Chapter 2

Statistical Concepts

In this chapter we review some probability and statistics results to be used in subsequent chapters. We focus on the properties of multivariate random vectors and discuss the basic ideas of parameter estimation and hypothesis testing. Most of statistics textbooks also cover similar topics but are in the univariate context; beginning graduate students in economics may find Amemiya (1994) a useful reference. More advanced topics in probability theory will be left to Chapter 5.

2.1 Distribution Functions

Given a random experiment, let Ω denote the collection of all its possible outcomes and $\mathbb{P}$ denote the probability measure assigned to sets of outcomes (subsets of Ω); a subset of Ω is referred to as an *event*. For the event A, $\mathbb{P}(A)$ is a measure of the likelihood of A such that $0 \leq \mathbb{P}(A) \leq 1$. The larger is $\mathbb{P}(A)$, the more likely is the event A to occur. A d-dimensional random vector ($\mathbb{R}^d$-valued random variable) is a function defined on Ω and takes values in $\mathbb{R}^d$. Thus, the value of a random vector depends on the random outcome $\omega \in \Omega$. Formal definitions of probability space and random variables are given in Section 5.1.

The (joint) *distribution function* of the $\mathbb{R}^d$-valued random variable $\boldsymbol{z}$ is the non-decreasing, right-continuous function $F_{\boldsymbol{z}}$ such that for $\boldsymbol{\zeta} = (\zeta_1 \ \ldots \ \zeta_d)' \in \mathbb{R}^d$,

$$F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \mathbb{P}\{\omega \in \Omega \colon z_1(\omega) \leq \zeta_1, \ldots, z_d(\omega) \leq \zeta_d\},$$

with

$$\lim_{\zeta_1 \to -\infty, \ldots, \zeta_d \to -\infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 0, \qquad \lim_{\zeta_1 \to \infty, \ldots, \zeta_d \to \infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 1.$$

Note that the distribution function of $\boldsymbol{z}$ is a standard point function defined on $\mathbb{R}^d$ and provides a convenient way to characterize the randomness of $\boldsymbol{z}$. The (joint) *density function*

of $F_{\boldsymbol{z}}$, if exists, is the non-negative function $f_{\boldsymbol{z}}$ such that

$$F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \int_{-\infty}^{\zeta_d} \cdots \int_{-\infty}^{\zeta_1} f_{\boldsymbol{z}}(s_1, \ldots, s_d)\, \mathrm{d}s_1 \cdots \mathrm{d}s_d,$$

where the right-hand side is a Riemann integral. Clearly, the density function $f_{\boldsymbol{z}}$ must be integrated to one on $\mathbb{R}^d$.

The *marginal distribution function* of the i^{th} component of $\boldsymbol{z}$ is

$$F_{z_i}(\zeta_i) = \mathbb{P}\{\omega \in \Omega \colon z_i(\omega) \leq \zeta_i\} = F_{\boldsymbol{z}}(\infty, \ldots, \infty, \zeta_i, \infty, \ldots, \infty).$$

Thus, the marginal distribution function of z_i is the joint distribution function without restrictions on the other elements z_j, $j \neq i$. The *marginal density function* of z_i is the non-negative function f_{z_i} such that

$$F_{z_i}(\zeta_i) = \int_{-\infty}^{\zeta_i} f_{z_i}(s)\, \mathrm{d}s.$$

It is readily seen that the marginal density function f_{z_i} can also be obtained from the associated joint density function by integrating out the other elements:

$$f_{z_i}(s_i) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\boldsymbol{z}}(s_1, \ldots, s_d)\, \mathrm{d}s_1 \cdots \mathrm{d}s_{i-1}\, \mathrm{d}s_{i+1} \cdots \mathrm{d}s_d.$$

If there are two random vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, they are said to be *independent* if, and only if, their joint distribution function is the product of all marginal distribution functions:

$$F_{\boldsymbol{z}_1, \boldsymbol{z}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = F_{\boldsymbol{z}_1}(\boldsymbol{\zeta}_1)\, F_{\boldsymbol{z}_2}(\boldsymbol{\zeta}_2);$$

otherwise, they are *dependent*. If random vectors possess density functions, they are independent if, and only if, their joint density function is also the product of marginal density functions. Intuitively, there exists absolutely no relationship between two independent random vectors. As a consequence, functions of independent random vectors remain independent, as stated in the result below.

Lemma 2.1 *If $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are independent random vectors, then their transformations, $h_1(\boldsymbol{z}_1)$ and $h_2(\boldsymbol{z}_2)$, are also independent random variables (vectors).*

2.2 Moments

Given the d-dimensional random vector $\boldsymbol{z}$ with the distribution function $F_{\boldsymbol{z}}$, the expectation of the i^{th} element z_i is defined as

$$\mathbb{E}(z_i) = \int_{\mathbb{R}^d} \cdots \int \zeta_i\, \mathrm{d}F_{\boldsymbol{z}}(\zeta_1, \ldots, \zeta_d),$$

© Chung-Ming Kuan, 2007, 2009, 2010

where the right-hand side is a Stieltjes integral; for more details about different integrals we refer to Rudin (1976). As this integral equals

$$\int_{\mathbb{R}} \zeta_i \, \mathrm{d}F_{\boldsymbol{z}}(\infty, \ldots, \infty, \zeta_i, \infty, \ldots, \infty) = \int_{\mathbb{R}} \zeta_i \, \mathrm{d}F_{z_i}(\zeta_i),$$

the expectation of z_i can be taken with respect to either the joint distribution function $F_{\boldsymbol{z}}$ or the marginal distribution function F_{z_i}. The expectation of a random variable is a *location* measure because it is a weighted average of all possible values of this variable, with the weights being associated probabilities.

We say that the random variable z_i has a finite expected value (or the expectation $\mathbb{E}(z_i)$ exists) if $\mathbb{E}\,|z_i| < \infty$. A random variable need not have a finite expected value; if it does, this random variable is said to be *integrable*. More generally, the expectation of a random vector is defined elementwise. That is, for a random vector $\boldsymbol{z}$, $\mathbb{E}(\boldsymbol{z})$ exists if all $\mathbb{E}(z_i)$, $i = 1, \ldots, d$, exist ($\boldsymbol{z}$ is integrable if all z_i, $i = 1, \ldots, d$, are integrable).

It is easily seen that the expectation operator does not have any effect on a constant; i.e., $\mathbb{E}(b) = b$ for any constant b. For integrable random variables z_i and z_j, the expectation operator is *monotonic* in the sense that

$$\mathbb{E}(z_i) \leq \mathbb{E}(z_j),$$

for any $z_i \leq z_j$ with probability one. Moreover, the expectation operator is *linear* in the sense that

$$\mathbb{E}(az_i + bz_j) = a\,\mathbb{E}(z_i) + b\,\mathbb{E}(z_j),$$

where a and b are two real numbers. This property immediately generalizes to integrable random vectors.

Lemma 2.2 *Let $\boldsymbol{A}$ $(n \times d)$ and $\boldsymbol{B}$ $(n \times c)$ be two non-stochastic matrices. Then for any integrable random vectors $\boldsymbol{z}$ $(d \times 1)$ and $\boldsymbol{y}$ $(c \times 1)$,*

$$\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{B}\,\mathbb{E}(\boldsymbol{y});$$

in particular, if $\boldsymbol{b}$ is an n-dimensional nonstochastic vector, then $\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{b}$.

More generally, let $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{z})$ be a well-defined, vector-valued function of $\boldsymbol{z}$. The expectation of $\boldsymbol{y}$ is

$$\mathbb{E}(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{g}(\boldsymbol{z})] = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{\zeta}) \; \mathrm{d}F_{\boldsymbol{z}}(\boldsymbol{\zeta}).$$

© Chung-Ming Kuan, 2007, 2009, 2010

When $\boldsymbol{g}(\boldsymbol{z}) = z_i^k$, $\mathbb{E}[\boldsymbol{g}(\boldsymbol{z})] = \mathbb{E}(z_i^k)$ is known as the k^{th} *moment* of z_i, where k need not be an integer. In particular, $\mathbb{E}(z_i)$ is the first moment of z_i. When a random variable has finite k^{th} moment, its moments of order less than k are also finite. When the k^{th} moment of a random variable does not exist, then the moments of order greater than k also fail to exist. See Section 2.3 for some examples of random variables that possess only low order moments. A random vector is said to have finite k^{th} moment if its elements all have finite k^{th} moment. A random variable with finite second moment is said to be *square integrable*; a random vector is square integrable if its elements are all square integrable.

The k^{th} *central moment* of z_i is $\mathbb{E}[z_i - \mathbb{E}(z_i)]^k$. In particular, the second central moment of the square integrable random variable z_i is

$$\mathbb{E}[z_i - \mathbb{E}(z_i)]^2 = \mathbb{E}(z_i^2) - [\mathbb{E}(z_i)]^2,$$

which is a measure of dispersion of the values of z_i. The second central moment is also known as *variance*, denoted as $\text{var}(\cdot)$. The square root of variance is *standard deviation*. It can be verified that, given the square integrable random variable z_i and real numbers a and b,

$$\text{var}(az_i + b) = \text{var}(az_i) = a^2 \text{ var}(z_i).$$

This shows that variance is location invariant but depends on the scale (measurement units) of random variables.

When $g(\boldsymbol{z}) = z_i z_j$, $\mathbb{E}[g(\boldsymbol{z})] = \mathbb{E}(z_i z_j)$ is the *cross moment* of z_i and z_j. The *cross central moment* of z_i and z_j is

$$\mathbb{E}[(z_i - \mathbb{E}(z_i))(z_j - \mathbb{E}(z_j))] = \mathbb{E}(z_i z_j) - \mathbb{E}(z_i)\,\mathbb{E}(z_j),$$

which is a measure of the co-variation between these two random variables. The cross central moment of two random variables is known as their *covariance*, denoted as $\text{cov}(\cdot,\cdot)$. Clearly, $\text{cov}(z_i, z_j) = \text{cov}(z_j, z_i)$ and $\text{cov}(z_i, z_i) = \text{var}(z_i)$. It can be seen that for real numbers a, b, c, d,

$$\text{cov}(az_i + b, cz_j + d) = \text{cov}(az_i, cz_j) = ac\,\text{cov}(z_i, z_j).$$

Thus, covariance is also location invariant but not scale invariant.

Observe that for any real numbers a and b,

$$\text{var}(az_i + bz_j) = a^2 \text{ var}(z_i) + b^2 \text{ var}(z_j) + 2ab \text{ cov}(z_i, z_j),$$

so that

$$\text{var}(z_i - az_j) = \text{var}(z_i) + a^2\,\text{var}(z_j) - 2a\,\text{cov}(z_i, z_j),$$

© Chung-Ming Kuan, 2007, 2009, 2010

which must be non-negative. Setting $a = \operatorname{cov}(z_i, z_j)/\operatorname{var}(z_j)$, we have

$$\operatorname{var}(z_i) - \operatorname{cov}(z_i, z_j)^2/\operatorname{var}(z_j) \geq 0.$$

In particular, when $z_i = az_j + b$ for some real numbers a and b, we have $\operatorname{var}(z_i) = a^2 \operatorname{var}(z_j)$ and $\operatorname{cov}(z_i, z_j) = a \operatorname{var}(z_j)$, so that

$$\operatorname{var}(z_i) - \operatorname{cov}(z_i, z_j)^2/\operatorname{var}(z_j) = 0.$$

This yields the *Cauchy-Schwarz inequality* for square integrable random variables.

Lemma 2.3 (Cauchy-Schwarz) *Let z_i, z_j be two square integrable random variables. Then,*

$$\operatorname{cov}(z_i, z_j)^2 \leq \operatorname{var}(z_i) \operatorname{var}(z_j),$$

where the equality holds when $z_i = az_j + b$ for some real numbers a and b.

cf. the Cauchy-Schwarz inequality (Lemma 1.1) in Section 1.2. This also suggests that when two random variables are square integrable, their covariance must be finite.

The *correlation coefficient* of z_i and z_j is defined as

$$\operatorname{corr}(z_i, z_j) = \frac{\operatorname{cov}(z_i, z_j)}{\sqrt{\operatorname{var}(z_i) \operatorname{var}(z_j)}}.$$

Clearly, a correlation coefficient provides the same information as its corresponding covariance. Moreover, we have from Lemma 2.3 that

$$-1 \leq \operatorname{corr}(z_i, z_j) \leq 1.$$

For two random variables z_i and z_j and real numbers a, b, c, d,

$$\operatorname{corr}(az_i + b, cz_j + d) = \operatorname{corr}(az_i, cz_j) = \frac{ac}{|a|\,|c|} \operatorname{corr}(z_i, z_j).$$

Thus, the correlation coefficient is not only location invariant but also scale invariant, apart from the sign change. If $\operatorname{corr}(z_i, z_j) = 0$, z_i and z_j are said to be *uncorrelated*. If $\operatorname{corr}(z_i, z_j) > 0$, z_i and z_j are said to be *positively correlated*; if $\operatorname{corr}(z_i, z_j) < 0$, z_i and z_j are *negatively correlated*. When $z_i = az_j + b$, $\operatorname{corr}(z_i, z_j) = 1$ if $a > 0$ and -1 if $a < 0$. In both cases, z_i and z_j are em perfectly correlated.

For a d-dimensional, square integrable random vector $\boldsymbol{z}$, its variance-covariance matrix is

$$\begin{aligned}\operatorname{var}(\boldsymbol{z}) &= \mathbb{E}[(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))'] \\ &= \begin{bmatrix} \operatorname{var}(z_1) & \operatorname{cov}(z_1, z_2) & \cdots & \operatorname{cov}(z_1, z_d) \\ \operatorname{cov}(z_2, z_1) & \operatorname{var}(z_2) & \cdots & \operatorname{cov}(z_2, z_d) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(z_d, z_1) & \operatorname{cov}(z_d, z_2) & \cdots & \operatorname{var}(z_d) \end{bmatrix},\end{aligned}$$

© Chung-Ming Kuan, 2007, 2009, 2010

which must be symmetric because $\mathrm{cov}(z_i, z_j) = \mathrm{cov}(z_j, z_i)$. As $(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))'$ is positive semi-definite with probability one, the monotonicity of the expectation operator implies that $\mathrm{var}(\boldsymbol{z})$ is positive semi-definite.

For two random vectors $\boldsymbol{y}$ $(c \times 1)$ and $\boldsymbol{z}$ $(d \times 1)$, the $d \times c$ covariance matrix of $\boldsymbol{z}$ and $\boldsymbol{y}$ is

$$\mathrm{cov}(\boldsymbol{z}, \boldsymbol{y}) = \mathbb{E}[(\boldsymbol{z} - \mathbb{E}\,\boldsymbol{z})(\boldsymbol{y} - \mathbb{E}\,\boldsymbol{y})'] = \mathbb{E}(\boldsymbol{z}\boldsymbol{y}') - \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{y}').$$

Two random vectors are uncorrelated if their covariance matrix is a zero matrix. If $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent, their joint distribution function is the product of individual distribution functions. It follows that the cross moment of $\boldsymbol{y}$ and $\boldsymbol{z}$ is the product of their individual first moment: that

$$\mathbb{E}(\boldsymbol{z}\boldsymbol{y}') = \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{y}').$$

This shows that independence implies $\mathrm{cov}(\boldsymbol{z}, \boldsymbol{y}) = \boldsymbol{0}$. Uncorrelated random vectors are not necessarily independent, however.

Based on the properties of variance and covariance for random variables, we have the following result for random vectors.

Lemma 2.4 *Let $\boldsymbol{A}$ $(n \times d)$, $\boldsymbol{B}$ $(n \times c)$, and $\boldsymbol{C}$ $(m \times c)$ be non-stochastic matrices and $\boldsymbol{b}$ an n-dimensional non-stochastic vector. Then for any square integrable random vectors $\boldsymbol{z}$ $(d \times 1)$ and $\boldsymbol{y}$ $(c \times 1)$,*

$$\mathrm{var}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\,\mathrm{var}(\boldsymbol{z})\boldsymbol{A}' + \boldsymbol{B}\,\mathrm{var}(\boldsymbol{y})\boldsymbol{B}' + 2\boldsymbol{A}\,\mathrm{cov}(\boldsymbol{z}, \boldsymbol{y})\boldsymbol{B}',$$
$$\mathrm{var}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}) = \mathrm{var}(\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}\ \mathrm{var}(\boldsymbol{z})\ \boldsymbol{A}'.$$

Given two square integrable random vectors $\boldsymbol{z}$ and $\boldsymbol{y}$, suppose that $\mathrm{var}(\boldsymbol{y})$ is positive definite. As the variance-covariance matrix of $(\boldsymbol{z}'\ \boldsymbol{y}')'$ must be a positive semi-definite matrix,

$$[I\ \ -\mathrm{cov}(\boldsymbol{z}, \boldsymbol{y})\,\mathrm{var}(\boldsymbol{y})^{-1}] \begin{bmatrix} \mathrm{var}(\boldsymbol{z}) & \mathrm{cov}(\boldsymbol{z}, \boldsymbol{y}) \\ \mathrm{cov}(\boldsymbol{y}, \boldsymbol{z}) & \mathrm{var}(\boldsymbol{y}) \end{bmatrix} \begin{bmatrix} I \\ -\mathrm{var}(\boldsymbol{y})^{-1}\,\mathrm{cov}(\boldsymbol{y}, \boldsymbol{z}) \end{bmatrix}$$
$$= \mathrm{var}(\boldsymbol{z}) - \mathrm{cov}(\boldsymbol{z}, \boldsymbol{y})\,\mathrm{var}(\boldsymbol{y})^{-1}\,\mathrm{cov}(\boldsymbol{y}, \boldsymbol{z})$$

is also a positive semi-definite matrix. This establishes the multivariate version of the Cauchy-Schwarz inequality for square integrable random vectors.

© Chung-Ming Kuan, 2007, 2009, 2010

Lemma 2.5 (Cauchy-Schwarz) *Let $\boldsymbol{y}, \boldsymbol{z}$ be two square integrable random vectors. Then,*

$$\operatorname{var}(\boldsymbol{z}) - \operatorname{cov}(\boldsymbol{z}, \boldsymbol{y}) \operatorname{var}(\boldsymbol{y})^{-1} \operatorname{cov}(\boldsymbol{y}, \boldsymbol{z})$$

is a positive semi-definite matrix.

A random vector is said to be *degenerate* (have a singular distribution) if its variance-covariance matrix is singular. Let $\boldsymbol{\Sigma}$ be the variance-covariance matrix of the d-dimensional random vector $\boldsymbol{z}$. If $\boldsymbol{\Sigma}$ is singular, then there exists a non-zero vector $\boldsymbol{c}$ such that $\boldsymbol{\Sigma c} = \boldsymbol{0}$. For this particular $\boldsymbol{c}$, we have

$$\boldsymbol{c}'\boldsymbol{\Sigma c} = \mathbb{E}[\boldsymbol{c}'(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))]^2 = 0.$$

It follows that $\boldsymbol{c}'[\boldsymbol{z} - \mathbb{E}(\boldsymbol{z})] = 0$ with probability one; i.e, the elements of $\boldsymbol{z}$ are linearly dependent with probability one. This implies that all the probability mass of $\boldsymbol{z}$ is concentrated in a subspace of dimension less than d.

2.3 Special Distributions

In this section we discuss several useful distributions. A random vector $\boldsymbol{z}$ is said to have a *multivariate normal* (*Gaussian*) *distribution* with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, denoted as $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if it has the density function

$$\frac{1}{(2\pi)^{d/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})\right).$$

For $d = 1$, this is just the density of the univariate normal random variable. Note that the multivariate normal density function is completely characterized by its mean vector and variance-covariance matrix. A normal random variable has moments of all orders; in particular, its even-order central moments are

$$\mathbb{E}(z - \mu)^k = (k-1) \cdots 3 \cdot 1 \operatorname{var}(z)^{k/2}, \qquad k \geq 2 \text{ and } k \text{ is even},$$

and its odd-order central moments are all zeros. A normal random variable with mean zero and variance one is usually called the *standard normal* random variable.

When $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements σ_{ii}, $i = 1, \ldots, d$, the elements of $\boldsymbol{z}$ are uncorrelated. In this case, the density function of $\boldsymbol{z}$ is simply the product of the marginal density functions of $z_1, \ldots, z_d$:

$$\frac{1}{(2\pi)^{d/2}(\prod_{i=1}^d \sigma_{ii})^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{(z_i - \mu_i)^2}{\sigma_{ii}}\right).$$

© Chung-Ming Kuan, 2007, 2009, 2010

That is, for random variables that are jointly normally distributed, uncorrelatedness implies independence. When $\sigma_{ii} = \sigma_o^2$, a constant, the joint density above further simplifies to

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

Note that uncorrelated random variables are not necessarily independent if they are not jointly normally distributed.

The result below shows that proper linear transformations of normal random vectors remain normally distributed.

Lemma 2.6 *Let $\boldsymbol{z}$ be a d-dimensional random vector distributed as $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also let $\boldsymbol{A}$ be an $n \times d$ non-stochastic matrix with full row rank $n < d$ and $\boldsymbol{b}$ be a d-dimensional non-stochastic vector. Then,*

$$\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}').$$

Lemma 2.6 implies that, when $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, any sub-vector (element) of $\boldsymbol{z}$ also has a multivariate (univariate) normal distribution; the converse need not be true, however. It is also easily seen that

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{0},\ \boldsymbol{I}_d),$$

where $\boldsymbol{\Sigma}^{-1/2}$ is such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{I}$, as defined in Section 1.7. Proper standardization of a normal random vector thus yields a normal random vector with independent elements. If $\boldsymbol{A}$ is not of full row rank, $\mathrm{var}(\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ does not have full rank, so that $\boldsymbol{A}\boldsymbol{z}$ is degenerate.

Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I}_d)$. The sum of squares of the elements of $\boldsymbol{z}$ is the *non-central chi-square* random variable with d degrees of freedom and the *non-centrality parameter* $\nu = \boldsymbol{\mu}'\boldsymbol{\mu}$, denoted as

$$\boldsymbol{z}'\boldsymbol{z} \sim \chi^2(d; \nu).$$

The density function of $\chi^2(d; \nu)$ is

$$f(x) = \exp\Big(-\frac{\nu + x}{2}\Big)x^{d/2-1}\frac{1}{2^{d/2}}\sum_{i=0}^{\infty}\frac{x^i\nu^i}{i!\,2^{2i}\,\Gamma(i + d/2)}, \qquad x > 0,$$

where Γ is the gamma function with

$$\Gamma(n) = \int_0^{\infty} e^{-x}x^{n-1}\ \mathrm{d}x.$$

© Chung-Ming Kuan, 2007, 2009, 2010

It can be shown that a $\chi^2(d;\nu)$ random variable has mean $(d+\nu)$ and variance $2d+4\nu$. When $\boldsymbol{\mu}=\mathbf{0}$, the non-centrality parameter $\nu=0$, and $\chi^2(d;0)$ is known as the central chi-square random variable, denoted as $\chi^2(d)$. The density of $\chi^2(d)$ is

$$f(x)=\exp\Big(-\frac{x}{2}\Big)x^{d/2-1}\frac{1}{2^{d/2}\,\Gamma(d/2)},\qquad x>0,$$

with mean d and variance $2d$. The result below follows directly from Lemma 2.6.

Lemma 2.7 *Let $\boldsymbol{z}$ be a d-dimensional random vector distributed as $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Then,*

$$\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\sim\chi^2(d;\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu});$$

in particular, if $\boldsymbol{\mu}=\mathbf{0}$, $\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\sim\chi^2(d)$.

Let w and x be two independent random variables such that $w\sim\mathcal{N}(\mu,1)$ and $x\sim\chi^2(n)$. Then

$$\frac{w}{\sqrt{x/n}}\sim t(n;\mu),$$

the non-central t distribution with n degrees of freedom and the non-centrality parameter μ. The density function of $t(n;\mu)$ is

$$f(x)=\frac{n^{n/2}\exp(-\mu^2/2)}{\Gamma(n/2)\Gamma(1/2)(n+x^2)^{(n+1)/2}}\sum_{i=0}^{\infty}\Gamma\Big(\frac{n+i+1}{2}\Big)\frac{\mu^i}{i!}\left(\frac{2x^2}{n+x^2}\right)^{i/2}(\text{sign }x)^i.$$

When $\mu=0$, $t(n;\mu)$ reduces to the central t distribution, denoted as $t(n)$, which has the density

$$f(x)=\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2)n^{1/2}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}.$$

The $t(n)$ random variable is symmetric about zero, and its k^{th} moment exists only for $k<n$; when $n>2$, its mean is zero and variance is $n/(n-2)$.

As n tends to infinity, it can be seen that

$$\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}=\left[\Big(1+\frac{x^2}{n}\Big)^{n/x^2}\right]^{-x^2/2}\Big(1+\frac{x^2}{n}\Big)^{-1/2}\to\exp(-x^2/2).$$

Also note that $\Gamma(1/2)=\pi^{1/2}$ and that for large n,

$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\approx(n/2)^{1/2}.$$

Thus, when n tends to infinity, the density of $t(n)$ converges to

$$\frac{1}{\sqrt{2\pi}}\exp(-x^2/2),$$

© Chung-Ming Kuan, 2007, 2009, 2010

the density of the standard normal random variable. When $n = 1$, the density for $t(1)$ becomes

$$f(x) = \frac{1}{\pi[1 + x^2]}.$$

This is also the density of the *Cauchy* random variable with the location parameter 0. The Cauchy random variable is a very special random variable because it does not even have finite first moment.

Let z_1 and z_2 be two independent random variables such that $z_1 \sim \chi^2(n_1; \nu_1)$ and $z_2 \sim \chi^2(n_2; \nu_2)$. Then,

$$\frac{z_1/n_1}{z_2/n_2} \sim F(n_1, n_2; \nu_1, \nu_2),$$

the non-central F distribution with the degrees of freedom n_1 and n_2 and the non-centrality parameters ν_1 and ν_2. The k^{th} moment of $F(n_1, n_2; \nu_1, \nu_2)$ exists when $k < n_2/2$. In many statistical applications we usually encounter $F(n_1, n_2; \nu_1, 0)$. When $n_2 > 2$, the mean of $F(n_1, n_2; \nu_1, 0)$ is

$$\frac{n_2(n_1 + \nu_1)}{n_1(n_2 - 2)};$$

when $n_2 > 4$, the variance is

$$2\Big(\frac{n_2}{n_1}\Big)^2 \frac{(n_1 + \nu_1)^2 + (n_1 + 2\nu_1)(n_2 - 2)}{(n_2 - 2)^2(n_2 - 4)}.$$

If both ν_1 and ν_2 are zero, we have the central F distribution $F(n_1, n_2)$. When $n_2 > 2$, $F(n_1, n_2)$ has mean $n_2/(n_2 - 2)$; when $n_2 > 4$, it has variance

$$\frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)}.$$

Note that if a random variable is distributed as $t(n)$, its square has the $F(1, n)$ distribution.

2.4 Likelihood

Suppose that we postulate p as the joint probability function of the discrete random variables $z_1, \ldots, z_T$ with the parameter vector $\boldsymbol{\theta}$. Plugging the observed values $\zeta_1, \ldots, \zeta_T$ of these random variables into p we obtain a function of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) := p(\zeta_1, \ldots, \zeta_T; \boldsymbol{\theta}).$$

This function represents the probability (likelihood) that those observed values are generated from the postulated probability function p; different parameter values of course result in different probability values. Thus, $L(\boldsymbol{\theta})$ is also known as the *likelihood function* of $\boldsymbol{\theta}$.

© Chung-Ming Kuan, 2007, 2009, 2010

Similarly, let f denote the postulated joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ with the parameter vector $\boldsymbol{\theta}$. Then given the observed values $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T$, the likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) := f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}).$$

In what follows, we will use L and f interchangeably. Note, however, that a postulated density function need not be the true density function that generates the random variables.

When f is differentiable and non-zero with probability one, the gradient vector of $\log L(\boldsymbol{\theta})$,

$$\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \frac{1}{L(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}),$$

is known as the *score* vector, denoted as $\boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})$. We can then write

$$\boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}) f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}).$$

For a given $\boldsymbol{\theta}$, the score vector varies with the observed values $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T$, so that it is also a random vector. We therefore denote the score vector as $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta})$.

When differentiation and integration can be interchanged, we have for each $\boldsymbol{\theta}$,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \ \ldots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \ \ldots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \nabla_{\boldsymbol{\theta}} \left(\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \ \ldots\ \mathrm{d}\boldsymbol{\zeta}_T \right) \\
&\quad = \nabla_{\boldsymbol{\theta}}\, 1 \\
&\quad = \mathbf{0}.
\end{aligned}$$

The left-hand side is in effect the expectation of the score vector with respect to f. If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the true density function, we immediately obtain the following result.

Lemma 2.8 *If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. Then under regularity conditions,*

$$\mathbb{E}[\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)] = \mathbf{0},$$

where $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)$ is the score evaluated at $\boldsymbol{\theta}_o$, and $\mathbb{E}$ is taken with respect to the true density function.

© Chung-Ming Kuan, 2007, 2009, 2010

Remark: The validity of Lemma 2.8 requires differentiability of the likelihood function and interchangeability of differentiation and integration; see, e.g., Amemiya (1985) for some sufficient conditions of these properties. Lemma 2.9 below also requires similar conditions.

It is easy to see that the Hessian matrix of the log-likelihood function is

$$\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \frac{1}{L(\boldsymbol{\theta})}\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \frac{1}{L(\boldsymbol{\theta})^2}[\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})][\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})]',$$

where the second term is just the outer product of the score vector. Again by interchanging differentiation and integration, we have for each $\boldsymbol{\theta}$ that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \frac{1}{L(\boldsymbol{\theta})}\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \cdots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \cdots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \nabla^2_{\boldsymbol{\theta}} \left(\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \cdots\ \mathrm{d}\boldsymbol{\zeta}_T \right) \\
&\quad = \nabla^2_{\boldsymbol{\theta}}\, 1 \\
&\quad = \mathbf{0}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&-\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \cdots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\qquad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})' f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots\ \mathrm{d}\boldsymbol{\zeta}_T,
\end{aligned}$$

where the left-hand side is the negative of the expected Hessian matrix and the right-hand side is the variance-covariance matrix of $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta})$, both with respect to the postulated density function f. If $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the true density function, the variance-covariance matrix of $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)$ is known as the *information matrix*. The result above, together with Lemma 2.8, yields the so-called *information matrix equality*.

Lemma 2.9 *If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. Then under regularity conditions,*

$$\mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}_o)] + \mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)) = \mathbf{0},$$

where $\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}_o)$ is the Hessian matrix of $\log L$ evaluated at $\boldsymbol{\theta}_o$, and $\mathbb{E}$ and var *are taken with respect to the true density function.*

Remark: When f is not the true density function, Lemma 2.8 and 2.9 need not hold. That is, neither $\mathbb{E}[\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta})]$ nor

$$\mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta})] + \mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}))$$

is necessarily zero.

© Chung-Ming Kuan, 2007, 2009, 2010

2.5 Estimation

2.5.1 Point Estimation

Let $\boldsymbol{\theta}_o$ denote a parameter vector associated with the joint distribution of T random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. A *point estimator* (or simply an estimator) for $\boldsymbol{\theta}_o$ is a function of these random vectors:

$$\hat{\boldsymbol{\theta}} = h(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T),$$

where h is some function. An estimator is clearly a random vector. Once the observed values of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ are plugged into this function, we obtain a *point estimate*. That is, a point estimate is just a particular value that an estimator may assume.

A simple principle of constructing estimators for moments is known as *analog estimation*. This principle suggests to estimate population moments using their finite-sample counterparts. For example, given a sample of T random variables $z_1, \ldots, z_T$ with the common k^{th} moment $\mathbb{E}(z_1^k)$, the analog estimator for $\mathbb{E}(z_1^k)$ is simply the sample average of z_i^k:

$$\frac{1}{T}\sum_{t=1}^{T} z_i^k.$$

In particular, the sample mean $\bar{z}$ is the analog estimator for the population mean.

To estimate the parameter vector $\boldsymbol{\theta}_o$, it is also natural to maximize the associated likelihood function $L(\boldsymbol{\theta})$. The resulting solution is known as the *maximum likelihood estimator* (MLE) for $\boldsymbol{\theta}_o$, denoted as $\tilde{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}_T$, where the subscript T indicates that this is an estimator based on a sample of T observations. As the maximum of a function is invariant with respect to monotonic transformations, it is quite common to compute the MLE by maximizing the log-likelihood function $\log L(\boldsymbol{\theta})$. It follows that the score vector evaluated at $\tilde{\boldsymbol{\theta}}$ must be zero; i.e., $\boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \tilde{\boldsymbol{\theta}}) = \mathbf{0}$.

2.5.2 Criteria for Point Estimators

Let $\hat{\boldsymbol{\theta}}$ be an estimator for $\boldsymbol{\theta}_o$. The difference $\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o$ is called the *bias* of $\hat{\boldsymbol{\theta}}$. An estimator is said to be *unbiased* if it has zero bias, i.e.,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_o;$$

otherwise, it is *biased*. Unbiasedness does not imply that an estimate must be close to the true parameter. In fact, it is even possible that all the values of an unbiased estimator deviate from the true parameter by a constant.

© Chung-Ming Kuan, 2007, 2009, 2010

Given two unbiased estimators, it is therefore natural to choose the one whose values are more concentrated around the true parameter. For real-valued unbiased estimators, this amounts to selecting an estimator with a smaller variance. If they are vector-valued, we adopt the following *efficiency* criterion. An unbiased estimator $\hat{\boldsymbol{\theta}}_1$ is said to be "better" (more efficient) than an unbiased estimator $\hat{\boldsymbol{\theta}}_2$ if

$$\mathrm{var}(\boldsymbol{a}'\hat{\boldsymbol{\theta}}_2) \geq \mathrm{var}(\boldsymbol{a}'\hat{\boldsymbol{\theta}}_1),$$

for all non-zero vectors $\boldsymbol{a}$. This is equivalent to the condition that

$$\boldsymbol{a}'[\mathrm{var}(\hat{\boldsymbol{\theta}}_2) - \mathrm{var}(\hat{\boldsymbol{\theta}}_1)]\boldsymbol{a} \geq 0,$$

for all non-zero vectors $\boldsymbol{a}$. Thus, an unbiased estimator $\hat{\boldsymbol{\theta}}_1$ is more efficient than an unbiased estimator $\hat{\boldsymbol{\theta}}_2$ if $\mathrm{var}(\hat{\boldsymbol{\theta}}_2) - \mathrm{var}(\hat{\boldsymbol{\theta}}_1)$ is a positive semi-definite matrix. Given a class of unbiased estimators, if one of them is better than all other estimators in that class, it is the "best" (most efficient) within this class.

More generally, we can compare estimators based on mean squared error (MSE):

$$\begin{aligned}
&\mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'] \\
&\quad = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o)(\hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o)'] \\
&\quad = \mathrm{var}(\hat{\boldsymbol{\theta}}) + [\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o]\,[\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o]',
\end{aligned}$$

where the second term is the outer product of the bias vector. An estimator $\hat{\boldsymbol{\theta}}_1$ (not necessarily unbiased) is said to be better (more efficient) than $\hat{\boldsymbol{\theta}}_2$ if $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_2) - \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1)$ is a positive semi-definite matrix. Clearly, the MSE criterion reduces to the previous variance-based criterion when estimators are unbiased.

The following result shows that the inverse of the information matrix is a lower bound, also known as the *Cramér-Rao lower bound*, for the variance-covariance matrix of any unbiased estimator.

Lemma 2.10 (Cramér-Rao) *If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. Let $\hat{\boldsymbol{\theta}}$ denote an unbiased estimator for $\boldsymbol{\theta}$ based on these random vectors. If $\mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o))$ is positive definite,*

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) - \mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o))^{-1}$$

is a positive semi-definite matrix.

© Chung-Ming Kuan, 2007, 2009, 2010

Proof: We first note that for any unbiased estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\, \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \hat{\boldsymbol{\theta}}\, \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \nabla_{\boldsymbol{\theta}} \left(\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \hat{\boldsymbol{\theta}}\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \right) \\
&\quad = \nabla_{\boldsymbol{\theta}} \boldsymbol{\theta} \\
&\quad = \boldsymbol{I},
\end{aligned}$$

where the third equality holds because $\hat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$ when $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})$ is the associated density function. Thus,

$$\operatorname{cov}(\hat{\boldsymbol{\theta}}, \boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)) = \boldsymbol{I}.$$

The assertion now follows from Lemma 2.5, the multivariate version of the Cauchy-Schwarz inequality. □

By Lemma 2.10, an unbiased estimator is the best if its variance-covariance matrix achieves the Cramér-Rao lower bound; the converse need not be true, however.

2.5.3 Interval Estimation

While a point estimate is a particular value representing the unknown parameter, *interval estimation* results in a range of values that may contain the unknown parameter with certain probability.

Suppose that there is an estimate $\hat{\theta}$ for the true parameter θ_o and a function $q(\hat{\theta}, \theta_o)$ whose distribution is known. Then, given a probability value γ, we can find suitable values a and b such that

$$\mathbb{P}\{a < q(\hat{\theta}, \theta_o) < b\} = \gamma.$$

Solving the inequality above for θ_o we may obtain an interval containing θ_o. This leads to the probability statement:

$$\mathbb{P}\{\alpha < \theta_o < \beta\} = \gamma,$$

where α and β depend on a, b, and $\hat{\theta}$. We can then conclude that we are $\gamma \times 100$ percent sure that the interval (α, β) contains θ_o. Here, γ is the *confidence coefficient*, and (α, β) is the associated *confidence interval* for θ_o. It is easily seen that the larger the value of γ,

© Chung-Ming Kuan, 2007, 2009, 2010

the wider is the associated confidence interval. Note that for a given confidence coefficient, there may exist different confidence intervals satisfying the same probability statement. It is then desirable to find the smallest possible confidence interval.

Let A_1 denote the event that a confidence interval contains θ_1 and A_2 the event that a confidence interval contains θ_2. The intersection $A = A_1 \cap A_2$ is thus the event that a confidence "box" covers both parameters. When A_1 and A_2 are independent such that $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \gamma$, we have $\mathbb{P}(A) = \gamma^2$. When these two events are not independent (e.g., the parameter estimators of θ_1 and θ_2 are correlated), it becomes difficult to determine $\mathbb{P}(A)$. As such, finding a proper confidence "box" based on individual confidence intervals is by no means an easy job. On the other hand, if a function $q(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2)$ with a known distribution is available, we can, for a given γ, find the values a and b such that

$$\mathbb{P}\{a < q(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2) < b\} = \gamma.$$

By solving the inequality above for θ_1 and θ_2 we may obtain a *confidence region* in which the point (θ_1, θ_2) lies with probability γ.

2.6 Hypothesis Testing

2.6.1 Basic Concepts

Given a sample of data, it is often desirable to check if certain characteristics of the underlying random mechanism (population) are supported by these data. For this purpose, a *hypothesis* of these characteristics must be specified, and a *test* is constructed so as to generate a rule of rejecting or accepting (not rejecting) the postulated hypothesis.

The hypothesis being tested is called the *null hypothesis*, denoted as H_0; the other states or values of the characteristics of interest form an *alternative hypothesis*, denoted as H_1. Hypotheses are usually formulated in terms of the parameters of models. For example, one may specify that $H_0 : \boldsymbol{\theta}_o = \boldsymbol{a}$ for some $\boldsymbol{a}$ and $H_1 : \boldsymbol{\theta}_o \neq \boldsymbol{a}$. Here, H_0 is a *simple hypothesis* in the sense that the parameter vector being tested takes a single value, but H_1 is a *composite hypothesis* in that the parameter vector may take more than one values. Given a sample of random variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$, a *test statistic* is a function of these random variables, denoted as $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$. The *critical region* C of $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ is the range of its possible values that lead to rejection of the null hypothesis. In what follows, the set

$$\Gamma = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T : \mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in C\}$$

will also be referred to as the critical region of $\mathcal{T}$. The complement of the critical region, C^c, is the region containing the values of $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ that lead to acceptance of the null

© Chung-Ming Kuan, 2007, 2009, 2010

hypothesis. We can also define

$$\Gamma^c = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T \colon \mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in C^c\}$$

as the acceptance region of $\mathcal{T}$.

A test may yield incorrect inferences. A test is said to commit the *type I error* if it rejects the null hypothesis when the null hypothesis is in fact true; a test is said to commit the *type II error* if it accepts the null hypothesis when the alternative hypothesis is true. Suppose that we are interested in testing $H_0 \colon \boldsymbol{\theta}_o = \boldsymbol{a}$ against $H_1 \colon \boldsymbol{\theta}_o = \boldsymbol{b}$. Let $\mathbb{P}_0$ be the probability asociated with $\boldsymbol{\theta}_o = \boldsymbol{a}$ and $\mathbb{P}_1$ the probability with $\boldsymbol{\theta}_o = \boldsymbol{b}$. The probability of the type I error is then

$$\alpha = \mathbb{P}_0\{(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \Gamma\} = \int_\Gamma f_0(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{a}) \, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T,$$

where $f_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{a})$ is the joint density with the parameter $\boldsymbol{\theta}_o = \boldsymbol{a}$. The value α is also known as the *size* or *significance level* of the test. The probability of the type II error is

$$\beta = \mathbb{P}_1\{(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \Gamma^c\} = \int_{\Gamma^c} f_1(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{b}) \, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T,$$

where $f_1(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{b})$ is the joint density with the parameter $\boldsymbol{\theta}_o = \boldsymbol{b}$. Clearly, α decreases when the critical region Γ is smaller; in the mean time, β increases due to a larger Γ^c. Thus, there is usually a trade-off between these two error probabilities.

Note, however, that the probability of the type II error cannot be defined as above when the alternative hypothesis is composite: $\boldsymbol{\theta}_o \in \Theta_1$, where Θ_1 is a set of parameter values in the parameter space. Consider now the probability $1 - \mathbb{P}_1(\Gamma^c) = \mathbb{P}_1(\Gamma)$, which is the probability of rejecting the null hypothesis when H_1 is true. Thus, both $\mathbb{P}_0(\Gamma)$ and $\mathbb{P}_1(\Gamma)$ are the probabilities of rejecting the null hypothesis under two different parameter values. More generally, define the *power function* of the test as

$$\pi(\boldsymbol{\theta}_o) = \mathbb{P}_{\boldsymbol{\theta}_o}\{(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \Gamma\},$$

where $\boldsymbol{\theta}_o$ varies in the parameter space. In particular, $\pi(\boldsymbol{a}) = \alpha$. For $\boldsymbol{\theta}_o \in \Theta_1$, $\pi(\boldsymbol{\theta}_o)$ describes the ability of a test that can correctly detect the falsity of the null hypothesis; these probabilities are also referred to as the *powers* of the test. The probability of the type II error under the composite alternative hypothesis $\boldsymbol{\theta}_o \in \Theta_1$ can now be defined as

$$\beta = \max_{\boldsymbol{\theta}_o \in \Theta_1} [1 - \pi(\boldsymbol{\theta}_o)].$$

© Chung-Ming Kuan, 2007, 2009, 2010

2.6.2 Construction of Tests

Given the null hypothesis $\boldsymbol{\theta}_o = \boldsymbol{a}$, the test statistic $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ is usually based on the comparison of an estimator of $\boldsymbol{\theta}_o$ and the hypothesized value $\boldsymbol{a}$. This statistic must have a known distribution under the null hypothesis, which will be referred to as the *null distribution*.

Given the statistic $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$, the probability $\mathbb{P}_0\{\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C\}$ can be determined by the null distribution of $\mathcal{T}$. If this probability is small, the event that $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C$ would be considered "unlikely" or "improbable" under the null hypothesis, while the event that $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C^c$ would be considered "likely" or "probable." When the former, unlikely event occurs (i.e., for data $\boldsymbol{z}_1 = \boldsymbol{\zeta}_1, \ldots, \boldsymbol{z}_T = \boldsymbol{\zeta}_T$, $\mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T)$ falls in C), it constitutes an evidence against the null hypothesis, so that the null hypothesis is rejected; otherwise, we accept (do not reject) the null hypothesis. We have seen that there is a trade-off between the error probabilities α and β. To construct a test, we may fix one of these two error probabilities at a small level. It is typical to specify a small significance level α and determine the associated critical region C by

$$\alpha = \mathbb{P}_0\{\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C\}.$$

As such, we shall write the critical region for the significance level α as C_α. This approach ensures that, even though the decision of rejection might be wrong, the probability of making the type I error is no greater than α. A test statistic is said to be *significant* if it is in the critical region; otherwise, it is *insignificant*.

Another approach is to reject the null hypothesis if

$$\mathbb{P}_0\{v \colon v > \mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T)\}$$

is small. This probability is the tail probability of the null distribution and also known as the p-value of the statistic $\mathcal{T}$. Although this approach does not require specifying the critical region, it is virtually the same as the previous approach.

The rationale of our test decision is that the null hypothesis is rejected because the test statistic takes an unlikely value. It is then natural to expect that the calculated statistic is relatively more likely under the alternative hypothesis. Given the null hypothesis $\boldsymbol{\theta}_o = \boldsymbol{a}$ and alternative hypothesis $\boldsymbol{\theta}_o \in \Theta_1$, we would like to have a test such that

$$\pi(\boldsymbol{a}) \leq \pi(\boldsymbol{\theta}_o), \qquad \boldsymbol{\theta}_o \in \Theta_1.$$

A test is said to be *unbiased* if its size is no greater than the powers under the alternative hypothesis. Moreover, we would like to have a test that can detect the falsity of the null

© Chung-Ming Kuan, 2007, 2009, 2010

hypothesis with probability approaching one when there is sufficient information. That is, for every $\boldsymbol{\theta}_o \in \Theta_1$,

$$\pi(\boldsymbol{\theta}_o) = \mathbb{P}_{\boldsymbol{\theta}_o}\{\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C\} \to 1,$$

as $T \to \infty$. A test is said to be *consistent* if its power approaches one when the sample size becomes infinitely large.

Example 2.11 Given the sample of i.i.d. normal random variables $z_1, \ldots, z_T$ with mean μ_o and variance one. We would like to test the null hypothesis $\mu_o = 0$. A natural estimator for μ_o is the sample average $\bar{z} = T^{-1}\sum_{t=1}^{T} z_t$. It is well known that

$$\sqrt{T}(\bar{z} - \mu_o) \sim \mathcal{N}(0, 1).$$

Hence, $\sqrt{T}\bar{z} \sim \mathcal{N}(0, 1)$ under the null hypothesis; that is, the null distribution of the statistic $\sqrt{T}\bar{z}$ is the standard normal distribution. Given the significance level α, we can determine the critical region C_α using

$$\alpha = \mathbb{P}_0(\sqrt{T}\bar{z} \in C_\alpha).$$

Let Φ denote the distribution function of the standard normal random variable. For $\alpha = 0.05$, we know

$$0.05 = \mathbb{P}_0(\sqrt{T}\bar{z} > 1.645) = 1 - \Phi(1.645).$$

The critical region is then $(1.645, \infty)$; the null hypothesis is rejected if the calculated statistic falls in this interval. When the null hypothesis is false, the distribution of $\sqrt{T}\bar{z}$ is no longer $\mathcal{N}(0, 1)$ but $\mathcal{N}(\mu_o, 1)$ for some non-zero μ_o. Suppose that $\mu_o > 0$. Then,

$$\mathbb{P}_1(\sqrt{T}\bar{z} > 1.645) = \mathbb{P}_1(\sqrt{T}(\bar{z} - \mu_o) > 1.645 - \sqrt{T}\mu_o).$$

Since $\sqrt{T}(\bar{z} - \mu_o) \sim \mathcal{N}(0, 1)$ under the alternative hypothesis, we have the power:

$$\mathbb{P}_1(\sqrt{T}\bar{z} > 1.645) = 1 - \Phi(1.645 - \sqrt{T}\mu_o).$$

Given that $\mu_o > 0$, this probability must be greater than the test size (0.05), so that the test is unbiased. On the other hand, when T increases, $1.645 - \sqrt{T}\mu_o$ becomes smaller so that $\Phi(1.645 - \sqrt{T}\mu_o)$ decreases. Thus, the test power improves when T increases. In particular, as T tends to infinity, the power of this test approaches one, showing that $\sqrt{T}\bar{z}$ is a consistent test. □

© Chung-Ming Kuan, 2007, 2009, 2010

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Amemiya, Takeshi (1994). *Introduction to Statistics and Econometrics*, Cambridge, MA: Harvard University Press.

Rudin, Walter (1976). *Principles of Mathematical Analysis*, Third edition, New York, NY: McGraw-Hill.

© Chung-Ming Kuan, 2007, 2009, 2010