{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \backsim \chi^2_{n-1}$$

Hence we have

$$\left\{q_1 < \frac{(n-1)S^2}{\sigma^2} < q_2\right\} \Leftrightarrow \left\{\frac{(n-1)S^2}{q_2} < \sigma^2 < \frac{(n-1)S^2}{q_1}\right\}$$

where q_1, q_2 are such that

$$P\left[q_1 < \frac{(n-1)S^2}{\sigma^2} < q_2\right] = \gamma$$

So the interval $\left(\frac{(n-1)S^2}{q_2}, \frac{(n-1)S^2}{q_1}\right)$ is a 100γ percent confidence interval for σ^2. The q_1, q_2 are often selected so that $P\left[q_2 < \frac{(n-1)S^2}{\sigma^2}\right] = P\left[\frac{(n-1)S^2}{\sigma^2} < q_1\right] = (1-\gamma)/2$. Such a confidence interval is referred to as **equal-tailed** confidence interval for σ^2.

Chapter 7

HYPOTHESIS TESTING

A **statistical hypothesis** is an assertion or conjecture, denoted by $\mathcal{H}$, about a distribution of one or more random variables. If the statistical hypothesis completely specifies the distribution is **simple**, otherwise is **composite**.

Example Let $X_1, X_2, ..., X_n$ be a random sample from $f(x;\theta) = \phi_{\mu,25}(x)$. The statistical hypothesis that the mean of the normal population is less or equal to 17 is denoted by: $\mathcal{H} : \theta \leq 17$. Such a hypothesis is composite, as it does not completely specify the distribution. On the other hand, the hypothesis $\mathcal{H} : \theta \leq 17$ is simple since it completely specifies the distribution.

A **test** of statistical hypothesis $\mathcal{H}$ is a rule or procedure for deciding whether to reject $\mathcal{H}$.

Example Let $X_1, X_2, ..., X_n$ be a random sample from $f(x;\theta) = \phi_{\mu,25}(x)$. Consider $\mathcal{H} : \theta \leq 17$. One possible test $\mathcal{Y}$ is as follows: Reject $\mathcal{H}$ if and only if $\overline{X} > 17 + 5/\sqrt{n}$.

In many hypotheses-testing problems two hypotheses are discussed. The first, the hypothesis being testing, is called the **null** hypothesis, denoted by $\mathcal{H}_0$, and the second is called the **alternative** hypothesis denoted by $\mathcal{H}_1$. We say that $\mathcal{H}_0$ is tested against or versus $\mathcal{H}_1$. The thinking is that if the null hypothesis is wrong the alternative hypothesis is true, and vice versa. We can make two types of errors:

Rejection of $\mathcal{H}_0$ when $\mathcal{H}_0$ is true is called a **Type I error**, and acceptance of $\mathcal{H}_0$ when $\mathcal{H}_0$ is false is called a **Type II error**. The **size of Type I error** is defined

to be the probability that a Type I error is made, and similarly the **size of a Type II error** is defined to be the probability that a Type II error is made.

Significance level or **size of a test**, denoted by α, is the supremum of the probability of rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is correct, i.e. it is the supremum of the Type I error. In general to perform a test we fix the size to a prespecified value in general 10%, 5% or 1%.

Example Let $X_1, X_2, ..., X_n$ be a random sample from $f(x;\theta) = \phi_{\mu,25}(x)$. Consider $\mathcal{H}_0 : \theta \leq 17$ and the test $\mathcal{Y}$: Reject $\mathcal{H}_0$ if and only if $\overline{X} > 17 + 5/\sqrt{n}$. Then of the test is

$$\sup_{\theta\leq 17} P[\overline{X} > 17 + 5/\sqrt{n}] = \sup_{\theta\leq 17} P\left[\frac{\overline{X}-\theta}{5/\sqrt{n}} > \frac{17+5/\sqrt{n}-\theta}{5/\sqrt{n}}\right] =$$
$$= \sup_{\theta\leq 17}\left\{1 - P\left[\frac{\overline{X}-\theta}{5/\sqrt{n}} \leq \frac{17+5/\sqrt{n}-\theta}{5/\sqrt{n}}\right]\right\} = \sup_{\theta\leq 17}\left\{1 - P\left[Z \leq \frac{17+5/\sqrt{n}-\theta}{5/\sqrt{n}}\right]\right\} =$$
$$= \sup_{\theta\leq 17}\left\{1 - \Phi\left[\frac{17+5/\sqrt{n}-\theta}{5/\sqrt{n}}\right]\right\} = 1 - \Phi(1) = 0.159$$

7.1 Testing Procedure

Let us establish a test procedure via an example. Assume that $n = 64$, $\overline{X} = 9.8$ and $\sigma^2 = 0.04$. We would like to test the hypothesis that $\mu = 10$.

1. Formulate the null hypothesis:

$\mathcal{H}_0 : \mu = 10$

2. Formulate the alternative:

$\mathcal{H}_1 : \mu \neq 10$

3. select the level of significance:

$\alpha = 0.01$

From tables find the critical values for Z, denoted by $c_Z = 2.58$.

4. Establish the rejection limits:

Reject $\mathcal{H}_0$ if $Z < -2.58$ or $Z > 2.58$.

5. Calculate Z:

$Z = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{9.8-10}{0.2/\sqrt{64}} = -8$

6. Make the decision:

Since Z is less than -2.58, reject $\mathcal{H}_0$.

To find the appropriate test for the mean we have to consider the following cases:

1. Normal population and known population variance (or standard deviation).

In this case the statistic we use is:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \backsim N(0,1)$$

2. Large samples in order to use the central limit theorem.

In this case the statistic we use is:

$$Z = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \backsim N(0,1)$$

3. Small samples from a normal population where the population variance (or standard deviation) is unknown.

In this case the statistic we use is:

$$t = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \backsim t_{n-1}$$

7.2 Testing Proportions

The null hypothesis will be of the form:

$\mathcal{H}_0 : \pi = \pi_0$

an the three possible alternatives are:

(1) $\mathcal{H}_1 : \pi \neq \pi_0$ two sided test, (2) $\mathcal{H}_1 : \pi < \pi_0$ one sided, (3) $\mathcal{H}_1 : \pi > \pi_0$ one sided. The appropriate statistic is based on the central limit theorem and is:

$$Z = \frac{p - \pi_0}{\frac{S}{\sqrt{n}}} \backsim N(0,1) \quad where \quad S^2 = \pi_0(1 - \pi_0)$$

Example: Mr. X believes that he will get more 60% of the votes. However, in a sample of 400 voters 252 indicate that they will vote for X. At a significance level of 5% test Mr. X belief.

$p = \frac{252}{400} = 0.63$, $S^2 = 0.6(1-0.6) = 0.24$. The $\mathcal{H}_0 : \pi = \pi_0$ and the alternative is $\mathcal{H}_1 : \pi > \pi_0$. The critical value is 1.64. Now $Z = \frac{p-\pi_0}{\frac{S}{\sqrt{n}}} = \frac{0.63-0.6}{0.489/\sqrt{400}} = 1.22$. Consequently, the null is not rejected as $Z < 1.64$. Thus Mr. X belief is wrong.

If fact we have the following possible outcomes when testing hypotheses:

	$\mathcal{H}_0$ is accepted	$\mathcal{H}_1$ is accepted
$\mathcal{H}_0$ is correct	Correct decision $(1-\alpha)$	Type I error (α)
$\mathcal{H}_1$ is correct	Type II error (β)	Correct decision $(1-\beta)$

An **operating characteristic curve** presents the probability of accepting a null hypothesis for various values of the population parameter at a given significance level α using a particular sample size. The **power** of the test is the inverse function of the operating characteristic curve, i.e. it is the probability of rejecting the null hypothesis for various possible values of the population parameter.

Part III

Asymptotic Theory

Chapter 8

MODES OF CONVERGENCE

We have a statistic T which is a measurable function of the data

$$T_n = T(X_1, ..., X_n),$$

and we would like to know what happens to T_n as $n \to \infty$. It turns out that the limit is easier to work with than T_n itself. The plan is to use the limit as approximation device. We think of a sequence $T_1, T_2, ...$ which have distribution functions $F_1, F_2, ...$

Definition A sequence of random variables $T_1, T_2, ...$ converges in probability to a random variable T (denoted by $T_n \xrightarrow{p} T$) if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P[|T_n - T| > \varepsilon] = 0.$$

Definition A sequence of random variables $T_1, T_2, ...$ converges in mean square to a random variable T (denoted by $T_n \xrightarrow{ms} T$) if,

$$\lim_{n\to\infty} E(T_n - T)^2 = 0$$

which is equivalent to (a) $Var(T_n) \to 0$ and (b) $E(T_n) - E(T) \to 0$ because of the triangle inequality.

Theorem 19 *Convergence in mean square implies convergence in probability.*

Proof. By the Markov/Chebychev inequality,

$$P[|T_n - T| \geq \varepsilon] \leq \frac{E|T_n - T|^2}{\varepsilon^2} \to 0.$$

■

Note that if $T_n > 0$,

$$P\left[T_n \geq \varepsilon\right] \leq \frac{E\left(T_n\right)^2}{\varepsilon^2} \leq \frac{E\left(T_n\right)}{\varepsilon}$$

so that if $\frac{E(T_n)}{\varepsilon} \to 0$, this is sufficient for $T_n \xrightarrow{p} 0$.

But the converse of the theorem is not necessarily true. To see this consider the following random variable

$$T_n = \begin{cases} n \quad with \quad probability \quad \frac{1}{n} \\ 0 \quad with \quad probability \quad 1 - \frac{1}{n} \end{cases}, \qquad T = 0.$$

Then $P\left[T_n \geq \varepsilon\right] = \frac{1}{n}$ for any $\varepsilon > 0$, and $P\left[T_n \geq \varepsilon\right] = \frac{1}{n} \to 0$. But $E\left(T_n\right)^2 = n^2 \frac{1}{n} = n \to \infty$.

A famous consequence of the theorem is the (Weak) Law of Large Numbers

Theorem 20 *WEAK LAW of LARGE NUMBERS Let $X_1, ..., X_n$ be i.i.d. with $E\left(X_i\right) = \mu$ and $Var\left(X_i\right) = \sigma^2 < \infty$, and $T_n = \overline{X}$. Then for all $\varepsilon > 0$,*

$$\lim_{n \to \infty} P[|T_n - \mu| > \varepsilon] = 0, \quad i.e., \quad T_n \xrightarrow{p} \mu.$$

The proof is easy because

$$E[(T_n - \mu)^2] = \frac{\sigma^2}{n} \to 0.$$

as we have shown. In fact, the result can be proved with only the hypothesis that $E|X| < \infty$ by using a truncation argument.

Another application of the previous theorem is to the empirical distribution function, i.e.,

$$F_n\left(x\right) = \frac{1}{n} \sum_{i=1}^{n} 1\left(X_i \leq x\right) \xrightarrow{p} F\left(x\right).$$

The next result is very important for applying the Law of Large Numbers beyond the simple sum of iid's.

Theorem 21 *CONTINUOUS MAPPING THEOREM If* $T_n \xrightarrow{p} \mu$ *a constant and* $g(.)$ *is a continuous function at* μ*, then*

$$g(T_n) \xrightarrow{p} g(\mu).$$

Proof. Let $\varepsilon > 0$. By the continuity of g at μ, $\exists\ \eta > 0$ such that if

$$|x - \mu| < \eta \Rightarrow |g(x) - g(\mu)| < \varepsilon$$

Let $A_n = \{|T_n - \mu| < \eta\}$ and $B_n = \{|g(T_n) - g(\mu)| < \varepsilon\}$. But when A_n is true so is B_n, i.e., $A_n \subset B_n$. Since $P(A_n) \to 1$, we must have that $P(B_n) \to 1$.

■

Now we look at the sample variance

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

We know that:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} E\left(X_i^2\right) = \sigma^2 + \mu^2$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu \Rightarrow \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \xrightarrow{p} \mu^2$$

by the continuous mapping theorem. Combining these two results we get

$$s^2 \xrightarrow{p} \sigma^2.$$

Finally, notice that when dealing with a vector $T_n = (T_{n1}, ..., T_{nk})^{/}$, we have that

$$\|T_n - T\| \xrightarrow{p} 0,$$

where $\|x\| = \left(x^{/}x\right)^{1/2}$ is the Euclidean norm, if and only if

$$|T_{nj} - T_j| \xrightarrow{p} 0$$

for all $j = 1, ..., k$. The if part is no surprise and follows from the continuous mapping theorem. The only if part follows as if $\|T_n - T\| < \varepsilon$ then $|T_{nj} - T_j| < \eta$ for each j and some $\eta > 0$.

Definition A sequence of random variables $T_1, T_2, ...$ converges almost surely to a random variable T (denoted by $T_n \xrightarrow{as} T$) if, for every $\varepsilon > 0$,

$$P[\lim_{n\to\infty} |T_n - T| < \varepsilon] = 1.$$

This result is generally harder to establish than convergence in probability, i.e., there are not simple sufficient conditions based on mean and variance. Almost sure convergence implies convergence in probability but not vice versa. Note again that vector convergence is equivalent to componentwise convergence. Continuous mapping theorem is obvious. Let

$$A = \{\omega : T_n(\omega) \to T(\omega)\}, P(A) = 1$$

On this set A, we have

$$g[T_n(\omega)] \to g[T(\omega)],$$

by ordinary continuity.

Theorem 22 *STRONG LAW of LARGE NUMBERS. If* $E|X| < \infty$, *then*

$$T_n(\omega) \xrightarrow{as} E(X).$$

We can have Strong Law of Large Numbers applied to empirical distribution functions and to sample variances (from the continuous mapping theorem) etc.

Chapter 9

ASYMPTOTIC THEORY 2

We can now establish the convergence in distribution and the central limit theorem, which is of great importance.

Definition A sequence of random variables $T_1, T_2, ...$ converges in distribution to a random variable T (denoted by $T_n \xrightarrow{D} T$) if

$$\lim_{n\to\infty} P[T_n \leq x] = P[T \leq x]$$

at all points of continuity of $F_T(x) = P[T \leq x]$.

Convergence in distribution is weaker than in probability, i.e.,

$$T_n \xrightarrow{P} T \Rightarrow T_n \xrightarrow{D} T$$

but not vice versa, except when the limit is nonrandom, i.e.,

$$T_n \xrightarrow{D} \alpha \Rightarrow T_n \xrightarrow{P} \alpha.$$

Theorem 23 *A sequence of random vectors* $(T_{n1}, T_{n2}, ..., T_{nk}) \xrightarrow{D} (T_1, T_2, ..., T_k)$ *iff*

$$c^{/} T_n \xrightarrow{D} c^{/} T$$

for any $c^{/} = (c_1, c_2, ..., c_k) \neq 0$.

This is known as the *Cramér − Wold* device. The main result is the following:

Theorem 24 *Central Limit Theorem of Lindemberg-Lévy. Let* $X_1, X_2, ..., X_n$ *be i.i.d. with* $E(X_i) = \mu$, $Var(X_i) = \sigma^2 < \infty$. *Then*

$$\begin{aligned} & \sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{D} N\left(0, \sigma^2\right) \\ T_n &= \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \xrightarrow{D} N(0,1) \end{aligned}$$

The vector version: Let $X_1, X_2, ..., X_n$ be i.i.d. with $E(X_i) = \mu$, $E\left[(X_i - \mu)(X_i - \mu)'\right] = \Sigma$ where $0 < \Sigma < \infty$. Then

$$T_n = \sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{D} N(0, \Sigma).$$

A modern proof of the result is based on characteristic functions.

Example. If $X_i \backsim N(\mu, \sigma^2)$, then

$$T_n = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \xrightarrow{D} N(0,1)$$

for all n a result which is trivial. Suppose instead that

$$X_i = \begin{cases} 1 & 0.5 \\ -1 & 0.5 \end{cases}$$

$E(X_i) = 0$, $Var(X_i) = 1$.

We know that $T_n \xrightarrow{D} N(0,1)$.

$$n = 2: \qquad \frac{X_1 + X_2}{\sqrt{2}} = \begin{cases} 2/\sqrt{2} & 1/4 \\ 0 & 1/2 \\ -2/\sqrt{2} & 1/4 \end{cases}$$

$$n = 3: \qquad \frac{X_1 + X_2 + X_3}{\sqrt{3}} = \begin{cases} 3/\sqrt{3} & 1/8 \\ 1/\sqrt{3} & 3/8 \\ -1/\sqrt{3} & 3/8 \\ -3/\sqrt{3} & 1/8 \end{cases}$$

etc. The Binomial distribution gets closer and closer to normal.

We can now approximately calculate for example

$$\begin{aligned} P\left(\overline{X} > 10\right) &= P\left(\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sigma} > \frac{\sqrt{n}\left(10-\mu\right)}{\sigma}\right) \\ &\cong P\left(Z > \frac{\sqrt{n}\left(10-\mu\right)}{\sigma}\right) \\ &= 1 - P\left(\frac{\sqrt{n}\left(10-\mu\right)}{\sigma} \leq Z\right) = 1 - \Phi\left(\frac{\sqrt{n}\left(10-\mu\right)}{\sigma}\right) \end{aligned}$$

CLT for non-identically distributed random variables.

Theorem 25 *(Lyapunov) Suppose that $X_1, X_2, ..., X_n$ are independent random variables with*

$$E\left(X_i\right) = \mu, \quad Var\left(X_i\right) = \sigma_i^2, \quad E\left|X_i - \mu_i\right|^3 = m_{3i}$$

and additionally

$$\left(\sum_{i=1}^{n} \sigma_i^2\right)^{-1/2} \left(\sum_{i=1}^{n} m_{3i}\right)^{1/3} \to 0$$

e.g. if

$$\frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 \to \sigma^2 \quad \frac{1}{n}\sum_{i=1}^{n} m_{3i} \to m_3$$

then

$$\begin{aligned} &\sqrt{n}\left(\overline{X}-\mu\right) \xrightarrow{D} N\left(0,\sigma^2\right) \\ T_n &= \frac{\overline{X} - E\left(\overline{X}\right)}{\sqrt{Var\left(\overline{X}\right)}} \xrightarrow{D} N\left(0,1\right) \end{aligned}$$

The Lindeberg-Feller CLT is even weaker.

Theorem 26 *(Lindeberg-Feller) Let x_i be independent with mean μ_i and variance σ_i^2, and distribution functions F_i. Suppose that $B_n^2 = \sum_{i=1}^{n} \sigma_i^2$ satisfies*

$$\frac{\sigma_n^2}{B_n^2} \to 0, \quad B_n^2 \to \infty, \quad as \quad n \to \infty.$$

Then

$$\frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu_i\right)}{\left[\frac{1}{n^2} B_n^2\right]^{1/2}} \xrightarrow{D} N\left(0,1\right)$$

if and only if the Lindeberg condition

$$\frac{\sum_{i=1}^{n}\int_{|t-\mu_i|>\varepsilon B_n}(t-\mu_i)^2\,dF_i(t)}{B_n^2}\to 0,\quad n\to\infty,\quad each\quad \varepsilon>0$$

is satisfied.

The key condition for the above CLT is the Lindeberg condition. Which basically ensures that no one term is so relatively large as to dominate the entire sample, in the limit. The following CLT gives some more transparent conditions that are sufficient for the Lindeberg condition to hold.

Theorem 27 *Let x_i be independent with mean μ_i and variance σ_i^2, and let $\bar{\sigma}_n^2 = \frac{1}{n}\sum\sigma_i^2$. If*

$$\frac{\max_{1\le i\le n}\left(E|x_i|^{2+\delta}\right)^{\frac{1}{2+\delta}}}{\bar{\sigma}_n}\le B<\infty\quad \delta>0,\quad \forall n\ge 1$$

then

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right)}{\bar{\sigma}_n}\xrightarrow{d}N(0,1)$$

The above condition although sufficient is not necessary. To see this assume that $y_t = 0$ with probability $\frac{1}{2}\left(1-\frac{1}{t^2}\right)$, 0 with the same probability and t with probability $\frac{1}{t^2}$. In this case y_t tends to a Bernoulli random variable, and the CLT certainly applies in this case. Yet the condition of the above theorem is not satisfied.

Furthermore, let assume that $y_t = 0$ with probability $1-\frac{1}{t^2}$ and t with probability $\frac{1}{t^2}$. Then $E(y_t) = 1/t \to 0$, and $Var(y_t) = 1-\frac{1}{t^2}\to 1$. Hence $\bar{\sigma}_n^2\to 1$. Despite this, it is clear that y_t is converging to a degenerate random variable which takes the value of 0 with probability 1 (in fact is $x_t = y_t - E(y_t)$ that is degenerate). However, it is verified that $\left(E|x_i|^{2+\delta}\right)^{\frac{1}{2+\delta}} = O\left(t^{\frac{\delta}{\delta+2}}\right)$ and consequently for any $\delta>0$ the condition of the above theorem must fail for n large enough.

Dependent random variables CLT's are available too.

9.1 Combination Properties

Theorem 28 *(Slutsky's) Suppose that and*$X_n \xrightarrow{D} X$ *and* $Y_n \xrightarrow{P} c$. *Then*

$$Y_n X_n \xrightarrow{D} cX$$
$$Y_n + X_n \xrightarrow{D} c + X$$
$$X_n Y_n / \xrightarrow{D} X/c \quad if \quad c \quad \text{nonzero}$$

Application: Suppose that we look at

$$\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{s_X},$$

when s_X is the sample variance. The CLT tell us that

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{D} N\left(0, \sigma^2\right)$$

The LLN and CMT say that

$$s_X \xrightarrow{p} \sigma.$$

Therefore,

$$\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{s_X} \xrightarrow{D} \frac{N\left(0, \sigma^2\right)}{\sigma} = N\left(0, 1\right)$$

Theorem 29 *Continuous Mapping Theorem II. Suppose that*

$$T_n = (T_{n1}, T_{n2}, .., T_{nk}) \xrightarrow{D} T = (T_1, T_2, .., T_k)$$

and $g : \mathbb{R}^k \to \mathbb{R}^q$. *Then*

$$g\left(T_n\right) \xrightarrow{D} g\left(T\right)$$

EXAMPLE.

$$Y_n + X_n \xrightarrow{D} X + Y$$
$$Y_n X_n \xrightarrow{D} YX$$

but notice that the assumption requires the joint convergence of (Y_n, X_n).

Theorem 30 *(Cramer) Assume that* $X_n \xrightarrow{d} N(\mu, \Sigma)$*, and* A_n *is a conformable matrix with plim*$A_n = A$*. Then* $A_n X_n \xrightarrow{d} N\left(\mu, A\Sigma A^{/}\right)$

Notice that if

$$\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{s_X} \xrightarrow{D} N(0,1)$$

then

$$\frac{n\left(\overline{X}-\mu\right)^2}{s_X^2} \xrightarrow{D} \chi_1^2$$

9.2 Delta Method.

Suppose that

$$\sqrt{n}\left(\overset{\wedge}{\theta}-\theta_0\right) \xrightarrow{D} X$$

where X has a *cdf* F and θ is a $p \times 1$ vector. Suppose that $g : \mathbb{R}^p \to \mathbb{R}^q$. Then

$$\sqrt{n}\left(g\left(\overset{\wedge}{\theta}\right)-g(\theta_0)\right) \xrightarrow{D} \underbrace{\frac{\partial g}{\partial \theta^{/}}(\theta_0)}_{q\times p} \cdot \underset{p\times 1}{X}$$

The proof is by the mean value theorem, i.e.,

$$g\left(\overset{\wedge}{\theta}\right) = g(\theta_0) + \frac{\partial g}{\partial \theta^{/}}\left(\overset{-}{\theta}\right)\left(\overset{\wedge}{\theta}-\theta_0\right)$$

where $\overset{-}{\theta}$ lies between θ_0 and $\overset{\wedge}{\theta}$. Now for $\frac{\partial g}{\partial \theta^{/}}$ continuous at θ_0, we have

$$\overset{-}{\theta} \xrightarrow{P} \theta_0 \Rightarrow \frac{\partial g}{\partial \theta^{/}}\left(\overset{-}{\theta}\right) \xrightarrow{P} \frac{\partial g}{\partial \theta^{/}}(\theta_0).$$

Therefore,

$$\sqrt{n}\left(g\left(\overset{\wedge}{\theta}\right)-g(\theta_0)\right) \xrightarrow{D} \frac{\partial g}{\partial \theta^{/}}(\theta_0)\cdot X$$

as required.

For example, $\sin\overline{X}$ when $\mu = 0$. Now $(\sin x)^{/} = \cos x$. Hence $\sqrt{n}\sin\overline{X} \xrightarrow{D} N(0,1)$.

In fact we can state the following theorem.

Theorem 31 *Suppose that X_n is asymptotically distributed as $N(\mu, \sigma_n^2)$, with $\sigma_n \to 0$. Let g be a real valued function differentiable m $(m \geq 1)$ times at $x = \mu$, with $g^m(\mu) \neq 0$ but $g^j(\mu) = 0$ for $j < m$. Then*

$$\frac{g(x_n) - g(\mu)}{\frac{1}{m!} g^m(\mu) \sigma_n^m} \xrightarrow{d} [N(0,1)]^m .$$

For example, let X_n be asymptotically $N(0, \sigma_n^2)$, with $\sigma_n \to 0$. Then

$$\frac{\log^2(1 + X_n)}{\sigma_n^2} \xrightarrow{d} \chi_1^2.$$

To see this apply the above theorem with $g(x) = \log^2(1+x)$, $m = 2$ and $\mu = 0$.

Chapter 10

ASYMPTOTIC ESTIMATION THEORY

Let $\overset{\wedge}{\theta}_n$ $(p \times 1)$ be an estimator, applied to a sample of size n, of a vector parameter θ_0. Both and must be elements of the set Θ of all admissible values of the parameters, called the parameter space, which can in principle be defined to be $\mathbb{R}^p$, or *p-dimensiona*l Euclidian space. For technical reasons, Θ must be a compact subset of $\mathbb{R}^p$ i.e. bounded and closed, i.e. it contains its boundary points. Furthermore, θ_0 must be an interior point of Θ. This is to say that $\theta_0 \in int(\Theta)$ if there exist a real number $\delta > 0$ such that $\theta \in \Theta$ whenever $\|\theta - \theta_0\| < \delta$. This excludes θ_0 being at the boundary of the set.

Definition 32 $\overset{\wedge}{\theta}_n$ *is a consistent estimator of* θ_0 *if* $plim\overset{\wedge}{\theta}_n = \theta_0$.

Consistency might be a minimum requirement for a useful estimator. Proofs of consistency play an important role in econometric theory. The forms that these proofs is that if $\lim_{n\to\infty} E\left(\overset{\wedge}{\theta}_n\right) = \theta_0$, i.e. the estimator is asymptotically unbiased, and $\lim_{n\to\infty} Var\left(\overset{\wedge}{\theta}_n\right) = 0$ suffices for the consistency of the estimator.

Now suppose $\overset{\wedge}{\theta}_n$ is consistent, and $n^k\left(\overset{\wedge}{\theta}_n - \theta_0\right) = O_p(1)$ for some $k > 0$, and has a non-degenerate limit distribution as $n \to \infty$. This distribution is called the asymptotic distribution of $\overset{\wedge}{\theta}_n$.

Definition 33 $\overset{\wedge}{\theta}_n$ *is said to be consistent and asymptotically normal (CAN) for* $\overset{\wedge}{\theta}_n \in int(\Theta)$ *if there exist* $k > 0$ *such that* $n^k\left(\overset{\wedge}{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, V)$, *where* V *is a finite variance-covariance matrix.*

In most applications $k = 1/2$, although it can be larger than this for models containing determinist trend terms. There are also case where $k > 1/2$ but the limiting distribution is not normal, e.g. when there are stochastic trends. Asymptotic normality is an important property to establish for an estimator, as it is often the only basis for constructing interval estimates and tests of hypotheses.

Let denote by $\mathcal{C}$ the class of CAN estimators of θ_0, and write $\overset{\wedge}{\theta}_n \in \mathcal{C}$ to denote that the estimator belongs to this class.

Definition 34 *$\overset{\wedge}{\theta}_n \in \mathcal{C}$ is said to best asymptotically normal for θ_0 (BAN) in the class if $AVar\left(\overset{\smile}{\theta}_n\right) - AVar\left(\overset{\wedge}{\theta}_n\right)$ is positive semi-definite for every $\overset{\smile}{\theta}_n \in \mathcal{C}$.*

This property is also called asymptotic efficiency. BAN can be seen as an asymptotic counterpart of the BLUE property.

10.1 Asymptotics of the Stochastic Regressor Model

Assume the regression model:

$$y_t = x_t^{/}\beta + u_t \quad t = 1, 2, ..., n$$

Let the following assumptions hold for all $n > k$

$$\begin{aligned} E(\mathbf{u}) &= 0 \quad a.s. \\ E\left(\mathbf{u}\mathbf{u}^{/}\right) &= \sigma^2 I_n \quad a.s. \\ rank(X) &= k \quad a.s. \end{aligned}$$

where $\mathbf{u}$ is the $(n \times 1)$ vector of the errors, X is the $(n \times k)$ matrix of the explanatory variables and I_n is the identity matrix of dimension n. These are the usual assumptions. Furthermore, assume that

$$p \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} x_t x_t^{/} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E\left(x_t x_t^{/}\right) = M_{xx} < \infty (p.d.)$$

and

$$E\left|\lambda^{/} x_t u_t\right|^{2+\delta} \leq B < \infty \quad \delta > 0, \quad \forall \quad fixed \quad \lambda$$

The first additional condition can be written as $plim n^{-1} X^{/} X = M_{xx}$ has two components. The weak law of large numbers must apply to the squares and the cross-products of the elements of x_t, and M_{xx} must have full rank. The latter can fail even if the matrix X has rank k for every finite n. To see this take the fairly trivial example $x_t = 1/t$. Then $\lim_{n\to\infty} \sum_{t=1}^{n} \frac{1}{t^2} = \frac{\pi^2}{6}$. Hence $\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} \frac{1}{t^2} = 0$.

The least squared estimator can be written as

$$\overset{\wedge}{\beta} = \left(\sum_{t=1}^{n} x_t x_t^{/} \right)^{-1} \sum_{t=1}^{n} x_t y_t = \beta + \left(\sum_{t=1}^{n} x_t x_t^{/} \right)^{-1} \sum_{t=1}^{n} x_t u_t$$

Consider now the $k \times 1$ vector $x_t u_t$. Since $E(u_t | x_t) = 0$ and $E(u_t^2 | x_t) = \sigma^2$ the Law of Iterated Expectations gives

$$Var\left(x_t u_t\right) = E\left[E\left(u_t^2 x_t x_t^{/} | x_t \right) \right] = \sigma^2 E\left(x_t x_t^{/} \right) < \infty.$$

Furthermore, the u_t's are independent hence, the Weak Law of Large Numbers can be applied on $x_t u_t$, i.e.

$$p \lim \frac{1}{n} \sum_{t=1}^{n} x_t u_t = 0$$

which is written as $p \lim \frac{1}{n} X^{/} \mathbf{u} = 0$. Then by the Continuous Mapping Theorem we have that

$$p \lim \overset{\wedge}{\beta} - \beta = \left(p \lim \frac{1}{n} X^{/} X \right)^{-1} p \lim \frac{1}{n} X^{/} \mathbf{u} = M_{xx}^{-1} 0 = 0$$

which is the consistency result.

Let us now consider the sequence $\lambda^{/} x_t u_t$. We have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} Var\left(\lambda^{/} x_t u_t \right) = \sigma^2 \lambda^{/} M_{xx} \lambda.$$

Since now M_{xx} is positive definite $0 < \sigma^2 \lambda^{/} M_{xx} \lambda < \infty$. Hence the denominator of the condition of Theorem 20 is bounded and bigger than 0. Furthermore, $E \left| \lambda^{/} x_t u_t \right|^{2+\delta} \leq B$ ensures the condition of the same Theorem and consequently we have that

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \lambda^{/} x_t u_t \xrightarrow{d} N\left(0, \sigma^2 \lambda^{/} M_{xx} \lambda \right)$$

for each specific λ. But this is equivalent to

$$\frac{1}{\sqrt{n}} X^{/}\mathbf{u} \xrightarrow{d} N\left(0, \sigma^2 M_{xx}\right).$$

Finally

$$\sqrt{n}\left(\overset{\wedge}{\beta} - \beta\right) = \left(\frac{1}{n} X^{/} X\right)^{-1} \frac{1}{\sqrt{n}} X^{/}\mathbf{u} \xrightarrow{d} N\left(0, \sigma^2 M_{xx}^{-1}\right).$$

Part IV

Likelihood Function

Chapter 11

MAXIMUM LIKELIHOOD ESTIMATION

Let the observations be $x = (x_1, x_2, ..., x_n)$, and the Likelihood Function be denoted by $L(\theta) = d(x;\theta)$, $\theta \in \Theta \subset \Re^k$. Then the Maximum Likelihood Estimator (MLE) is:

$$\overset{\wedge}{\theta} = \arg\max_{\theta\in\Theta} p(x;\theta) \Leftrightarrow d\left(x;\overset{\wedge}{\theta}\right) \geq d(x;\theta) \quad \forall\theta \in \Theta.$$

If $\overset{\wedge}{\theta}$ is unique, then the model is identified. Let $\ell(\theta) = \ln d(x;\theta)$, then for local identification we have the following Lemma.

Lemma 35 *The model is Locally Identified iff the Hessian is negative definite with probability 1, i.e.*

$$\Pr\left[H\left(\overset{\wedge}{\theta}\right) = \frac{\partial^2 \ell\left(\overset{\wedge}{\theta}\right)}{\partial\theta\partial\theta^{/}} < 0\right] = 1.$$

Assume that the log-Likelihood Function can be written in the following form:

$$\ell(\theta) = \sum_{i=1}^{n} \ln d(x_i;\theta).$$

Then we usually make the following assumptions:

Assumption A1. The range of the random variable x, say C, i.e.

$$C = \{x \in \Re^n : d(x;\theta) > 0\},$$

be independent of the parameter θ.

Assumption A2. The Log-Likelihood Function $\ln d(x;\theta)$ has partial derivatives with respect of θ up to third order, which are bounded and integrable with respect to x.

The vector

$$s(x;\theta) = \frac{\partial \ell(x;\theta)}{\partial \theta}$$

which is $k \times 1$, is called the score vector. We can state the following Lemma.

Lemma 36 *Under the assumptions A1. and A2. we have that*

$$E(s) = 0, \quad E\left(ss^{/}\right) = -E\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta^{/}}\right].$$

Proof: As $L(x,\theta)$ is a density function it follows that

$$\int_C L(x,\theta)\, dx = 1$$

where $C = \{x \in \Re^n : L(x,\theta) > 0\} \subset \Re^n$. Under A1. C is independent of θ. Consequently, the derivative, with respect to θ, of the integral is equal to the integral of the derivative.* Consequently taking derivatives of the above integral we have:

$$\begin{aligned}\frac{\partial}{\partial \theta}\int_C L(x,\theta)\, dx &= 0 \Rightarrow \int_C \frac{\partial L(x,\theta)}{\partial \theta} dx = 0 \Rightarrow \int_C \frac{\partial L}{\partial \theta}\frac{1}{L} L dx = 0 \Rightarrow \int_C \frac{\partial \ln L}{\partial \theta} L dx = 0 \\ &\Rightarrow \int_C \frac{\partial \ell}{\partial \theta} L dx = 0 \Rightarrow \int_C sL dx = 0 \Rightarrow E(s) = 0.\end{aligned}$$

*In case that C was dependent on θ, we would have to apply the Second Fundamental Theorm of Analysis to find the derivative of the integral.

Hence, we also have that $E\left(s^{/}\right)=0$ and taking derivatives with respect to θ we have

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int_{C}s^{/}Ldx &= 0\Rightarrow\frac{\partial}{\partial\theta}\int_{C}\frac{\partial\ell}{\partial\theta^{/}}Ldx=0\Rightarrow\int_{C}\frac{\partial}{\partial\theta}\left(\frac{\partial\ell}{\partial\theta^{/}}L\right)dx=0\\
&\Rightarrow \int_{C}\left[\left(\frac{\partial}{\partial\theta}\frac{\partial\ell}{\partial\theta^{/}}\right)L+\frac{\partial L}{\partial\theta}\left(\frac{\partial\ell}{\partial\theta^{/}}\right)\right]dx=0\\
&\Rightarrow \int_{C}\left[\frac{\partial^{2}\ell}{\partial\theta\partial\theta^{/}}L+\frac{\partial L}{\partial\theta}\frac{1}{L}L\left(\frac{\partial\ell}{\partial\theta^{/}}\right)\right]dx=0\\
&\Rightarrow \int_{C}\left[\frac{\partial^{2}\ell}{\partial\theta\partial\theta^{/}}L+\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta^{/}}L\right]dx=0\\
&\Rightarrow \int_{C}\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta^{/}}Ldx=-\int_{C}\frac{\partial^{2}\ell}{\partial\theta\partial\theta^{/}}Ldx\\
&\Rightarrow \int_{C}ss^{/}Ldx=-\int_{C}\frac{\partial^{2}\ell}{\partial\theta\partial\theta^{/}}Ldx\Rightarrow E\left(ss^{/}\right)=-E\left(\frac{\partial^{2}\ell}{\partial\theta\partial\theta^{/}}\right)
\end{aligned}$$

which is the second result. ■

The matrix

$$J\left(\theta\right)=E\left(ss^{/}\right)=E\left(\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta^{/}}\right)=-E\left(\frac{\partial^{2}\ell}{\partial\theta\partial\theta^{/}}\right)$$

is called (Fisher) Information Matrix and is a measure of the information that the sample x contains about the parameters in θ. In case that $\ell\left(x,\theta\right)$ can be written as $\ell\left(x,\theta\right)=\sum_{i=1}^{n}\ell\left(x_{i},\theta\right)$ we have that

$$J\left(\theta\right)=-E\left(\frac{\partial^{2}}{\partial\theta\partial\theta^{/}}\sum_{i=1}^{n}\ell\left(x_{i},\theta\right)\right)=-\sum_{i=1}^{n}E\left(\frac{\partial^{2}\ell\left(x_{i},\theta\right)}{\partial\theta\partial\theta^{/}}\right)=-nE\left(\frac{\partial^{2}\ell\left(x_{i},\theta\right)}{\partial\theta\partial\theta^{/}}\right).$$

Consequently the Information matrix is proportional to the sample size. Furthermore, by Assumption A2. $E\left(\frac{\partial^{2}\ell(x_{i},\theta)}{\partial\theta\partial\theta^{/}}\right)$ is bounded. Hence

$$J\left(\theta\right)=O\left(n\right).$$

Now we can state the following Lemma that will be needed in the sequel.

Lemma 37 *Under assumptions A1. and A2. we have that for any unbiased estimator of θ, say $\widetilde{\theta}$,*

$$E\left(\widetilde{\theta}s^{/}\right) = I_k,$$

where I_k is the identity matrix of order k.

Proof: As $\widetilde{\theta}$ is an unbiased estimator of θ, we have that

$$E\left(\widetilde{\theta}\right) = \theta \Rightarrow \int_C \widetilde{\theta} L\left(x;\theta\right) dx = \theta.$$

Taking derivatives, with respect to $\theta^{/}$, we have that

$$\begin{aligned}
\frac{\partial}{\partial \theta^{/}} \int_C \widetilde{\theta}\left(x\right) L\left(x;\theta\right) dx &= I_k \Rightarrow \int_C \widetilde{\theta}\left(x\right) \frac{\partial L\left(x;\theta\right)}{\partial \theta^{/}} \frac{1}{L\left(x;\theta\right)} L\left(x;\theta\right) dx = I_k \\
&\Rightarrow \int_C \widetilde{\theta}\left(x\right) \frac{\partial \ell\left(x,\theta\right)}{\partial \theta^{/}} L\left(x;\theta\right) dx = I_k \Rightarrow \int_C \widetilde{\theta}\left(x\right) s^{/} L\left(x;\theta\right) dx = I_k \\
&\Rightarrow E\left(\widetilde{\theta}s^{/}\right) = I_k
\end{aligned}$$

which is the result. ■

We can now prove the Cramer-Rao Theorem.

Theorem 38 *Under the regularity assumptions A1. and A2. we have that for any unbiased estimator $\widetilde{\theta}$ we have that*

$$J^{-1}\left(\theta\right) \leq V\left(\widetilde{\theta}\right).$$

Proof: Let us define the following $2k \times 1$ vector $\xi^{/} = \left(\delta^{/}, s^{/}\right) = \left(\widetilde{\theta}^{/} - \theta^{/}, s^{/}\right)$. Now we have that

$$V\left(\xi\right) = E\left(\xi\xi^{/}\right) = \begin{bmatrix} \delta\delta^{/} & \delta s^{/} \\ s\delta^{/} & ss^{/} \end{bmatrix} = \begin{bmatrix} V\left(\widetilde{\theta}\right) & I_k \\ I_k & J\left(\theta\right) \end{bmatrix}.$$

For the above result we took into consideration that $\widetilde{\theta}$ is unbiased, i.e. $E\left(\widetilde{\theta}\right) = \theta$, the above Lemma, i.e. $E\left(\widetilde{\theta}s^{/}\right) = I_k$, $E\left(s\right) = 0$, and $E\left(ss^{/}\right) = J\left(\theta\right)$. It is known

that all variance-covariance matrices are positive semi-definite. Hence $V(\xi) \geq 0$. Let us define the following matrix

$$B = \left[\begin{array}{cc} I_k & -J^{-1}(\theta) \end{array} \right].$$

The matrix B is $k \times 2k$ and of rank k. Consequently, as $V(\xi) \geq 0$, we have that $BV(\xi) B^{/} \geq 0$. Hence, as I_k and $J^{-1}(\theta)$ are symmetric we have

$$\begin{aligned} BV(\xi) B^{/} &= \left[\begin{array}{cc} I_k & -J^{-1}(\theta) \end{array} \right] \left[\begin{array}{cc} V\left(\widetilde{\theta}\right) & I_k \\ I_k & J(\theta) \end{array} \right] \left[\begin{array}{c} I_k \\ -J^{-1}(\theta) \end{array} \right] \\ &= \left[\begin{array}{cc} V\left(\widetilde{\theta}\right) - J^{-1}(\theta) & 0 \end{array} \right] \left[\begin{array}{c} I_k \\ -J^{-1}(\theta) \end{array} \right] = V\left(\widetilde{\theta}\right) - J^{-1}(\theta) \geq 0. \end{aligned}$$

Consequently, $V\left(\widetilde{\theta}\right) \geq J^{-1}(\theta)$, in the sense that $V\left(\widetilde{\theta}\right) - J^{-1}(\theta)$ is a positive semi-definite matrix. ■

The matrix $J^{-1}(\theta)$ is called the Cramer-Rao Lower Bound as it is the lower bound for any unbiased estimator (either linear or non-linear).

Let, now, θ_0 denote the true parameter values of θ and $d(x; \theta_0)$ be the likelihood function evaluated at the true parameter values. Then for any function $f(x)$ we define

$$E_0[f(x)] = \int f(x) d(x; \theta_0) dx.$$

Let $\ell(\theta)/n$ be the average log-likelihood and define a function $z : \Theta \rightarrow \Re$ as

$$z(\theta) = E_0(\ell(\theta)/n) = \frac{1}{n} \int \ell(\theta) d(x; \theta_0) dx.$$

Then we can state the following Lemma:

Lemma 39 *$\forall \theta \in \Theta$ we have that*

$$z(\theta) \leq z(\theta_0)$$

with strict inequality if

$$\Pr[x \in S : d(x; \theta) \neq d(x; \theta_0)] > 0$$

Proof: From the definition of $z(\theta)$ we have that

$$\begin{aligned} n[z(\theta) - z(\theta_0)] &= E_0[\ell(\theta) - \ell(\theta_0)] = E_0\left[\ln \frac{d(x;\theta)}{d(x;\theta_0)}\right] \leq \ln E_0\left[\frac{d(x;\theta)}{d(x;\theta_0)}\right] \\ &= \ln \int \frac{d(x;\theta)}{d(x;\theta_0)} d(x;\theta_0)\, dx = \ln \int d(x;\theta)\, dx = \ln 1 = 0 \end{aligned}$$

where the inequality is due to Jensen. The inequality is strict when the ratio $\frac{d(x;\theta)}{d(x;\theta_0)}$ is non-constant with probability greater than 0. ■

When the observations $x = (x_1, x_2, ..., x_n)$ are randomly sampled, we have that

$$\ell(\theta) = \sum_{i=1}^{n} z_i \quad where\ z_i = \ln d(x_i;\theta)$$

and the z_i random variables are independent and have the same distribution with mean $E(z_i) = z(\theta)$. Then from the Weak Law of Large Numbers we have that

$$p\lim \frac{1}{n}\sum_{i=1}^{n} z_i = p\lim \frac{1}{n}\ell(\theta) = z(\theta).$$

However, the above is true under weaker assumptions, e.g. for dependent observations or for non-identically distributed random variables etc. To avoid a lengthy exhibition of various cases we make the following assumption:

Assumption A3. Θ is a compact subset of $\Re^k$, and

$$p\lim \frac{1}{n}\ell(\theta) = z(\theta) \quad \forall \theta \in \Theta.$$

Recall that a closed and bounded subset of $\Re^k$ is compact.

Theorem 40 *Under the above assumption and if the statistical model is identified we have that*

$$\overset{\wedge}{\theta} \xrightarrow{p} \theta_0$$

where $\overset{\wedge}{\theta}$ *is the MLE and* θ_0 *the true parameter values.*

Proof: Let N is an open sphere with centre θ_0 and radius ε, i.e. $N = \{\theta \in \Theta : \|\theta - \theta_0\| < \varepsilon\}$. Then $\overline{N}$ is closed and consequently $A = \overline{N} \cap \Theta$ is closed and bounded, i.e. compact. Hence

$$\max_{\theta \in A} z(\theta)$$

exist and we can define

$$\delta = z(\theta_0) - \max_{\theta \in A} z(\theta) \tag{11.1}$$

Let $T_\delta \subset S$ the event (a subset of the sample space) which defined by

$$\forall \theta \in \Theta \quad \left| \frac{1}{n} \ell(\theta) - z(\theta) \right| < \frac{\delta}{2}. \tag{11.2}$$

Hence (11.2) applies for $\theta = \overset{\wedge}{\theta}$ as well. Hence

$$for \quad T_\delta \Rightarrow \quad z\left(\overset{\wedge}{\theta}\right) > \frac{1}{n} \ell\left(\overset{\wedge}{\theta}\right) - \frac{\delta}{2}.$$

Given now that $\ell\left(\overset{\wedge}{\theta}\right) \geq \ell(\theta_0)$ we have that

$$for \quad T_\delta \Rightarrow \quad z\left(\overset{\wedge}{\theta}\right) > \frac{1}{n} \ell(\theta_0) - \frac{\delta}{2}.$$

Furthermore, as $\theta_0 \in \Theta$, we have that

$$for \quad T_\delta \Rightarrow \quad \frac{1}{n} \ell(\theta_0) > z(\theta_0) - \frac{\delta}{2}.$$

from the relationship that if $|x| < d \Rightarrow -d < x < d$. Adding the above two inequalities we have that

$$for \quad T_\delta \Rightarrow \quad z\left(\overset{\wedge}{\theta}\right) > z(\theta_0) - \delta.$$

Substituting out δ, employing (11.1) we get

$$for \quad T_\delta \Rightarrow \quad z\left(\overset{\wedge}{\theta}\right) > \max_{\theta \in A} z(\theta).$$

Hence

$$\overset{\wedge}{\theta} \notin A \Rightarrow \overset{\wedge}{\theta} \notin \overline{N} \cap \Theta \Rightarrow \overset{\wedge}{\theta} \in N \cap \Theta \Rightarrow \overset{\wedge}{\theta} \in N.$$

Consequently we have shown that when T_δ is true then $\overset{\wedge}{\theta} \in N$. This implies that

$$\Pr\left(\overset{\wedge}{\theta} \in N\right) \geq \Pr\left(T_\delta\right)$$

and taking limits, as $n \to \infty$ we have

$$\lim_{n\to\infty} \Pr\left(\|\theta - \theta_0\| < \varepsilon\right) \geq \lim_{n\to\infty} \Pr\left(T_\delta\right) = 1$$

by the definition of N and by assumption A3. Hence, as ε is any small positive number, we have

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \Pr\left(\|\theta - \theta_0\| < \varepsilon\right) = 1$$

which is the definition of probability limit. ■

When the observations $x = (x_1, x_2, ..., x_n)$ are randomly sampled, we have :

$$\begin{aligned}
\ell\left(\theta\right) &= \sum_{i=1}^{n} \ln d\left(x_i;\theta\right) \\
s\left(\theta\right) &= \frac{\partial \ell\left(\theta\right)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln d\left(x_i;\theta\right)}{\partial \theta} \\
H\left(\theta\right) &= \frac{\partial^2 \ell\left(\theta\right)}{\partial \theta \partial \theta^{/}} = \sum_{i=1}^{n} \frac{\partial^2 \ln d\left(x_i;\theta\right)}{\partial \theta \partial \theta^{/}}.
\end{aligned}$$

Given now that the observations x_i are independent and have the same distribution, with density $d\left(x_i;\theta\right)$, the same is true for the vectors $\frac{\partial \ln d(x_i;\theta)}{\partial \theta}$ and the matrices $\frac{\partial^2 \ln d(x_i;\theta)}{\partial \theta \partial \theta^{/}}$. Consequently we can apply a Central Limit Theorem to get:

$$n^{-1/2} s\left(\theta\right) = n^{-1/2} \sum_{i=1}^{n} \frac{\partial \ln d\left(x_i;\theta\right)}{\partial \theta} \xrightarrow{d} N\left(0, \bar{J}\left(\theta\right)\right)$$

and from the Law of Large Numbers

$$n^{-1} H\left(\theta\right) = n^{-1} \sum_{i=1}^{n} \frac{\partial^2 \ln d\left(x_i;\theta\right)}{\partial \theta \partial \theta^{/}} \xrightarrow{p} -\bar{J}\left(\theta\right)$$

where

$$\bar{J}\left(\theta\right) = n^{-1} J\left(\theta\right) = n^{-1} J\left(\theta\right) = -n^{-1} E\left(\frac{\partial^2 \ell\left(\theta\right)}{\partial \theta \partial \theta^{/}}\right),$$

i.e., the average Information matrix. However, the two asymptotic results apply even if the observations are dependent or identically distributed. To avoid a lengthy exhibition of various cases we make the following assumptions. As $n \to \infty$ we have:

Assumption A4.

$$n^{-1/2} s(\theta) \xrightarrow{d} N\left(0, \bar{J}(\theta)\right)$$

and

Assumption A5.

$$n^{-1} H(\theta) \xrightarrow{p} -\bar{J}(\theta)$$

where

$$\bar{J}(\theta) = J(\theta)/n = E\left(ss'\right)/n = -E(H)/n.$$

We can now state the following Theorem

Theorem 41 *Under assumptions A2. and A3.the above two assumptions and identification we have:*

$$\sqrt{n}\left(\overset{\wedge}{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, \left(\bar{J}(\theta_0)\right)^{-1}\right]$$

where $\overset{\wedge}{\theta}$ *is the MLE and* θ_0 *the true parameter values.*

Proof: As $\overset{\wedge}{\theta}$ maximises the Likelihood Function we have from the first order conditions that

$$s\left(\overset{\wedge}{\theta}\right) = \frac{\partial \ell\left(\overset{\wedge}{\theta}\right)}{\partial \theta} = 0.$$

From the Mean Value Theorem, around θ_0, we have that

$$s(\theta_0) + H(\theta_*)\left(\overset{\wedge}{\theta} - \theta_0\right) = 0 \tag{11.3}$$

where $\theta_* \in \left[\overset{\wedge}{\theta}, \theta_0\right]$, i.e. $\|\theta_* - \theta_0\| \leq \left\|\overset{\wedge}{\theta} - \theta_0\right\|$.

Now from the consistency of the MLE we have that

$$\overset{\wedge}{\theta} = \theta_0 + o_p(1)$$

where$o_p(1)$ is a random variable that goes to 0 in probability as $n \to \infty$. As now $\theta_* \in \left[\overset{\wedge}{\theta}, \theta_0\right]$, we have that

$$\theta_* = \theta_0 + o_p(1)$$

as well. Hence from 11.3 we have that

$$\sqrt{n}\left(\overset{\wedge}{\theta} - \theta_0\right) = -\left[H(\theta_*)/n\right]^{-1} s(\theta_0)/\sqrt{n}.$$

As now $\theta_* = \theta_0 + o_p(1)$ and under the second assumption we have that

$$\sqrt{n}\left(\overset{\wedge}{\theta} - \theta_0\right) = \left[\bar{J}(\theta_0)\right]^{-1} s(\theta_0)/\sqrt{n} + o_p(1).$$

Under now the first assumption the above equation implies that $\sqrt{n}\left(\overset{\wedge}{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, \left(\bar{J}(\theta_0)\right)^{-1}\right]$. ■

Example: Let $y_t \backsim N(\mu, \sigma^2)$ i.i.d for $t = 1, ..., T$. Then

$$\ell(\theta) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \mu)^2$$

where $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$. Now

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{t=1}^{T}(y_t - \mu)$$

and

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{-T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T}(y_t - \mu)^2.$$

Now

$$\frac{\partial \ell\left(\overset{\wedge}{\theta}\right)}{\partial \mu} = \frac{\partial \ell\left(\overset{\wedge}{\theta}\right)}{\partial \sigma^2} = 0.$$

Hence

$$\overset{\wedge}{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t \quad and \quad \overset{\wedge}{\sigma^2} = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \overset{\wedge}{\mu}\right)^2.$$

Now

$$H(\theta) = \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'} = \begin{bmatrix} -\frac{T}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{t=1}^{T}(y_t-\mu) \\ -\frac{1}{\sigma^4}\sum_{t=1}^{T}(y_t-\mu) & \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{T}(y_t-\mu)^2 \end{bmatrix},$$

and consequently evaluating $H(\theta)$ at $\overset{\wedge}{\theta}$ we have

$$H\left(\overset{\wedge}{\theta}\right) = \begin{bmatrix} -\frac{T}{\overset{\wedge}{\sigma^2}} & 0 \\ 0 & -\frac{T}{2\overset{\wedge}{\sigma^4}} \end{bmatrix}$$

which is clearly negative definite. Now the Information matrix is:

$$\begin{aligned} J(\theta) &= -E(H(\theta)) = E\left(\begin{bmatrix} \frac{T}{\sigma^2} & \frac{1}{\sigma^4}\sum_{t=1}^{T}(y_t-\mu) \\ \frac{1}{\sigma^4}\sum_{t=1}^{T}(y_t-\mu) & -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{t=1}^{T}(y_t-\mu)^2 \end{bmatrix}\right) \\ &= \begin{bmatrix} \frac{T}{\sigma^2} & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix} \end{aligned}$$

Chapter 12

RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION

Let us assume that the $k \times 1$ vector of parametersθ_0 satisfy r constrains, i.e.

$$\varphi(\theta) = 0$$

where $\varphi(\theta)$ and 0 are $r \times 1$ vectors. Let us also assume that

$$F(\theta) = \frac{\partial \varphi(\theta)}{\partial \theta^{/}}$$

the $r \times k$ matrix of derivatives has rank r, i.e. there no redundant constrain.

Under these conditions, MLE is still consistent but not is not asymptotically efficient, as it ignores the information in the constrains. To get an asymptotically efficient estimator we have to take into consideration the information of the constrains. Hence we form the Lagrangian:

$$L(\theta, \lambda) = \ell(\theta) + \lambda^{/}\varphi(\theta) = \ell(\theta) + \varphi^{/}(\theta)\lambda$$

where λ is the $r \times 1$ vector of Lagrange Multipliers. The first order conditions are:

$$\begin{aligned} \frac{\partial L}{\partial \theta} &= \frac{\partial \ell}{\partial \theta} + F^{/}(\theta)\lambda = s(\theta) + F^{/}(\theta)\lambda = 0 \\ \frac{\partial L}{\partial \lambda} &= \frac{\partial \lambda^{/}}{\partial \lambda}\varphi(\theta) = I_k \varphi(\theta) = \varphi(\theta) = 0. \end{aligned}$$

Let $\widetilde{\theta}$ and $\widetilde{\lambda}$ the constrained ML estimators, i.e. the solution of the above first order conditions, i.e.

$$s\left(\widetilde{\theta}\right) + F^{/}\left(\widetilde{\theta}\right)\widetilde{\lambda} = 0 \tag{12.1}$$

$$\varphi\left(\widetilde{\theta}\right)=0. \tag{12.2}$$

Applying the Mean Value Theorem around θ_0 we have to $s\left(\widetilde{\theta}\right)$ and $\varphi\left(\widetilde{\theta}\right)$ we get

$$s\left(\widetilde{\theta}\right)=s\left(\theta_0\right)+\left[\frac{\partial s\left(\theta_*\right)}{\partial\theta^{/}}\right]\left(\widetilde{\theta}-\theta_0\right)=s\left(\theta_0\right)+H\left(\theta_*\right)\left(\widetilde{\theta}-\theta_0\right) \tag{12.3}$$

$$\varphi\left(\widetilde{\theta}\right)=\varphi\left(\theta_0\right)+\left[\frac{\partial\varphi\left(\theta_{**}\right)}{\partial\theta^{/}}\right]\left(\widetilde{\theta}-\theta_0\right)=\varphi\left(\theta_0\right)+F\left(\theta_{**}\right)\left(\widetilde{\theta}-\theta_0\right) \tag{12.4}$$

where θ_* and θ_{**} are defined as

$$\begin{aligned}\left\|\theta_*-\theta_0\right\| &\leq \left\|\widetilde{\theta}-\theta_0\right\| \\ \left\|\theta_{**}-\theta_0\right\| &\leq \left\|\widetilde{\theta}-\theta_0\right\|.\end{aligned}$$

Notice that θ_* and θ_{**} are not necessarily the same.

Substituting (12.3) into (12.1), (12.4) into (12.2) and taking into account that under the null $\varphi\left(\theta_0\right)=0$ we get

$$s\left(\theta_0\right)+H\left(\theta_*\right)\left(\widetilde{\theta}-\theta_0\right)+F^{/}\left(\widetilde{\theta}\right)\widetilde{\lambda}=0$$

and

$$F\left(\theta_{**}\right)\left(\widetilde{\theta}-\theta_0\right)=0.$$

Hence we get

$$\frac{1}{\sqrt{n}}s\left(\theta_0\right)+\frac{1}{n}H\left(\theta_*\right)\sqrt{n}\left(\widetilde{\theta}-\theta_0\right)+\frac{1}{\sqrt{n}}F^{/}\left(\widetilde{\theta}\right)\widetilde{\lambda}=0 \tag{12.5}$$

and

$$\sqrt{n}F\left(\theta_{**}\right)\left(\widetilde{\theta}-\theta_0\right)=0. \tag{12.6}$$

Now $\widetilde{\theta}$ is consistent and so are θ_* and θ_{**}, i.e.

$$\widetilde{\theta}=\theta_0+o_p\left(1\right),\quad \theta_*=\theta_0+o_p\left(1\right),\quad \textit{and}\quad \theta_{**}=\theta_0+o_p\left(1\right).$$

Furthermore, according to our assumptions we have that

$$\frac{1}{n}H\left(\theta_*\right)=-\bar{J}\left(\theta_0\right)+o_p\left(1\right)$$

where $\bar{J}(\theta_0)$ the average information matrix.

Hence, equations (12.5) and (12.6) become:

$$\frac{s(\theta_0)}{\sqrt{n}} - \bar{J}(\theta_0)\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) + F^{/}(\theta_0)\frac{\widetilde{\lambda}}{\sqrt{n}} = o_p(1) \tag{12.7}$$

and

$$F(\theta_0)\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) = o_p(1). \tag{12.8}$$

Let us now define the matrix

$$P(\theta_0) = F(\theta_0)\left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) > 0$$

as $\bar{J}(\theta_0) > 0$ and $F(\theta_0)$ has rank r. Now multiplying (12.7) by $F(\theta_0)\left[\bar{J}(\theta_0)\right]^{-1}$ and add (12.8) we get:

$$F(\theta_0)\left[\bar{J}(\theta_0)\right]^{-1}\frac{s(\theta_0)}{\sqrt{n}} + P(\theta_0)\frac{\widetilde{\lambda}}{\sqrt{n}} = o_p(1).$$

Hence, as $P(\theta_0) > 0$ we have that

$$\frac{\widetilde{\lambda}}{\sqrt{n}} = -\left[P(\theta_0)\right]^{-1} F(\theta_0)\left[\bar{J}(\theta_0)\right]^{-1}\frac{s(\theta_0)}{\sqrt{n}} + o_p(1). \tag{12.9}$$

Furthermore, substituting into (12.7) we get

$$\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) = \left\{I_k - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0)\left[P(\theta_0)\right]^{-1} F(\theta_0)\right\}\left[\bar{J}(\theta_0)\right]^{-1}\frac{s(\theta_0)}{\sqrt{n}} + o_p(1). \tag{12.10}$$

Now we can state the following Theorem:

Theorem 42 *Under the assumptions A4., A5., model identification and that the true parameter values satisfy the constrains, we have*

$$\frac{\widetilde{\lambda}}{\sqrt{n}} \xrightarrow{d} N\left(0, \left[P(\theta_0)\right]^{-1}\right)$$

and

$$\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \left[\bar{J}(\theta_0)\right]^{-1} - A\right)$$

where

$$A = \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0)\left[P(\theta_0)\right]^{-1} F(\theta_0)\left[\bar{J}(\theta_0)\right]^{-1} > 0.$$

Proof: From assumption A4. we have that

$$n^{-1/2}s(\theta) \xrightarrow{d} N\left(0, \bar{J}(\theta)\right) \Rightarrow \left[\bar{J}(\theta_0)\right]^{-1/2} \frac{s(\theta_0)}{\sqrt{n}} \xrightarrow{d} N(0, I_k).$$

Hence from (12.9) we have

$$\frac{\widetilde{\lambda}}{\sqrt{n}} = -\left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1/2} \left[\bar{J}(\theta_0)\right]^{-1/2} \frac{s(\theta_0)}{\sqrt{n}} + o_p(1).$$

Hence

$$\frac{\widetilde{\lambda}}{\sqrt{n}} \xrightarrow{d} N(0, \Omega_1)$$

where

$$\Omega_1 = \left\{-\left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1/2}\right\} \left\{-\left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1/2}\right\}^{/} = \left[P(\theta_0)\right]^{-1}.$$

Furthermore from (12.10) we have

$$\begin{aligned}\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) &= \left\{I_k - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0)\right\} \left[\bar{J}(\theta_0)\right]^{-1/2} \\ &\quad \times \left[\bar{J}(\theta_0)\right]^{-1/2} \frac{s(\theta_0)}{\sqrt{n}} + o_p(1).\end{aligned}$$

Hence

$$\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) \xrightarrow{d} N(0, \Omega_2)$$

where

$$\begin{aligned}\Omega_2 &= \left\{\left\{I_k - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0)\right\} \left[\bar{J}(\theta_0)\right]^{-1/2}\right\} \\ &\quad \times \left\{\left\{I_k - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0)\right\} \left[\bar{J}(\theta_0)\right]^{-1/2}\right\}^{/} \\ &= \left[\bar{J}(\theta_0)\right]^{-1} - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1} \\ &\quad - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1} \\ &\quad \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1} \\ &= \left[\bar{J}(\theta_0)\right]^{-1} - \left[\bar{J}(\theta_0)\right]^{-1} F^{/}(\theta_0) \left[P(\theta_0)\right]^{-1} F(\theta_0) \left[\bar{J}(\theta_0)\right]^{-1}\end{aligned}$$

which is the result. ■

Hence we can reach the following Conclusion:

Corollary 43 *The Restricted MLE is at least as efficient as the MLE, i.e.*

$$AsyVar\left(\widetilde{\theta}\right) \leq AsyVar\left(\overset{\wedge}{\theta}\right).$$

Part V

Neyman or Ratio of the Likelihoods Tests

Let $x = (x_1, x_2, ..., x_n)$ be a random sample having a density function $d(x, \theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$. We would like to test a simple hypothesis

$$H_0 : \theta = \theta_0$$

against a simple alternative

$$H_1 : \theta = \theta_1 \in \Theta - \{\theta_0\}.$$

To simplify the notation we write

$$d_0(x) = d(x, \theta_0), \quad d_1(x) = d(x, \theta_1).$$

The Neyman Ratio is the statistic

$$\lambda(x) = \frac{d_1(x)}{d_0(x)} = \frac{d(x, \theta_1)}{d(x, \theta_0)}.$$

If S is the sample space of x, then the Neyman Ratio Test is defined by the Rejection Region

$$R = \{x \in S : \lambda(x) \geq c_\alpha\}$$

where $c_\alpha > 0$ is a constant such that the Type I Error (Size) of the test is equal to $\alpha \in (0, 1/2)$, i.e.

$$P(R|\theta = \theta_0) = \int_R d_0(x)\, dx = \alpha,$$

i.e. the probability of rejecting the null when it is correct. Notice that the Power of the above test, $\pi_R(\alpha)$ is

$$\pi_R(\alpha) = P(R|\theta = \theta_1) = \int_R d_1(x)\, dx,$$

i.e. the probability of rejecting the null when the alternative is correct.

Lemma 44 *(Neyman-Pearson) Let A be the Rejection Region of the test of H_0 with size less or equal to α, i.e.*

$$P(A|\theta = \theta_0) = \int_A d_0(x)\, dx \leq \alpha.$$

Then the Power of this test is less or equal to the Power of the Neyman Test, i.e.

$$P(A|\theta = \theta_1) \leq P(R|\theta = \theta_1).$$

Proof: The difference in the Powers of the 2 tests is

$$\pi_R(\alpha) - \pi_A(\alpha) = P(R|\theta = \theta_1) - P(A|\theta = \theta_1) =$$

$$\int_R d_1(x)\,dx - \int_A d_1(x)\,dx =$$

$$\int_{\overline{A}\cap R} d_1(x)\,dx + \int_{A\cap R} d_1(x)\,dx - \int_{\overline{R}\cap A} d_1(x)\,dx - \int_{R\cap A} d_1(x)\,dx =$$

$$\int_{\overline{A}\cap R} d_1(x)\,dx - \int_{\overline{R}\cap A} d_1(x)\,dx.$$

Form the definition of the Rejection Region R we have that

$$\forall x \in R \quad d_1(x) \geq c_\alpha d_0(x), \quad \textit{and} \quad \forall x \in \overline{R} \quad d_1(x) < c_\alpha d_0(x).$$

Substituting to the above integrals we get

$$\pi_R(\alpha) - \pi_A(\alpha) \geq \int_{\overline{A}\cap R} c_\alpha d_0(x)\,dx - \int_{\overline{R}\cap A} c_\alpha d_0(x)\,dx =$$

$$c_\alpha\left(\int_{\overline{A}\cap R} d_0(x)\,dx + \int_{A\cap R} d_0(x)\,dx - \int_{\overline{R}\cap A} d_0(x)\,dx - \int_{A\cap R} d_0(x)\,dx\right) =$$

$$c_\alpha\left(\int_R d_0(x)\,dx - \int_A d_0(x)\,dx\right) \geq c_\alpha(\alpha - \alpha) = 0.$$

Hence $\pi_R(\alpha) \geq \pi_A(\alpha)$.

The above Lemma says that when we test a simple null versous a simple alternative the Neyman Test is the Most Powerful Test of all tests that have the same or smaller size.

However, the usual tests are of simple H_0 versous composite alternative. In such cases the power of the test is, in general, a function of the parameters in the alternative, i.e. $\pi_A(\alpha, \theta)$ the power of the test is a function of the parameters as well.

Definition 45 *Let A be the rejection region of a test. Then if $\pi_A(\alpha, \theta) \geq \alpha$ for all $\theta \in \Theta$ the test is unbiased.*

The comparison of unbiased tests is very difficult as the one is more powerful in one region of the parametric space and the other in another.

Definition 46 *If for any test, say B, we have that $\pi_A(\alpha,\theta) \geq \pi_B(\alpha,\theta)$ for all $\theta \in \Theta$ the test A is called uniformly most powerful test.*

There is one way to find uniformly most powerful tests or alternatively to compare the power of any given test. Assume that we have to test the null $H_0 : \theta = 0$ versous the alternative $H_1 : \theta \neq 0$. We can calculate the power of the Neyman Ratio Test of the simple null $H_0 : \theta = 0$ versous the simple alternative $H_1 : \theta = 1$. and graph the point in a $\pi_R(\alpha,\theta) - \theta$ diagram. We repeat for the simple alternatives $H_1 : \theta = -1$, $H_1 : \theta = 2$, $H_1 : \theta = -2$, etc. The line we get by joining all these points together is called the Envelope Power Function. As an immediate consequence of the Neyman-Pearson Lemma we have:

Theorem 47 *Let $\pi_A(\alpha,\theta)$ be the power function of a test of size α of the null hypothesis $H_0 : \theta = \theta_0$ versous the alternative $H_1 : \theta \in \Theta - \{\theta_0\}$. If $e_\alpha(\theta)$ is the Envelope Power Function we have that*

$$\forall \theta \in \Theta \quad \pi_A(\alpha,\theta) \leq e_\alpha(\theta).$$

For hypothesis testing the Envelope Power Function is the analogue of the Cramer-Rao bound in estimation. It is obvious that if the power function of a test is identical to the Envelope Power Function then this test is uniformly most powerful in the class of unbiased tests. Hence we have the following Corollary

Corollary 48 *If the Rejection Region of a Neyman Ratio test of a simple null $H_0 : \theta = \theta_0$ versous the alternative $H_1 : \theta \in \Theta_1$ does not depend on θ_1, a random point is Θ_1, then the Neyman test is uniformly most powerful test.*

The proof is based on the fact that if the rejection region R does not depend on θ_1, the power of the test is identical to the Envelope Power Function.

Example 49 *Let $x = (x_1, x_2, ..., x_n)$ be a random sample from a $N(\mu, \sigma^2)$ with unknown μ and known σ^2. We want to test the hypothesis $H_0 : \mu = \mu_0$ versous the*

alternative $H_1 : \mu > \mu_0$. *For any* $\mu_1 > \mu$ *the Likelihood Functions for* μ_j $(j = 0, 1)$ *is given by*

$$d_j(\theta) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu_j\right)^2\right]$$

and the ratio of the likelihoods is

$$\lambda(x) = d_1(x)/d_0(x) = \exp\left[\frac{1}{\sigma^2}(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\left(\mu_1^2 - \mu_0^2\right)\right].$$

Hence $\lambda(x) \geq c_\alpha$ *is equivalent to*

$$\frac{1}{\sigma^2}(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}\left(\mu_1^2 - \mu_0^2\right) \geq \ln c_\alpha \Leftrightarrow$$

$$\sum_{i=1}^{n} x_i \geq \frac{\sigma^2}{\mu_1 - \mu_0}\ln c_\alpha + \frac{n(\mu_1 + \mu_0)}{2} \Leftrightarrow \overline{x} \geq c_\alpha^{/} = \frac{\sigma^2 \ln c_\alpha}{n(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2}$$

where the constants c_α *and* $c_\alpha^{/}$ *are determined by the size of the test, which should be* α, *i.e.*

$$P\left(\overline{x} \geq c_\alpha^{/} | \mu = \mu_0\right) = \alpha.$$

Under H_0, *we have that* $\overline{x} \backsim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ *and* $\sqrt{n}\frac{\overline{x} - \mu_0}{\sigma} \backsim N(0, 1)$. *The above probability is equivalent to*

$$P\left(\sqrt{n}\frac{\overline{x} - \mu_0}{\sigma} \geq \sqrt{n}\frac{c_\alpha^{/} - \mu_0}{\sigma} | \mu = \mu_0\right) = \alpha.$$

Let z_α *be the critical value that leaves* $\alpha\%$ *in the tail of a* $N(0, 1)$ *distribution. Hence*

$$P\left(\sqrt{n}\frac{\overline{x} - \mu_0}{\sigma} \geq z_\alpha | \mu = \mu_0\right) = \alpha.$$

Hence we have shown that

$$\lambda(x) \geq c_\alpha \Leftrightarrow \sqrt{n}\frac{\overline{x} - \mu_0}{\sigma} \geq z_\alpha.$$

The Rejection Region of the Ratio of the Neyman Test is

$$R = \{x \in S : \lambda(x) \geq c_\alpha\} = \left\{x \in S : \sqrt{n}\frac{\overline{x} - \mu_0}{\sigma} \geq z_\alpha\right\}$$

which is independent of μ_1. *Consequently, the test is Uniformly Most Powerful.*

The Neyman Ratio Test is generalised and for the testing of a composite null versous composite alternative. Let $x = (x_1, x_2, ..., x_n)$ be a sample with Likelihood Function

$$d(x;\theta), \quad \theta \in \Theta \subset \mathbb{R}^k,$$

and $\Omega \subset \mathbb{R}^\nu$ $(\nu < k)$ be a subset of the parametric space Θ. We would like to test the composite null

$$H_0 : \theta \in \Omega$$

versous the alternative

$$H_1 : \theta \in \Theta - \Omega.$$

The Neyman Ratio is the function

$$\lambda(x) = \frac{\sup_{\theta\in\Theta} d(x;\theta)}{\sup_{\theta\in\Omega} d(x;\theta)}$$

and the Neyman Ratio Test is defined by the Rejection Region of H_0

$$R = \{x \in S : \lambda(x) \geq c_\alpha\},$$

where $c_\alpha > 0$ is a constant such that the probability of Type I Error is α, i.e.

$$\sup_{\theta\in\Omega} \{P(R|\theta \in \Omega)\} = \sup_{\theta\in\Omega} \int_R d(x;\theta)\, dx = \alpha.$$

Example 50 *Let $x = (x_1, x_2, ..., x_n)$ be a random sample from a $N(\mu, \sigma^2)$ with unknown μ and σ^2. We want to test the hypothesis $H_0 : \mu = \mu_0$ versous the alternative $H_1 : \mu \neq \mu_0$. The vector of parameters is*

$$\theta' = \left(\mu, \sigma^2\right) \in \Theta = \mathbb{R} \times \mathbb{R}_+.$$

Under the null the parameter sample is $\Omega = \mu_0 \times \mathbb{R}_+$. Consequently we have that $H_0 : \theta \in \Omega$ versous $H_1 : \theta \in \Theta - \Omega$. For any θ the likelihood function is

$$d(x;\theta) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

However, the maximum of $\sup_{\theta\in\Omega}\left[d\left(x;\theta\right)\right]=\sup_{\sigma^2}\left[d\left(x;\mu_0,\sigma^2\right)\right]$. *The maximum of* $d\left(x;\mu_0,\sigma^2\right)$ *is at*

$$\sigma^2=s_0^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2$$

Hence

$$\sup_{\theta\in\Omega}\left[d\left(x;\theta\right)\right]=\left(2\pi s_0^2\right)^{-n/2}\exp\left[-\frac{n}{2}\right].$$

With the same reasoning we find that

$$\sup_{\theta\in\Theta}\left[d\left(x;\theta\right)\right]=\left(2\pi s^2\right)^{-n/2}\exp\left[-\frac{n}{2}\right],\quad s^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2.$$

Hence the Ratio of the Likelihoods is

$$\lambda\left(x\right)=\frac{\sup_{\theta\in\Theta-\Omega}\left[d\left(x;\theta\right)\right]}{\sup_{\theta\in\Omega}\left[d\left(x;\theta\right)\right]}=\left(s^2/s_0^2\right)^{-n/2},$$

and the Rejection Region of the test is

$$R=\left\{x\in S:\left(s_0^2/s^2\right)^{n/2}\geq c_\alpha\right\}=\left\{x\in S:\frac{s_0^2-s^2}{s^2}\geq c_\alpha^{2/n}-1\right\}.$$

Notice that

$$\begin{aligned}\frac{s_0^2-s^2}{s^2} &= \frac{\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2-\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}=\frac{n\left(\overline{x}-\mu_0\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}=>\\ \frac{s_0^2-s^2}{s^2} &= \frac{n\left(\overline{x}-\mu_0\right)^2}{\left(n-1\right)\overset{\wedge}{\sigma^2}},\quad where\quad \overset{\wedge}{\sigma^2}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\end{aligned}$$

Hence

$$R=\left\{x\in S:\frac{n\left(\overline{x}-\mu_0\right)^2}{\overset{\wedge}{\sigma^2}}\geq\left(n-1\right)\left(c_\alpha^{2/n}-1\right)\right\}.$$

It is easy to show that under H_0 *the ratio*

$$F=\frac{n\left(\overline{x}-\mu_0\right)^2}{\overset{\wedge}{\sigma^2}}\backsim F_{1,n-1}.$$

Hence the null is rejected when

$$F=\frac{n\left(\overline{x}-\mu_0\right)^2}{\overset{\wedge}{\sigma^2}}\geq f_\alpha$$

where f_α *the* $\alpha\%$ *critical value of an* $F_{1,n-1}$.

Chapter 13

χ-SQUARE TESTS

Let $\theta \in \Theta \subset \mathbb{R}^k$ be a vector of parameters and $\Omega \subset \mathbb{R}^\nu$ $(\nu < k)$ be a set subset of the space of Θ from points that satisfy $r = \nu - k$ non-linear equations constraints such that

$$\varphi(\theta) = 0$$

where φ is a $r \times 1$ vector of functions and 0 an $r \times 1$ vector of zeros. Furthermore, let $\overset{\wedge}{\theta}$ be a Consistent Asymptotically Normal estimator, i.e.

$$\sqrt{n}\left(\overset{\wedge}{\theta} - \theta\right) \xrightarrow{d} N(0, V),$$

where V is a known or consistently estimated variance matrix.

We want to test $H_0 : \varphi(\theta) = 0$ versous $H_1 : \varphi(\theta) \neq 0$. Assume that the $r \times k$ matrix

$$F(\theta) = \frac{\partial \varphi}{\partial \theta'} = \left\{ \frac{\partial \varphi_i}{\partial \theta_j}, \quad i = 1, ..., r, \quad j = 1, 2, ..k \right\}$$

has full row rank, i.e.

$$rank\left[F(\theta)\right] = r.$$

This is fulfilled if there are no redundant restrictions.

Theorem 51 *For a Consistent Asymptotically Normal (CAN) estimator $\overset{\wedge}{\theta}$, assuming that $rank\left[F(\theta)\right] = r$ and under $H_0 : \varphi(\theta) = 0$, we have that*

$$n\varphi\left(\overset{\wedge}{\theta}\right)' \left(FVF'\right)^{-1} \varphi\left(\overset{\wedge}{\theta}\right) \xrightarrow{d} \chi^2_r$$

Proof. From the Delta Method we have that if $\sqrt{n}\left(\overset{\wedge}{\theta}-\theta\right) \xrightarrow{d} X$ then $\sqrt{n}\left(\varphi\left(\overset{\wedge}{\theta}\right)-\varphi(\theta)\right) \xrightarrow{d} F(\theta)\cdot X$. But under H_0 $\varphi(\theta)=0$. Hence $\sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right) \xrightarrow{d} F(\theta)\cdot X \Rightarrow \sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right) \xrightarrow{d} N\left(0, FVF^{/}\right)$. It follows that $n\varphi\left(\overset{\wedge}{\theta}\right)^{/}\left(FVF^{/}\right)^{-1}\varphi\left(\overset{\wedge}{\theta}\right) \xrightarrow{d} \chi^2_r$ where $r=rank\left(FVF^{/}\right)=r$. In case that the matrices F and V are functions of the unknown parameter vector $\overset{\wedge}{\theta}$, can be substituted by consistent estimators, e.g. $F\left(\overset{\wedge}{\theta}\right)$ and $V\left(\overset{\wedge}{\theta}\right)$. ■

The χ^2 test does not necessarily have the optimal properties of the test of the likelihood ratio, however, it has the great advantage that we do not have to know the exact distribution of the sample. What is necessary is a CAN estimator. Furthermore, it is fairly easy to construct and evaluate such a test. Consequently, it is not a surprise that χ^2 tests are the most popular tests in statistics.

Chapter 14

THE CLASSICAL TESTS

Let the null hypothesis be represented by

$$\Omega = \{\theta \in \Theta : \varphi(\theta) = 0\}$$

where θ is the vector of parameters and $\varphi(\theta) = 0$ are the restrictions. Consequently the Neyman ratio test is given by:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)} = \frac{L\left(\overset{\wedge}{\theta}\right)}{L\left(\widetilde{\theta}\right)}$$

where $L(\theta)$ is the Likelihood function. As now the $\ln(\cdot)$ is a monotonic, strictly increasing, an equivalent test can be based on

$$LR = 2\ln(\lambda(x)) = 2\left[\ell\left(\overset{\wedge}{\theta}\right) - \ell\left(\widetilde{\theta}\right)\right]$$

where where LR is the well known Likelihood Ratio test and $\ell(\theta)$ is the log-likelihood function.

Using a Taylor expansion of $\ell\left(\widetilde{\theta}\right)$ around $\overset{\wedge}{\theta}$ and employing the Mean Value Theorem we get:

$$\ell\left(\widetilde{\theta}\right) = \ell\left(\overset{\wedge}{\theta}\right) + \frac{\partial \ell\left(\overset{\wedge}{\theta}\right)}{\partial \theta^{/}}\left(\widetilde{\theta} - \overset{\wedge}{\theta}\right) + \frac{1}{2}\left(\widetilde{\theta} - \overset{\wedge}{\theta}\right)^{/} \frac{\partial^2 \ell(\theta_*)}{\partial \theta \partial \theta^{/}}\left(\widetilde{\theta} - \overset{\wedge}{\theta}\right)$$

where $\left\|\theta_* - \overset{\wedge}{\theta}\right\| \leq \left\|\widetilde{\theta} - \overset{\wedge}{\theta}\right\|$. Now, $\frac{\partial \ell\left(\overset{\wedge}{\theta}\right)}{\partial \theta^{/}} = s\left(\overset{\wedge}{\theta}\right) = 0$ due to the fact the the first order conditions are satisfied by the ML estimator $\overset{\wedge}{\theta}$. Consequently, the LR test is

given by:

$$LR = 2\left[\ell\left(\overset{\wedge}{\theta}\right) - \ell\left(\widetilde{\theta}\right)\right] = -\left(\widetilde{\theta} - \overset{\wedge}{\theta}\right)^{/} \frac{\partial^2 \ell\left(\theta_*\right)}{\partial\theta\partial\theta^{/}} \left(\widetilde{\theta} - \overset{\wedge}{\theta}\right).$$

Now we know that

$$\sqrt{n}\left(\widetilde{\theta} - \theta_0\right) = \left\{I_k - \left[\bar{J}\left(\theta_0\right)\right]^{-1} F^{/}\left(\theta_0\right)\left[P\left(\theta_0\right)\right]^{-1} F\left(\theta_0\right)\right\}\left[\bar{J}\left(\theta_0\right)\right]^{-1} \frac{s\left(\theta_0\right)}{\sqrt{n}} + o_p\left(1\right)$$

and

$$\sqrt{n}\left(\overset{\wedge}{\theta} - \theta_0\right) = \left[\bar{J}\left(\theta_0\right)\right]^{-1} \frac{s\left(\theta_0\right)}{\sqrt{n}} + o_p\left(1\right).$$

Hence

$$\sqrt{n}\left(\widetilde{\theta} - \overset{\wedge}{\theta}\right) = -\left[\bar{J}\left(\theta_0\right)\right]^{-1} F^{/}\left(\theta_0\right)\left[P\left(\theta_0\right)\right]^{-1} F\left(\theta_0\right)\left[\bar{J}\left(\theta_0\right)\right]^{-1} \frac{s\left(\theta_0\right)}{\sqrt{n}} + o_p\left(1\right)$$

and consequently

$$\begin{aligned} LR &= \left(\left[\bar{J}\left(\theta_0\right)\right]^{-1} F^{/}\left(\theta_0\right)\left[P\left(\theta_0\right)\right]^{-1} F\left(\theta_0\right)\left[\bar{J}\left(\theta_0\right)\right]^{-1} \frac{s\left(\theta_0\right)}{\sqrt{n}}\right)^{/} \left(-\frac{1}{n}\frac{\partial^2 \ell\left(\theta_*\right)}{\partial\theta\partial\theta^{/}}\right) \\ &\quad \left[\bar{J}\left(\theta_0\right)\right]^{-1} F^{/}\left(\theta_0\right)\left[P\left(\theta_0\right)\right]^{-1} F\left(\theta_0\right)\left[\bar{J}\left(\theta_0\right)\right]^{-1} \frac{s\left(\theta_0\right)}{\sqrt{n}} + o_p\left(1\right). \end{aligned}$$

Now from assumption A5. we have

$$n^{-1} H\left(\theta\right) = -\bar{J}\left(\theta\right) + o_p\left(1\right),$$

$$\theta_* = \theta_0 + o_p\left(1\right) \quad \textit{and} \quad P\left(\theta_0\right) = F\left(\theta_0\right)\left[\bar{J}\left(\theta_0\right)\right]^{-1} F^{/}\left(\theta_0\right).$$

Hence

$$LR = \left(\frac{s\left(\theta_0\right)}{\sqrt{n}}\right)^{/} \left[\bar{J}\left(\theta_0\right)\right]^{-1} F^{/}\left(\theta_0\right)\left[P\left(\theta_0\right)\right]^{-1} F\left(\theta_0\right)\left[\bar{J}\left(\theta_0\right)\right]^{-1} \frac{s\left(\theta_0\right)}{\sqrt{n}} + o_p\left(1\right).$$

We can now state the following Theorem:

Theorem 52 *Under the usual assumptions and under the null Hypothesis we have that*

$$LR = 2\left[\ell\left(\overset{\wedge}{\theta}\right) - \ell\left(\widetilde{\theta}\right)\right] \xrightarrow{d} \chi^2_r.$$

Proof: The Likelihood Ratio is written as

$$LR = (\xi_0)^{/} Z_0 \left[Z_0^{/} Z_0 \right]^{-1} Z_0^{/} \xi_0 + o_p(1)$$

where

$$\left[\bar{J}(\theta_0) \right]^{-1/2} \frac{s(\theta_0)}{\sqrt{n}} = \xi_0, \quad and \quad Z_0 = \left[\bar{J}(\theta_0) \right]^{-1/2} F^{/}(\theta_0).$$

Now

$$\left[\bar{J}(\theta_0) \right]^{-1/2} \frac{s(\theta_0)}{\sqrt{n}} \xrightarrow{d} N(0, I_k)$$

and $Z_0 \left[Z_0^{/} Z_0 \right]^{-1} Z_0^{/}$ is symmetric idempotent. Hence

$$r\left(Z_0 \left[Z_0^{/} Z_0 \right]^{-1} Z_0^{/} \right) = tr\left(Z_0 \left[Z_0^{/} Z_0 \right]^{-1} Z_0^{/} \right) = tr\left(\left[Z_0^{/} Z_0 \right]^{-1} Z_0^{/} Z_0 \right) = tr(I_r) = r.$$

Consequently, we get the result. ■

The Wald test is based on the idea that if the restrictions are correct the vector $\varphi\left(\overset{\wedge}{\theta}\right)$ should be close to zero.

Expanding $\varphi\left(\overset{\wedge}{\theta}\right)$ around $\varphi(\theta_0)$ we get:

$$\varphi\left(\overset{\wedge}{\theta}\right) = \varphi(\theta_0) + \frac{\partial \varphi(\theta_*)}{\partial \theta^{/}} \left(\overset{\wedge}{\theta} - \theta_0\right) = F(\theta_*) \left(\overset{\wedge}{\theta} - \theta_0\right)$$

as under the null $\varphi(\theta_0) = 0$. Hence

$$\sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right) = \sqrt{n} F(\theta_*) \left(\overset{\wedge}{\theta} - \theta_0\right)$$

and consequently

$$\sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right) = \sqrt{n} F(\theta_0) \left(\overset{\wedge}{\theta} - \theta_0\right) + o_p(1).$$

Furthermore recall that

$$\sqrt{n}\left(\overset{\wedge}{\theta} - \theta_0\right) \xrightarrow{d} N\left[0, \left(\bar{J}(\theta_0)\right)^{-1}\right].$$

Hence,

$$\sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right) \xrightarrow{d} N\left[0, F(\theta_0) \left(\bar{J}(\theta_0)\right)^{-1} F^{/}(\theta_0)\right].$$

Let us now consider the following quadratic:

$$n\left[\varphi\left(\overset{\wedge}{\theta}\right)\right]^{/}\left[F\left(\theta_0\right)\left(\bar{J}\left(\theta_0\right)\right)^{-1}F^{/}\left(\theta_0\right)\right]^{-1}\varphi\left(\overset{\wedge}{\theta}\right),$$

which is the square of the Mahalanobis distance of the $\sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right)$ vector. However the above quantity can not be considered as a statistic as it is a function of the unknown parameter θ_0. The Wald test is given by the above quantity if the unknown vector of parameters θ_0 is substituted by the ML estimator $\overset{\wedge}{\theta}$, i.e.

$$\begin{aligned} W &= \left[\varphi\left(\overset{\wedge}{\theta}\right)\right]^{/}\left[F\left(\overset{\wedge}{\theta}\right)\left(n\bar{J}\left(\overset{\wedge}{\theta}\right)\right)^{-1}F^{/}\left(\overset{\wedge}{\theta}\right)\right]^{-1}\varphi\left(\overset{\wedge}{\theta}\right) \\ &= \left[\varphi\left(\overset{\wedge}{\theta}\right)\right]^{/}\left[F\left(\overset{\wedge}{\theta}\right)\left(J\left(\overset{\wedge}{\theta}\right)\right)^{-1}F^{/}\left(\overset{\wedge}{\theta}\right)\right]^{-1}\varphi\left(\overset{\wedge}{\theta}\right), \end{aligned}$$

where $J\left(\overset{\wedge}{\theta}\right)$ is the estimated information matrix. In case that $J\left(\overset{\wedge}{\theta}\right)$ does not have an explicit formula it can be substituted by a consistent estimator, e.g. by

$$\hat{J} = -\sum_{i=1}^{n}\frac{\partial^2\ell\left(\overset{\wedge}{\theta}\right)}{\partial\theta\partial\theta^{/}}$$

or by the asymptotically equivalent

$$\hat{J} = \sum_{i=1}^{n}\frac{\partial\ell\left(\overset{\wedge}{\theta}\right)}{\partial\theta}\frac{\partial\ell\left(\overset{\wedge}{\theta}\right)}{\partial\theta^{/}}.$$

Hence the Wald statistic is given by

$$W = \left[\varphi\left(\overset{\wedge}{\theta}\right)\right]^{/}\left[F\left(\overset{\wedge}{\theta}\right)\left(\hat{J}\right)^{-1}F^{/}\left(\overset{\wedge}{\theta}\right)\right]^{-1}\varphi\left(\overset{\wedge}{\theta}\right).$$

Now we can prove the following Theorem:

Theorem 53 *Under the usual regularity assumptions and the null hypothesis we that*

$$W = \left[\varphi\left(\overset{\wedge}{\theta}\right)\right]^{/}\left[F\left(\overset{\wedge}{\theta}\right)\left(\hat{J}\right)^{-1}F^{/}\left(\overset{\wedge}{\theta}\right)\right]^{-1}\varphi\left(\overset{\wedge}{\theta}\right)\xrightarrow{d}\chi^2_r.$$

Proof: For any consistent estimator of θ_0 we have that

$$F\left(\overset{\wedge}{\theta}\right)\left(\hat{J}\right)^{-1}F^{/}\left(\overset{\wedge}{\theta}\right)=F\left(\theta_0\right)\left(n\bar{J}\left(\theta_0\right)\right)^{-1}F^{/}\left(\theta_0\right)+o_p\left(1\right).$$

Hence

$$W=n\left[\varphi\left(\overset{\wedge}{\theta}\right)\right]^{/}\left[F\left(\theta_0\right)\left(\bar{J}\left(\theta_0\right)\right)^{-1}F^{/}\left(\theta_0\right)\right]^{-1}\varphi\left(\overset{\wedge}{\theta}\right)+o_p\left(1\right).$$

Furthermore,

$$\sqrt{n}\varphi\left(\overset{\wedge}{\theta}\right)\overset{d}{\rightarrow}N\left[0,F\left(\theta_0\right)\left(\bar{J}\left(\theta_0\right)\right)^{-1}F^{/}\left(\theta_0\right)\right],$$

and the result follows. ■

The Lagrange Multiplier (LM) test considers the distance from zero of the estimated Lagrange Multipliers. Recall that

$$\frac{\widetilde{\lambda}}{\sqrt{n}}\overset{d}{\rightarrow}N\left(0,\left[P\left(\theta_0\right)\right]^{-1}\right).$$

Consequently, the square Mahalanobis distance is

$$\left(\frac{\widetilde{\lambda}}{\sqrt{n}}\right)^{/}P\left(\theta_0\right)\left(\frac{\widetilde{\lambda}}{\sqrt{n}}\right)=\left(\widetilde{\lambda}\right)^{/}F\left(\theta_0\right)\left[n\bar{J}\left(\theta_0\right)\right]^{-1}F^{/}\left(\theta_0\right)\left(\widetilde{\lambda}\right).$$

Again, the above quantity is not a statistic as it is a function of the unknown parameters θ_0. However, we can employ the restricted ML estimates of θ_0 to find the the unknown quantities, i.e. $\widetilde{F}=F\left(\widetilde{\theta}\right)$ and $\widetilde{J}=J\left(\widetilde{\theta}\right)$. Hence we can prove the following:

Theorem 54 *Under the usual regularity assumptions and the null hypothesis we have*

$$LM=\left(\widetilde{\lambda}\right)^{/}\widetilde{F}\left[\widetilde{J}\right]^{-1}\widetilde{F}^{/}\left(\widetilde{\lambda}\right)\overset{d}{\rightarrow}\chi_r^2.$$

Proof: Again we have that for any consistent estimator of θ_0, as is the restricted MLE $\widetilde{\theta}$, we have that

$$LM=\left(\widetilde{\lambda}\right)^{/}\widetilde{F}\left[\widetilde{J}\right]^{-1}\widetilde{F}^{/}\left(\widetilde{\lambda}\right)=\left(\frac{\widetilde{\lambda}}{\sqrt{n}}\right)^{/}P\left(\theta_0\right)\left(\frac{\widetilde{\lambda}}{\sqrt{n}}\right)+o_p\left(1\right)$$

and by the asymptotic distribution of the Lagrange Multipliers we get the result. ■

Now we have that the Restricted MLE satisfy the first order conditions of the Lagrangian, i.e.

$$s\left(\widetilde{\theta}\right)+F^{/}\left(\widetilde{\theta}\right)\widetilde{\lambda}=0.$$

Consequently the LM test can be expressed as:

$$LM=\left(s\left(\widetilde{\theta}\right)\right)^{/}\left[\widetilde{J}\right]^{-1}s\left(\widetilde{\theta}\right).$$

Now Rao has suggested to find the score vector and the information matrix of the unrestricted model and evaluate them at the restricted MLE. Under this form the LM statistic is called **efficient score statistic** as it measures the distance of the score vector, evaluated at the restricted MLE, from zero.

14.1 The Linear Regression

Let us consider the classical linear regression model:

$$y=X\beta+u,\quad u|X\backsim N\left(0,\sigma^{2}I_{n}\right)$$

where y is the $n\times1$ vector of endogenous variables, X is the $n\times k$ matrix of weakly exogenous explanatory variables, β is the $k\times1$ vector of mean parameters and u is the $n\times1$ vector of errors. Let us call the vector of parameters θ, i.e. $\theta^{/}=\left(\beta^{/},\sigma^{2}\right)$ a $(k+1)\times1$ vector. The log-likelihood function is:

$$\ell\left(\theta\right)=-\frac{n}{2}\ln\left(2\pi\right)-\frac{n}{2}\ln\left(\sigma^{2}\right)-\frac{1}{2}\frac{\left(y-X\beta\right)^{/}\left(y-X\beta\right)}{\sigma^{2}}.$$

The first order conditions are:

$$\frac{\partial\ell\left(\theta\right)}{\partial\beta}=\frac{X^{/}\left(y-X\beta\right)}{\sigma^{2}}=0$$

and

$$\frac{\partial\ell\left(\theta\right)}{\partial\sigma^{2}}=-\frac{n}{2\sigma^{2}}+\frac{1}{2}\frac{\left(y-X\beta\right)^{/}\left(y-X\beta\right)}{\sigma^{4}}=0.$$

Solving the equations we get:

$$\begin{aligned} \overset{\wedge}{\beta} &= \left(X^{/}X\right)^{-1}X^{/}y \\ \overset{\wedge}{\sigma^2} &= \frac{\hat{u}^{/}\hat{u}}{n}, \quad \hat{u} = y - X\overset{\wedge}{\beta}. \end{aligned}$$

Notice that the MLE of β is the same as OLS estimator. Something which is not true for the MLE of σ^2.

The Hessian is

$$H\left(\theta\right) = \frac{\partial^2 \ell\left(\theta\right)}{\partial\theta\partial\theta^{/}} = \begin{pmatrix} \frac{\partial^2 \ell(\theta)}{\partial\beta\partial\beta^{/}} & \frac{\partial^2 \ell(\theta)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 \ell(\theta)}{\partial\sigma^2\partial\beta^{/}} & \frac{\partial^2 \ell(\theta)}{\partial(\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2}X^{/}X & -\frac{1}{2\sigma^4}X^{/}u \\ -\frac{1}{2\sigma^4}u^{/}X & \frac{n}{2\sigma^4} - \frac{u^{/}u}{\sigma^6} \end{pmatrix}.$$

Hence the Information matrix is

$$J\left(\theta\right) = E\left[-H\left(\theta\right)\right] = \begin{pmatrix} \frac{1}{\sigma^2}X^{/}X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

and the Cramer-Rao limit

$$J^{-1}\left(\theta\right) = \begin{pmatrix} \sigma^2\left(X^{/}X\right)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

Notice that under normality, of the errors, the OLS estimator is asymptotically efficient.

Let us now consider r linear constrains on the parameter vector β, i.e.

$$\varphi\left(\beta\right) = Q\beta - q = 0 \tag{14.1}$$

where Q is the $r \times k$ matrix of the restrictions (with $r < k$) and q a known vector. Let us now form the Lagrangian, i.e.

$$L = \ell\left(\theta\right) + \lambda^{/}\varphi\left(\beta\right) = \ell\left(\theta\right) + \varphi^{/}\left(\beta\right)\lambda = \ell\left(\theta\right) + \left(Q\beta - q\right)^{/}\lambda,$$

where λ is the vector of the r Lagrange Multipliers. The first order conditions are:

$$\frac{\partial L}{\partial\beta} = \frac{\partial\ell\left(\theta\right)}{\partial\beta} + Q^{/}\lambda = \frac{X^{/}\left(y - X\beta\right)}{\sigma^2} + Q^{/}\lambda = 0 \tag{14.2}$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{\partial \ell(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\frac{(y - X\beta)^{/}(y - X\beta)}{\sigma^4} = 0 \tag{14.3}$$

and

$$\frac{\partial L}{\partial \lambda} = Q\beta - q = 0. \tag{14.4}$$

Now from (14.2) we have that

$$X^{/}y = X^{/}X\beta - \sigma^2 Q^{/}\lambda \tag{14.5}$$

and it follows that

$$Q\left(X^{/}X\right)^{-1}X^{/}y = Q\left(X^{/}X\right)^{-1}X^{/}X\beta - \sigma^2 Q\left(X^{/}X\right)^{-1}Q^{/}\lambda.$$

Hence

$$Q\left(X^{/}X\right)^{-1}X^{/}y = Q\beta - \sigma^2 Q\left(X^{/}X\right)^{-1}Q^{/}\lambda.$$

It follows that

$$Q\beta - QVQ^{/}\lambda = Q\overset{\wedge}{\beta},$$

where

$$\overset{\wedge}{\beta} = \left(X^{/}X\right)^{-1}X^{/}y \quad and \quad V = \sigma^2\left(X^{/}X\right)^{-1}.$$

Now from (14.4) we have that $Q\beta = q$. Hence we get

$$\lambda = -\left[QVQ^{/}\right]^{-1}\left(Q\overset{\wedge}{\beta} - q\right). \tag{14.6}$$

Substituting out λ from (14.5) employing the above and solving for β we get:

$$\widetilde{\beta} = \overset{\wedge}{\beta} - \left(X^{/}X\right)^{-1}Q^{/}\left[Q\left(X^{/}X\right)^{-1}Q^{/}\right]^{-1}\left(Q\overset{\wedge}{\beta} - q\right).$$

Solving (14.3) we get that

$$\widetilde{\sigma^2} = \frac{(\widetilde{u})^{/}\widetilde{u}}{n}, \quad \widetilde{u} = y - X\widetilde{\beta},$$

and from (14.6) we get:

$$\widetilde{\lambda} = -\left[Q\widetilde{V}Q^{/}\right]^{-1}\left(Q\overset{\wedge}{\beta} - q\right), \quad \widetilde{V} = \widetilde{\sigma^2}\left(X^{/}X\right)^{-1}.$$

The above 3 formulae give the restricted MLEs.

Now the Wald test for the linear restrictions in (14.1) is given by

$$W = \left(Q\hat{\beta} - q\right)' \left[Q\hat{V}Q'\right]^{-1} \left(Q\hat{\beta} - q\right).$$

The restricted and unrestricted residuals are given by

$$\tilde{u} = y - X\tilde{\beta}, \quad and \quad \hat{u} = y - X\hat{\beta}.$$

Hence

$$\tilde{u} = \hat{u} + X\left(\hat{\beta} - \tilde{\beta}\right)$$

and consequently, if $X'\hat{u} = 0$, i.e. the regression has a constant we have that

$$\tilde{u}'\tilde{u} = \hat{u}'\hat{u} + \left(\hat{\beta} - \tilde{\beta}\right)' X'X \left(\hat{\beta} - \tilde{\beta}\right).$$

It follows that

$$\tilde{u}'\tilde{u} - \hat{u}'\hat{u} = \left(Q\hat{\beta} - q\right)' \left[Q\left(X'X\right)^{-1}Q'\right]^{-1} \left(Q\hat{\beta} - q\right).$$

Hence the Wald test is given by

$$W = n\frac{\tilde{u}'\tilde{u} - \hat{u}'\hat{u}}{\hat{u}'\hat{u}}.$$

The LR test is given by

$$LR = 2\left[\ell\left(\hat{\theta}\right) - \ell\left(\tilde{\theta}\right)\right] = n\ln\left(\frac{\tilde{u}'\tilde{u}}{\hat{u}'\hat{u}}\right)$$

and the LM test is

$$LM = n\frac{\tilde{u}'\tilde{u} - \hat{u}'\hat{u}}{\tilde{u}'\tilde{u}}$$

as

$$LM = \left(\tilde{\lambda}\right)' \tilde{F}\left[\tilde{J}\right]^{-1} \tilde{F}'\left(\tilde{\lambda}\right) = \left(Q\hat{\beta} - q\right)' \left[Q\tilde{V}Q'\right]^{-1} \left(Q\hat{\beta} - q\right).$$

We can now state a well known result.

Theorem 55 *Under the classical assumptions of the Linear Regression Model we have that*

$$W \geq LR \geq LM.$$

Proof: The three test can be written as

$$W = n(r-1), \quad LR = n\ln(r), \quad LM = n\left(1 - \frac{1}{r}\right),$$

where $r = \frac{\widetilde{u}'\widetilde{u}}{\widehat{u}'\widehat{u}} \geq 1$. Now we know that $\ln(x) \geq \frac{x-1}{x}$ and the result follows by considering $x = r$ and $x = 1/r$.

14.2 Autocorrelation

Apply the LM test to test the hypothesis that $\rho = 0$ in the following model

$$y_t = x_t'\beta + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{i.i.d.}{\backsim} N\left(0, \sigma^2\right).$$

Discuss the advantages of this LM test over the Wald and LR tests of this hypothesis.

First notice that from $u_t = \rho u_{t-1} + \varepsilon_t$ we get that

$$E(u_t) = \rho E(u_{t-1}) + E(\varepsilon_t) = \rho E(u_{t-1})$$

as $E(\varepsilon_t) = 0$ and for $|\rho| < 1$ we get that

$$E(u_t) - \rho E(u_{t-1}) = 0 \Rightarrow E(u_t) = 0$$

as $E(u_t) = E(u_{t-1})$ independent of t. Furthermore

$$Var(u_t) = E\left(u_t^2\right) = \rho^2 E\left(u_{t-1}^2\right) + E\left(\varepsilon_t^2\right) + 2\rho E(u_{t-1}\varepsilon_t) = \rho^2 E\left(u_{t-1}^2\right) + \sigma^2$$

as the first equality follows from the fact that $E(u_t) = 0$, and the last from the fact that

$$E(u_{t-1}\varepsilon_t) = E\left[u_{t-1}E(\varepsilon_t | I_{t-1})\right] = E\left[u_{t-1}0\right] = 0$$

where I_{t-1} the information set at time $t-1$, i.e. the sigma-field generated by $\{\varepsilon_{t-1}, \varepsilon_{t-2}, ...\}$. Hence

$$E\left(u_t^2\right) - \rho^2 E\left(u_{t-1}^2\right) = \sigma^2 \Rightarrow E\left(u_t^2\right) = \frac{\sigma^2}{1-\rho^2}$$

as $E\left(u_t^2\right) = E\left(u_{t-1}^2\right)$ independent of t.

Substituting out u_t we get

$$y_t = x_t^{/}\beta + \rho u_{t-1} + \varepsilon_t,$$

and observing that $u_{t-1} = y_{t-1} - x_{t-1}^{/}\beta$ we get

$$y_t = x_t^{/}\beta + \rho\left(y_{t-1} - x_{t-1}^{/}\beta\right) + \varepsilon_t \Rightarrow \varepsilon_t = y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho$$

where by assumption the $\varepsilon_t's$ are i.i.d. Hence the log-likelihood function is

$$l\left(\theta\right) = -\frac{T}{2}\ln\left(2\pi\right) - \frac{T}{2}\ln\left(\sigma^2\right) - \sum_{t=1}^{T}\frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)^2}{2\sigma^2},$$

where we assume that $y_{-1} = 0$, and $x_{-1} = 0$. as we do not have any observations for $t = -1$. In any case, given that $|\rho| < 1$, the first observation will not affect the distribution LM test, as it is based in asymptotic theory, i.e. $T \to \infty$. The first order conditions are:

$$\frac{\partial l}{\partial \beta} = \sum_{t=1}^{T}\frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)\left(x_t - x_{t-1}\rho\right)}{\sigma^2}$$

$$\frac{\partial l}{\partial \rho} = \sum_{t=1}^{T}\frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)\left(y_{t-1} - x_{t-1}^{/}\beta\right)}{\sigma^2} = \sum_{t=1}^{T}\frac{\varepsilon_t u_{t-1}}{\sigma^2},$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \sum_{t=1}^{T}\frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)^2}{2\sigma^4} = -\frac{T}{2\sigma^2} + \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^4}.$$

The second derivatives are:

$$\frac{\partial^2 l}{\partial \beta \partial \beta^{/}} = -\sum_{t=1}^{T}\frac{\left(x_t - x_{t-1}\rho\right)\left(x_t^{/} - x_{t-1}^{/}\rho\right)}{\sigma^2}$$

$$\frac{\partial^2 l}{\partial \rho^2} = -\sum_{t=1}^{T} \frac{\left(y_{t-1} - x_{t-1}^{/}\beta\right)^2}{\sigma^2} = -\sum_{t=1}^{T} \frac{u_{t-1}^2}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2} = \frac{T}{2\sigma^4} - \sum_{t=1}^{T} \frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)^2}{\sigma^6} = \frac{T}{2\sigma^4} - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{\sigma^6}.$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \beta \partial \rho} &= -\sum_{t=1}^{T} \frac{\left(y_{t-1} - x_{t-1}^{/}\beta\right)\left(x_t^{/} - x_{t-1}^{/}\rho\right) + \left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)\left(x_{t-1}^{/}\right)}{\sigma^2} = \\ &= -\sum_{t=1}^{T} \frac{u_{t-1}\left(x_t^{/} - x_{t-1}^{/}\rho\right) + \varepsilon_t\left(x_{t-1}^{/}\right)}{\sigma^2} \end{aligned}$$

$$\frac{\partial^2 l}{\partial \rho \partial \sigma^2} = -\sum_{t=1}^{T} \frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)\left(y_{t-1} - x_{t-1}^{/}\beta\right)}{\sigma^4} = -\sum_{t=1}^{T} \frac{\varepsilon_t u_{t-1}}{\sigma^4},$$

$$\frac{\partial^2 l}{\partial \beta \partial \sigma^2} = -\sum_{t=1}^{T} \frac{\left(y_t - x_t^{/}\beta - \rho y_{t-1} + x_{t-1}^{/}\beta\rho\right)\left(x_t^{/} - x_{t-1}^{/}\rho\right)}{\sigma^4} = -\sum_{t=1}^{T} \frac{\varepsilon_t\left(x_t^{/} - x_{t-1}^{/}\rho\right)}{\sigma^4}$$

Notice now that the Information Matrix J is

$$J(\theta) = -E\left[H(\theta)\right] = \begin{bmatrix} \sum_{t=1}^{T} \frac{(x_t - x_{t-1}\rho)\left(x_t^{/} - x_{t-1}^{/}\rho\right)}{\sigma^2} & 0 & 0 \\ 0 & \frac{T}{1-\rho^2} & 0 \\ 0 & 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

as $E\left[\frac{u_{t-1}\left(x_t^{/} - x_{t-1}^{/}\rho\right) + \varepsilon_t\left(x_{t-1}^{/}\right)}{\sigma^2}\right] = 0$, $E\left[\frac{\varepsilon_t\left(x_t^{/} - x_{t-1}^{/}\rho\right)}{\sigma^4}\right] = 0$, $E\left[\sum_{t=1}^{T} \frac{\varepsilon_t u_{t-1}}{\sigma^4}\right] = 0$, $E\left[\frac{\varepsilon_t^2}{\sigma^6}\right] = \frac{1}{\sigma^4}$, $E\left[\frac{u_{t-1}^2}{\sigma^2}\right] = \frac{E\left(u_{t-1}^2\right)}{\sigma^2} = \frac{1}{1-\rho^2}$, i.e. the matrix is block diagonal between β, ρ, and σ^2. Consequently the LM test has the form

$$LM = s_\rho^{/} J_{\rho\rho}^{-1} s_\rho = \frac{\left(s_\rho\right)^2}{J_{\rho\rho}}$$

as $s_\rho = \sum_{t=1}^{T} \frac{\varepsilon_t u_{t-1}}{\sigma^2}$, $J_{\rho\rho} = \frac{T}{1-\rho^2}$. All these quantities evaluated under the null.

Hence under $H_0 : \rho = 0$ we have that

$$J_{\rho\rho} = T, \quad and \quad u_t = \varepsilon_t$$

i.e. there is no autocorrelation. Consequently, we can estimate β by simple OLS, as OLS and ML result in the same estimators and σ^2 by the ML estimator, i.e.

$$\widetilde{\beta} = \left(\sum_{t=1}^{T} x_t x_t^{/} \right)^{-1} \sum_{t=1}^{T} x_t y_t, \quad and \quad \widetilde{\sigma^2} = \frac{\widetilde{u}^{/}\widetilde{u}}{T} = \frac{\sum_{t=1}^{T} \widetilde{u}_t^2}{T},$$

where $\widetilde{u}_t = y_t - x_t^{/}\widetilde{\beta} = \widetilde{\varepsilon}_t$ the OLS residuals. Hence

$$LM = \frac{\left(\sum_{t=1}^{T} \frac{\widetilde{u}_t \widetilde{u}_{t-1}}{\widetilde{\sigma^2}} \right)^2}{T} = T \left(\sum_{t=1}^{T} \widetilde{u}_t \widetilde{u}_{t-1} \right)^2 \left(\sum_{t=1}^{T} \widetilde{u}_t^2 \right)^{-2}.$$

Book References

1. T. Amemiya: Advanced Econometrics.

2. E. Berndt: The Practice of Econometrics: Classic and Cotemporary

3. G. Box and G. Jenkins (1976) TimeSeries Analysis forecasting and Control.

K. Cuthbertson, S.G. Hall and M. P. Taylor: Applied Econometric Techniques.

4. R. Davidson and J. MacKinnon: Econometric Theory and Methods.

5. C. Gourieroux and A. Monfort: Statistics and Econometric Models, Vol I and II.

6. W.H. Greene: Econometric Analysis.

7. J. Hamilton Time Series Analysis

8. A. Harvey: The Econometric Analysis of Time Series.

9. A. Harvey: Time Series Models.

10. J. Johnston: Econometric Methods.

11. G. Judge, R. Hill, H. Lutkepohl and T. Lee: Introduction to the Theory and Practice of Econometrics.

12. R. Pindyck and D. Rubinfeld: Econometric Models and Economic Forecasts.

13. P. Ruud: An Introduction to Classical Econometric Theory.

14. R. Serfling: Approximation Theorems of Mathematical Statistics.

15. H White: Asymptotic Theory for Econometricians.