

BASIC STATISTICS

[1] Random Variable (RV)

- RV are usually denoted by capital: X, Y, Z
- A specific possible value of X is denoted by low case: x.

EX:

X = # faced up when you toss a die; $x = 1, 2, \dots, 6$.

Note that there is a rule (probability) generating X.

Definition of RV:

RV is a variable which can take different values with some probability.

[2] Single RV

1. Probability and Cumulative density functions (pdf, cdf):

(1) Discrete RV

X: a RV with $x = a_1, a_2, \dots, a_n$ (n could be ∞ .)

Definition: Pdf: $f(x) = \Pr(X=x)$.

Cdf: $F(x) = \Pr(X \leq x)$.

Conditions for pdf: 1) $f(x) \geq 0$ for any x.

2) $\sum_x f(x) = 1$.

3) $F(x) \leq 1$.

EX: X = # faced up (a die) with pdf: $f(x) = 1/6$, where $x = 1, \dots, 6$.

(2) Continuous RV

X: a RV with pdf, $f(x)$, and cdf, $F(x)$ where

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^x f(v)dv .$$

- Conditions for pdf:
- 1) $f(x) \geq 0$, for any x .
 - 2) $\int_{\Omega} f(v) dv = 1$, where Ω denotes the range of x .
 - 3) $F(x) \leq 1$.

Computation of $\Pr(a \leq X \leq b)$: $Pr(a \leq X \leq b) = \int_a^b f(v) dv$.

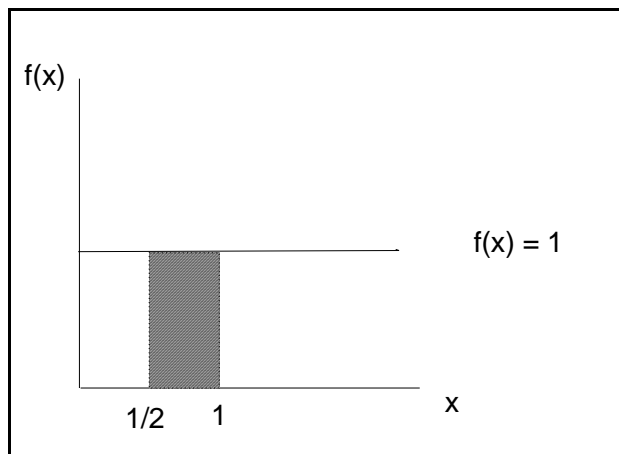
Note: In cases where X is continuous, $\Pr(a \leq X) = \Pr(a < X)$.

EX: (Uniform distribution: Ω)

Ω : $0 \leq x \leq 1$; $f(x) = 1$.

$\Pr(1/2 < x < 1) = \int_{1/2}^1 f(v) dv = [v]_{1/2}^1 = 1 - 1/2 = 1/2$.

$\Pr(1/2 < X < 1)$ = shaded area in the graph below.



2. Expectations:

- General Definition of Expectation:

- $g(X)$ is a function of a RV, X .
- $E[g(x)] = \sum_x g(x)f(x)$ (or $\int_{\Omega} g(x)f(x) dx$).

EX: $g(x) = x, (x - \mu_x)^2, \ln(x)$, etc.

Population Mean: $\mu_x = E(x) = \sum_x xf(x)$ [or $\int_{\Omega} xf(x)dx$].

Population variance: $\sigma_x^2 = \text{var}(x) = E[(x - \mu_x)^2] = \sum_x (x - \mu_x)^2 f(x)$ [or $\int_{\Omega} (x - \mu_x)^2 f(x)dx$]

Standard Deviation (Error): $\sigma_x = \sqrt{\sigma_x^2}$.

Question: What do μ_x and σ_x^2 mean?

[An answer]

- $X = \#$ faced up when you toss a die ($f(x) = 1/6$, $x = 1, 2, \dots, 6$).
- Toss the die repeatedly billions and billions (b) times: $x^{(1)}, x^{(2)}, \dots, x^{(b)}$ [a population].
- Mean of these = $(1/b)\sum_{j=1}^b x^{(j)} = \mu_x$, almost surely (a.s.).
- Mean dispersion of these = $(1/b)\sum_{j=1}^b (x^{(j)} - \mu_x)^2 = \sigma_x^2$, a.s.
- Similarly, $(1/b)\sum_{j=1}^b g[x^{(j)}] = E[g(x)]$, a.s.

Median:

Median of $X = m_x$ such $\Pr(X \leq m_x) \geq 1/2$ and $\Pr(X \geq m_x) \geq 1/2$.

→ Order $x^{(1)}, \dots, x^{(b)}$: $x^{[1]} \leq x^{[2]} \leq \dots \leq x^{[b]}$.

→ $m_x =$ the middle point of this order, a.s.

Fact: If $f(x)$ is symmetric around μ_x , $\mu_x = m_x$.

Some useful theorems:

X : RV; a, b, c : constants.

- $E(ax+b) = aE(x) + b$.
- $\text{var}(x) = E(x^2) - \mu_x^2$.
- $\text{var}(ax+b) = a^2\text{var}(x)$.

Definition:

Let $\mu_3 = E[(x-\mu_x)^3]$; and $\mu_4 = E[(x-\mu_x)^4]$.

Skewness coefficient (SC) = μ_3/σ_x^3 ; Kurtosis coefficient (KC) = $\mu_4/\sigma_x^4 - 3$.

Note:

- SC measures the asymmetry of the distribution of x around μ_x .
 - If $f(x)$ is symmetrically distributed around μ_x , $SC = 0$.
 - If $SC > 0$, the “long tail” is in the $(x \geq \mu_x)$ direction.

- KC measures the thickness of the tails of a distribution:

If X is normally distributed, $KC = 0$.

Exercise for $E(x)$, $\text{var}(x)$ and $E[g(x)]$:

- $X = 1, 0$ with $f(x) = 1/2$.

$$E(x) = \sum_x xf(x) = 0 \times (1/2) + 1 \times (1/2) = 1/2; \text{var}(x) = (0-1/2)^2 \times (1/2) + (1-1/2)^2 \times (1/2) = 1/4 .$$

- $g(x) = (1/2)x^2 + (1/2)x + 2$.

$$E[g(x)] = [1/2+1/2+2] \times (1/2) + [0+0+2] \times (1/2) = 5/2.$$

- Compute SC and KC. Do this by yourself.

A Digression for Fun

- $X = \#$ faced up when you toss a die ($f(x) = 1/6$, $x = 1, 2, \dots, 6$).
- Consider a repeated game:
 - You are a statistician hired by a Mafia.
 - Should forecast the outcome from the die: $\hat{x} =$ your forecast of x .
 - Lose money whenever your forecast is wrong: $s = (x - \hat{x})^2$ [loss function].
 - Should Repeat this game billions and billions times.
 - Wish to choose \hat{x} which minimizes your average loss:

$$\min E(s) = E[(x - \hat{x})^2] .$$
 → Best choice of $\hat{x} = \mu_x!!!$
 - Average loss from choosing $\mu_x = E[(x - \mu_x)^2] = \text{var}(x)$.
- What if $s = |x - \hat{x}|$? → Best choice = m_x .

End of digression

3. Examples of pdf's:

(1) Poisson Distribution:

- EX: # of times to visit doctors; # of job offers; # of patents.
- Pdf: $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, \dots$
- $E(x) = \text{var}(x) = \lambda$.

(2) Normal distribution

- $X \sim N(\mu_x, \sigma_x^2)$, where $E(x) = \mu_x$ and $\text{var}(x) = \sigma_x^2$.
- Pdf: $f(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right]$, $-\infty < x < \infty$.
- $f(x)$ is symmetric around $x = \mu_x$.

Standard Normal Distribution: $z \sim N(0,1)$.

- Pdf: $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$, $-\infty < z < \infty$.
- Fact: $x \sim N(\mu_x, \sigma_x^2) \rightarrow (x-\mu_x)/\sigma_x \sim N(0,1)$.

(3) χ^2 (chi-square) distribution

- Z_1, \dots, Z_k are RVs iid with $N(0,1)$.
 $y = \sum_{i=1}^k Z_i^2 \sim \chi^2(k)$, $y > 0$, with degrees of freedom (df) = k .
- $E(y) = k$; $\text{var}(y) = 2k$.

(4) Student t distribution

- Let $z \sim N(0,1)$ and $y \sim \chi^2(k)$. Z and Y are sto. indep. Then,

$$t = \frac{z}{\sqrt{y/k}} \sim t(k).$$

- $E(t) = 0$, $k > 1$; $\text{var}(t) = k/(k-2)$, $k > 2$.
- As $k \rightarrow \infty$, $\text{var}(t) \rightarrow 1$: In fact, $t \rightarrow z$.
- The pdf of t is similar to that of z , but t has thicker tails.
- $f(t)$ is symmetric around $t = 0$.

(5) F distribution.

- Let $y_1 \sim \chi^2(k_1)$ and $y_2 \sim \chi^2(k_2)$ be sto. indep. Then,

$$f = \frac{y_1/k_1}{y_2/k_2} \sim f(k_1, k_2) .$$

- $f(1, k_2) = t(k_2)^2$.
- $f \sim f(k_1, k_2) \Rightarrow k_1 f \rightarrow \chi^2(k_1)$ as $k_2 \rightarrow \infty$.

[3] Bivariate Distributions

Consider two RVs, X, Y with joint pdf: $f(x,y) = \Pr(X=x, Y=y)$.

Marginal (unconditional) pdf:

$$f_x(x) = \sum_y f(x,y) = \Pr(X=x) \text{ regardless of } Y; f_y(y) = \sum_x f(x,y) = \Pr(Y=y) \text{ regardless of } X.$$

Conditional pdf:

$$f(x|y) = \Pr(X = x, \text{ given } Y = y) = f(x,y)/f_y(y).$$

Stochastic Independence:

- X and Y are sto. indep. iff $f(x,y) = f_x(x)f_y(y)$, for all x,y.
- Under this condition, $f(x|y) = f(x,y)/f_y(y) = [f_x(x)f_y(y)]/f_y(y) = f_x(x)$.

EX:

- Tossing two coins, A and B.
- X = 1 if head from A; = 0 if tail from A.

Y = 1 if head from B; = 0 if tail from B.

$f(x,y) = 1/4$ for any $x,y = 0, 1$. (4 possible cases)

- Marginal pdf of x:

$$f_x(0) = \Pr(X=0) \text{ regardless of } y = f(0,1) + f(0,0) = 1/4 + 1/4 = 1/2.$$

$$f_x(1) = \Pr(X=1) \text{ regardless of } y = f(1,1) + f(1,0) = 1/4 + 1/4 = 1/2.$$

$$\rightarrow f_x(x) = 1/2, x = 0, 1.$$

Similarly, $f_y(y) = 1/2, y = 0, 1$.

- Conditional pdf:

$$f(x=1|y=1) = f(1,1)/f_y(1) = (1/4)/(1/2) = 1/2; f(x=0|y=1) = f(0,1)/f_y(1) = 1/2.$$

$$\rightarrow f(x|y=1) = 1/2, x = 0, 1.$$

- Find $f(y|x=0)$ by yourself.

- Stochastic independence:

$$f_x(x) = f_y(y) = 1/2; f_x(x)f_y(y) = 1/4 = f(x,y), \text{ for any } x \text{ and } y.$$

\rightarrow x and y are stochastically independent.

EX:

The joint probability distribution of x and y is given by the following table: (e.g., $f(4,9) = 0$.)

$x \backslash y$	1	3	9
2	1/8	1/24	1/12
4	1/4	1/4	0
6	1/8	1/24	1/12

- (1) Find the marginal pdf of y .
- (2) Are x and y stochastically independent?
- (3) Find the conditional pdf of y given that $x = 2$.

Expectation: $E[g(x,y)] = \sum_x \sum_y g(x,y)f(x,y)$ [or $\int \int_{\Omega} g(x,y) f(x,y) dx dy$].

Covariance: $\sigma_{xy} = \text{cov}(x,y) = E[(x-\mu_x)(y-\mu_y)]$.

Note: $\sigma_{xy} = \text{cov}(x,y) > 0 \Rightarrow$ positively linearly related; $\sigma_{xy} = \text{cov}(x,y) < 0 \Rightarrow$ negatively linearly related;
 $\sigma_{xy} = \text{cov}(x,y) = 0 \Rightarrow$ no linear relation.

Correlation Coefficient:

The correlation coefficient between x and y is defined by:

$$\rho_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} .$$

Note: $\sigma_{xy} = \rho_{xy}\sigma_x\sigma_y$.

Theorem: $-1 \leq \rho_{xy} \leq 1$.

Note: $\rho_{xy} \rightarrow 1$: highly positively linearly related; $\rho_{xy} \rightarrow -1$; highly negatively linearly related;
 $\rho_{xy} \rightarrow 0$: no linear relation.

Theorem: If X & Y are stoch. indep., $\text{cov}(x,y) = 0$. But not vice versa.

An exercise for computing $E[g(x,y)]$:

$x, y = 1, 0$, with $f(x,y) = 1/4$.

$$E(xy) = \sum_x \sum_y xyf(x,y) = 0 \times 0 \times (1/4) + 0 \times 1 \times (1/4) + 1 \times 0 \times (1/4) + 1 \times 1 \times (1/4) = 1/4.$$

Conditioning in a Bivariate Distribution:

X,Y: RVs with $f(x,y)$. (Y = consumption, X = income)

Population of billions and billions: $\{(x^{(1)}, y^{(1)}), \dots, (x^{(b)}, y^{(b)})\}$.

Average of $y^{(j)} = E(y)$.

For people earning a specific income x, what is the average of y?

Conditional Mean and Variance:

$$E(y|x) = E(y|X=x) = \sum_y yf(y|x).$$

$$\text{var}(y|x) = E[(y-E(y|x))^2|x] = \sum_y (y-E(y|x))^2 f(y|x).$$

Regression model:

$$\epsilon = y - E(y|x).$$

→ $y = y - E(y|x) + E(y|x) = E(y|x) + \epsilon$ (regression model).

→ $E(y|x)$ = explained part of y by x.

→ ϵ = unexplained part of y (called disturbance term).

→ $E(\epsilon|x) = 0$ and $\text{var}(\epsilon|x) = \text{var}(y|x)$.

Note:

- $E(y|x)$ may vary with x , i.e., $E(y|x)$ is a function of x .
- Thus, we can define $E_x[E(y|x)]$, where $E_x(\cdot)$ is the expectation over $x = \sum_x \cdot f_x(x)$ or $\int_x \cdot f_x(x) dx$.

Theorem: (Law of Iterative Expectations)

$$E(y) \text{ [unconditional mean]} = E_x[E(y|x)] .$$

Proof:

$$E(y) = \sum_x \sum_y y f(x,y) = \sum_x \sum_y y f(y|x) f_x(x) = \sum_x [\sum_y y f(y|x)] f_x(x).$$

Note: For discrete RV, X with $x = x_1, \dots$,

$$E(y) = \sum_x E(y|x) f_x(x) = E(y|x=x_1) f_x(x_1) + E(y|x=x_2) f_x(x_2) + \dots .$$

Implication:

If you know conditional mean of y and marginal distribution of x , you can also find unconditional mean of y too.

EX 1: Suppose $E(y|x) = 0$, for all x . $\rightarrow E(y) = E_x[E(y|x)] = E_x(0) = 0$.

EX 2: $E(y|x) = \beta_1 + \beta_2 x$ (linear regression line). $\rightarrow E(y) = E_x(\beta_1 + \beta_2 x) = \beta_1 + \beta_2 E(x)$.

Question: When can $E(y|x)$ be linear? Answered later.

Definition: We say that y is homoskedastic if $\text{var}(y|x)$ is constant.

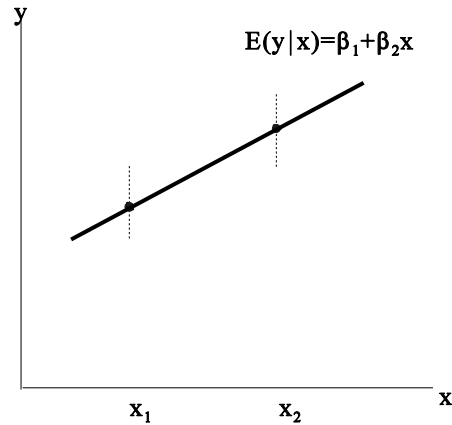
EX: $y = E(y|x) + \epsilon$ with $\text{var}(\epsilon|x) = \sigma^2$ (constant).

$$\rightarrow \text{var}(y|x) = \sigma^2$$

$\rightarrow y$ is homoskedastic.

Graphical Interpretation of Conditional Means and Variances

- Consider the following population:



- $E(y|x=x_1)$ measures the average value of y for the group of $x = x_1$.
- $\text{var}(y|x=x_1)$ measures the dispersion of y given $x = x_1$.
- If $\text{var}(y|x=x_1) = \text{var}(y|x=x_2) = \dots$, we say that y is homoskedastic.
- Law of iterative expectation:

$$E(y) = \sum_x E(y|x)f_x(x) = E(y|x=x_1)\Pr(x=x_1) + E(y|x=x_2)\Pr(x=x_2) + \dots$$

Question: It is worth finding $E(y|x)$?

Theorem: (Decomposition of Variance)

$$\text{var}(y) = \text{var}_x[E(y|x)] + E_x[\text{var}(y|x)].$$

Implication: $\text{var}_x[E(y|x)] \leq \text{var}(y)$, since $E_x[\text{var}(y|x)] \geq 0$.

Coefficient of Determination:

$$R^2 = \text{var}_x[E(y|x)]/\text{var}(y).$$

→ measure of worthiness of knowing $E(y|x)$.

→ $0 \leq R^2 \leq 1$.

Note:

- $\text{var}(y)$ = total variation of y .
- $\text{var}_x[E(y|x)]$ → a part of variation in y due to variation in $E(y|x)$
 = variation in y explained by $E(y|x)$.
- R^2 = variation in y explained by $E(y|x)$ /total variation of y .

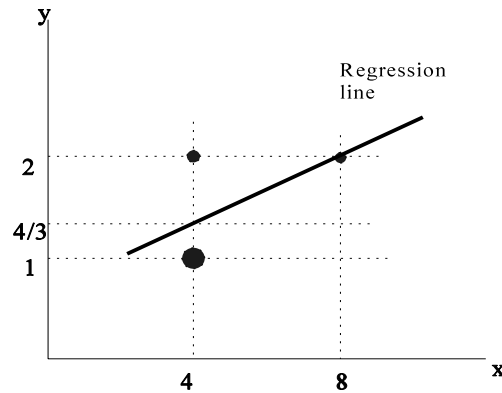
- Wish R^2 close to 1.

Summarizing Exercise:

- A population with X (income=\$10,000) and Y (consumption=\$10,000).
- Joint Pdf:

Y\X	4	8
1	1/2	0
2	1/4	1/4

- Graph for this population:



- Marginal Pdf:

Y\X	4	8	$f_y(y)$
1	1/2	0	1/2
2	1/4	1/4	1/2
$f_x(x)$	3/4	1/4	

- Means of X and Y:

- $E(x) \equiv \mu_x = \sum_x x f_x(x) = 4 \times f_x(4) + 8 \times f_x(8) = 4 \times (3/4) + 8 \times (1/4) = 5.$
- $E(y) \equiv \mu_y = \sum_y y f_y(y) = 1.5$

- Variances of X and Y:

- $\text{var}(x) \equiv \sigma_x^2 = \sum_x (x - \mu_x)^2 f_x(x) = (4 - 5)^2 f_x(4) + (8 - 5)^2 f_x(8) = 1 \times (3/4) + 9 \times (1/4) = 3.$
- $\text{var}(y) \equiv \sigma_y^2 = 1/4.$

- Covariance between X and Y:

- $\text{cov}(x,y) \equiv E[(x-\mu_x)(y-\mu_y)] = E(xy) - \mu_x\mu_y = \sum_x \sum_y xyf(x,y) - \mu_x\mu_y$
 $= 4 \times 1 \times f(4,1) + 4 \times 2 \times f(4,2) + 8 \times 1 \times f(8,1) + 8 \times 2 \times f(8,2) - 5 \times 1.5 = 0.5.$

- $\rho_{xy} \equiv \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} = \frac{0.5}{\sqrt{3}\sqrt{1/4}} \approx 0.58.$

- Conditional Probabilities

Y\X	4	8	f_y(y)
1	1/2	0	1/2
2	1/4	1/4	1/2
f _x (x)	3/4	1/4	

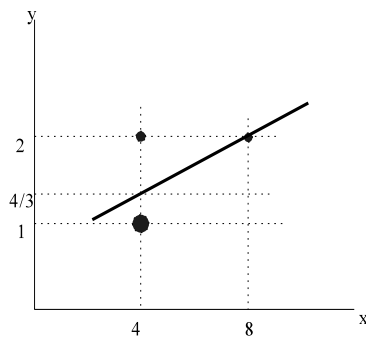
- f(y|x):

Y\X	4	8
1	2/3	0
2	1/3	1

- Conditional mean:

- $E(y|x=4) = \sum_y yf(y|x=4) = 1 \times f(y=1|x=4) + 2 \times f(y=2|x=4) = 1 \times (2/3) + 2 \times (1/3) = 4/3$

- $E(y|x=8) = 2.$



- Conditional variance of Y:

- $\text{var}(y|x=4) = \sum_y [y - E(y|x=4)]^2 f(y|x=4) = 6/27.$

- $\text{var}(y|x=8) = 0.$

- Law of iterative expectation:

- $E_x[E(y|x)] = \sum_x E(y|x)f_x(x)$

$$= E(y|x=4)f_x(4) + E(y|x=8)f_x(8)$$

$$= (4/3) \times (3/4) + 2 \times (1/4) = 1.5 = E(y)!!!$$

[4] Bivariate Normal Distribution

Definition: (Bivariate Normal Distribution)

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right).$$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp \left[-\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right\} \right], x,y \in \mathbb{R}.$$

Here, $\text{cov}(x,y) = \sigma_{xy} = \rho\sigma_x\sigma_y$.

Facts:

- 1) $f_x(x) \sim N(\mu_x, \sigma_x^2)$ and $f_y(y) \sim N(\mu_y, \sigma_y^2)$.
- 2) $E(y|x) = \beta_1 + \beta_2x$ and $\text{var}(y|x) = \sigma^2$ (constant) [See Greene.]
 → $E(y|x)$ is linear in x and y is homoskedastic.
- 3) If $\rho = 0$ ($\sigma_{xy} = 0$), x and y are stochastically independent.

[5] Multivariate Distributions

1. Mean vector and covariance matrix:

Definition: X_1, \dots, X_n : random variables.

Let $x = [x_1, \dots, x_n]'$ ($n \times 1$ vector). Then,

$$E(x) = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_n) \end{bmatrix}; \text{Cov}(x) = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3) & \dots & \text{cov}(x_1, x_n) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) & \text{cov}(x_2, x_3) & \dots & \text{cov}(x_2, x_n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_n, x_1) & \text{cov}(x_n, x_2) & \text{cov}(x_n, x_3) & \dots & \text{var}(x_n) \end{bmatrix}.$$

→ $\text{Cov}(x)$ is symmetric.

Note: **In Greene, Cov(x) is denoted by Var(x).**

Definition: (Expectation of random matrix)

Suppose that B_{ij} are RVs. Then,

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{nn} \end{bmatrix} \Rightarrow E(B) = \begin{bmatrix} E(B_{11}) & E(B_{12}) & \cdots & E(B_{1n}) \\ E(B_{21}) & E(B_{22}) & \cdots & E(B_{2n}) \\ \vdots & \vdots & & \vdots \\ E(B_{n1}) & E(B_{n2}) & \cdots & E(B_{nn}) \end{bmatrix}.$$

Theorem: $\text{Cov}(x) = E[(x-\mu_x)(x-\mu_x)'] = E(xx') - \mu_x\mu_x'$.

Proof: See Greene.

EX: If x is scalar, $\text{Cov}(x) = E[(x-\mu)^2] = \text{var}(x)$.

EX: $x = [x_1, x_2]'$; $E(x) = \mu = [\mu_1, \mu_2]'$

$$\rightarrow x - \mu = [x_1 - \mu_1, x_2 - \mu_2]'$$

$$\rightarrow (x - \mu)(x - \mu)' = \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} [x_1 - \mu_1, x_2 - \mu_2] = \begin{bmatrix} (x_1 - \mu_1)^2 & (x_1 - \mu_1)(x_2 - \mu_2) \\ (x_1 - \mu_1)(x_2 - \mu_2) & (x_2 - \mu_2)^2 \end{bmatrix}$$

$$\rightarrow E[(x-\mu)(x-\mu)'] = \text{Cov}(x).$$

2. Mean and Variance of a linear combination of RVs:

Definition:

Let $X = [X_1, \dots, X_n]'$ be a random vector and let $c = [c_1, \dots, c_n]'$ be a $n \times 1$ vector of fixed constants.

Then,

$$c'x = x'c = c_1x_1 + \dots + c_nx_n = \sum_j c_jx_j \text{ (scalar)}.$$

Theorem:

$$(1) E(c'x) = c'E(x)$$

$$(2) \text{var}(c'x) = c'\text{Cov}(x)c.$$

Proof:

$$(1) E(c'x) = E(\sum_j c_j x_j) = \sum_j E(c_j x_j) = \sum_j E(c_j x_{j1} + \dots + c_j x_{jn}) = c_1 E(x_1) + \dots + c_n E(x_n) = \sum_j c_j E(x_j) = c' E(x).$$

$$(2) \text{var}(c'x) = E[(c'x - E(c'x))^2] = E[\{c'x - c'E(x)\}^2] = E[\{c'(x-E(x))\}^2] \\ = E[\{c'(x-E(x))\} \{c'(x-E(x))\}] = E[\{c'(x-E(x))\} \{(x-E(x))'c\}] \\ = E[c'(x-E(x))(x-E(x))'c] = c'E[(x-E(x))(x-E(x))']c = c' \text{Cov}(x)c.$$

Remark:

(2) implies that $\text{Cov}(x)$ is always positive semidefinite.

→ $c' \text{Cov}(x)c \geq 0$ for any nonzero vector c .

Proof:

For any nonzero vector c , $c' \text{Cov}(x)c = \text{var}(c'x) \geq 0$.

Remark:

- $\text{Cov}(x)$ is symmetric and positive semidefinite.
- Usually, $\text{Cov}(x)$ is positive definite, that is, $c' \text{Cov}(x)c > 0$, for any nonzero vector c .

Digression to Definite Matrices

Definition:

Let $B = [b_{ij}]_{n \times n}$ be a symmetric matrix, and $c = [c_1, \dots, c_n]'$. Then, the scalar, $c' B c$, is called a quadratic form of B .

Definition:

If $c' B c > (<) 0$ for any nonzero vector c , B is called positive (negative) definite.

If $c' B c \geq (\leq) 0$ for any nonzero c , B is called positive (negative) semidefinite.

Theorem:

Let B be a symmetric and square matrix given by:

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{12} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{1n} & b_{2n} & \dots & b_{nn} \end{bmatrix}.$$

Define the principal minors by:

$$|B_1| = b_{11} ; |B_2| = \begin{vmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{vmatrix} ; |B_3| = \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{12} & b_{22} & b_{23} \\ b_{13} & b_{23} & b_{33} \end{vmatrix} ; \dots$$

B is positive definite iff $|B_1|, |B_2|, \dots, |B_n|$ are all positive. B is negative definite iff $|B_1| < 0, |B_2| > 0, |B_3| < 0, \dots$

EX:

Show that B is positive definite:

$$B = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

End of Digression

Theorem:

Let X be a $n \times 1$ random vector and let A be a $m \times n$ matrix of constants (Ax is a $m \times 1$ random vector). Then,

$$E(Ax) = AE(x); \text{Cov}(Ax) = ACov(x)A'$$

[6] Multivariate Normal distribution

Definition:

$x = [x_1, \dots, x_n]'$ is a normal vector, i.e., each of the x_i 's is normal.

Let $E(x) = \mu = [\mu_1, \dots, \mu_n]'$ and $\text{Cov}(x) = \Sigma = [\sigma_{ij}]_{n \times n}$. Then,

$$x \sim N(\mu, \Sigma).$$

Pdf of x:

$$f(x) = f(x_1, \dots, x_n) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp[-(1/2)(x-\mu)' \Sigma^{-1}(x-\mu)] ,$$

where $|\Sigma| = \det(\Sigma)$.

EX:

Let X be a single RV with $N(\mu_x, \sigma_x^2)$. Then,

$$f(x) = (2\pi)^{-1/2}(\sigma_x^2)^{-1/2}\exp[-(1/2)(x-\mu_x)(\sigma_x^2)^{-1}(x-\mu_x)] = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right].$$

EX:

Assume that all the x_i are iid with $N(\mu_x, \sigma_x^2)$. Then,

$$(1) \mu = E(x) = [\mu_x, \dots, \mu_x]'$$
 ;

$$(2) \Sigma = \text{Cov}(x) = \text{diag}(\sigma_x^2, \dots, \sigma_x^2) = \sigma_x^2 \mathbf{I}_n .$$

Using (1) and (2), we can show that $f(x) = f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$, where,

$$f(x_i) = = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left[-\frac{(x_i - \mu_x)^2}{2\sigma_x^2}\right].$$

1. Conditional normal distribution

$[y, x_2, \dots, x_k]'$ is a normal vector. Then,

$$E(y|x_2, \dots, x_k) = \beta_1 + \beta_2 x_2 + \dots + \beta_k x_k = \mathbf{x}^* \beta$$

$$[\mathbf{x}^{*'} = (1, x_2, \dots, x_k) \text{ and } \beta = (\beta_1, \dots, \beta_k)']$$

$$\text{var}(y|\mathbf{x}^*) = \sigma^2 .$$

→ The regression of y on x_1, \dots, x_k is linear & homoskedatic.

Proof: See Greene.

2. Distributions of linear functions of a normal vector

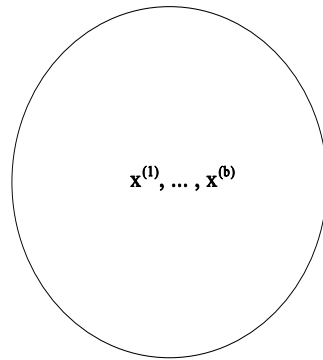
$$x_{n \times 1} \sim N(\mu, \Sigma).$$

$$y = \mathbf{A}x + \mathbf{b}, \text{ where } \mathbf{A}_{m \times n} \text{ and } \mathbf{b}_{m \times 1} \text{ are fixed.}$$

$$\rightarrow y \sim N(\mathbf{A}\mu + \mathbf{b}, \mathbf{A}\Sigma\mathbf{A}').$$

[7] Sample and Estimator

(1) A population (of billions and billions)



- A unknown characteristic of the population is denoted by $\theta \in \mathbb{R}$.
 (θ is called a unknown parameter of interest.)
 (θ could be the population mean or population variance.)
- Wish to estimate θ .
- $\{x_1, \dots, x_T\}$: a sample of size T from the population.
- $\hat{\theta}$: an estimator of θ , which is a function of the sample.
 (e.g, $\hat{\theta} = \bar{x} = (1/T)\sum_{t=1}^T x_t$.)
- A sample is random, in the sense that there are many possible samples of size T.

Set of all possible samples	Estimates
(Set of all possible samples)	
SAM 1: $\{x_1^{[1]}, \dots, x_t^{[1]}, \dots, x_T^{[1]}\}$	$\rightarrow \hat{\theta}^{[1]}$
SAM 2: $\{x_1^{[2]}, \dots, x_t^{[2]}, \dots, x_T^{[2]}\}$	$\rightarrow \hat{\theta}^{[2]}$
:	:
SAM b': $\{x_1^{[b']}, \dots, x_t^{[b']}, \dots, x_T^{[b']}\}$	$\rightarrow \hat{\theta}^{[b']}$
Since $\{x_1, \dots, x_T\}$ is random, so is $\hat{\theta}$.	\rightarrow We can define $E(\hat{\theta})$ and $\text{var}(\hat{\theta})$.

(2) Meaning of “a random sample (RS) from a distribution f(x)”

- Means that x_1, \dots, x_T are iid.
- EX: $\{x_1, \dots, x_T\}$ a RS from $N(\mu, \sigma^2)$.
 → $x_t \sim N(\mu, \sigma^2)$ for any $t = 1, \dots, T$.
 → $E(x_t) = \mu$ and $\text{var}(x_t) = \sigma^2$, for any $t = 1, \dots, T$.

Note:

A sample need not be iid.

→ Let x_t be the height of the t 'th person (cross-section data)

→ Likely to be independent of others' height.

→ Likely to be identically distributed.

→ Let x_t be US GNP at time t (time-series data)

→ x_t and x_{t-1} are likely to be correlated.

→ x_1, \dots, x_T are not iid.

(3) Criteria for a “good” estimator

1) Minimum Variance Unbiased Estimator

Definition: $E(\hat{\theta}) = \theta \rightarrow \hat{\theta}$ is called a unbiased estimator of θ .

Implication: $\hat{\theta}^{[1]}, \dots, \hat{\theta}^{[b']}$ $\rightarrow (1/b') \sum_{j=1}^{b'} \hat{\theta}^{[j]} = \theta$, a.s.

EX:

$\{x_1, \dots, x_T\}$: RS from a dist. with μ and σ^2 .

$\bar{x} = (1/T) \sum_{t=1}^T x_t$; $s_x^2 = [1/(T-1)] \sum_{t=1}^T (x_t - \bar{x})^2$.

→ $E(\bar{x}) = \mu$ and $E(s_x^2) = \sigma^2$.

→ So, \bar{x} and s_x^2 are unbiased estimators of μ and σ^2 , respectively.

Definition:

Let $\hat{\theta}$ and $\tilde{\theta}$ be unbiased estimators of θ .

$\text{var}(\tilde{\theta}) > \text{var}(\hat{\theta}) \Rightarrow \hat{\theta}$ is more efficient than $\tilde{\theta}$.

Implication:

$\hat{\theta}$: $\hat{\theta}^{[1]}, \dots, \hat{\theta}^{[b']}$;

$\tilde{\theta}$: $\tilde{\theta}^{[1]}, \dots, \tilde{\theta}^{[b']}$.

$\text{var}(\tilde{\theta}) > \text{var}(\hat{\theta}) \rightarrow \text{Dispersion of } \tilde{\theta}^{[1]}, \dots, \tilde{\theta}^{[b']} > \text{Dispersion of } \hat{\theta}^{[1]}, \dots, \hat{\theta}^{[b']}$

→ $\hat{\theta}$ is less sensitive to the chosen sample.

EX: $\{x_1, \dots, x_T\}$: RS from a dist. with μ and σ^2 .

$$\tilde{x} = x_1.$$

$$\rightarrow E(\tilde{x}) = E(x_1) = \mu \text{ (unbiased).}$$

$$\rightarrow \text{var}(\tilde{x}) = \text{var}(x_1) = \sigma^2.$$

$$\rightarrow \text{But, } \text{var}(\bar{x}) = \sigma^2/T.$$

$\rightarrow \bar{x}$ is more efficient than \tilde{x} .

Definition:

$\hat{\theta}$: a unbiased estimator.

$\hat{\theta}$ is MVUE iff $\text{var}(\tilde{\theta}) \geq \text{var}(\hat{\theta})$ for any unbiased estimator $\tilde{\theta}$.

\rightarrow Say that $\hat{\theta}$ is efficient.

2) Minimum Mean Square Error (MMSE) Estimator

Definition:

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

Note: If $E(\hat{\theta}) = \theta$, $\text{var}(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2] = E[(\hat{\theta} - \theta)^2] = \text{MSE}(\hat{\theta})$.

Theorem:

Let $\text{Bias}(\hat{\theta}) = E(\hat{\theta} - \theta)$. Then, $\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$.

Definition:

The MMSE estimator minimizes $\text{MSE}(\hat{\theta})$.

Note:

1) MMSE estimator could be biased.

2) MMSE is usually a function of θ .

\rightarrow To get MMSE, need to know θ .

\rightarrow If you know θ , why do you estimate?

\rightarrow If we wish to test for some hypotheses regarding θ , MVUE is more meaningful.

(3) How to find MVUE

Notational Change:

- From now on, we denote the true value of θ as θ_o .
- Then, view θ as a variable.

Definition: (Likelihood function)

- joint pdf of $x_1, \dots, x_T = f(x_1, \dots, x_T, \theta_o)$.
- $L_T(\theta) = f(x_1, \dots, x_T, \theta)$ (likelihood function).

Remark:

- $L_T(\theta)$ is a joint pdf of x_1, \dots, x_T replacing θ_o by θ .
- View $L_T(\theta)$ as a function of θ given x_1, \dots, x_T .

Definition: (log-likelihood function)

$$l_T(\theta) = \ln[f(x_1, \dots, x_T, \theta)].$$

EX:

$\{x_1, \dots, x_T\}$: RS from a dist. with $f(x, \theta_o)$.

- $x_t \sim f(x_t, \theta_o)$.
- $f(x_1, \dots, x_T, \theta_o) = \prod_{t=1}^T f(x_t, \theta_o)$.
- $f(x_1, \dots, x_T, \theta) = \prod_{t=1}^T f(x_t, \theta)$.
- $\ln[f(x_1, \dots, x_T, \theta)] = \sum_{t=1}^T \ln[f(x_t, \theta)]$.
- $l_T(\theta) = \sum_{t=1}^T \ln[f(x_t, \theta)]$.

Definition: (Maximum Likelihood Estimator)

MLE $\hat{\theta}$ maximizes $l_T(\theta)$ given data points x_1, \dots, x_T .

Theorem:

If $\hat{\theta}$ is MLE and $E(\hat{\theta}) = \theta_o$, $\hat{\theta}$ is an efficient estimator.

Theorem:

Let $\hat{\theta}$ be MLE. Suppose $E(\hat{\theta}) \neq \theta_o$. Suppose $\exists g(\hat{\theta}) \ni E[g(\hat{\theta})] = \theta_o$. Then, $g(\hat{\theta})$ is efficient.

EX:

$\{x_1, \dots, x_T\}$: RS from a Poisson dist., $f(x, \theta) = e^{-\theta} \theta^x / x!$ [Suppressing subscript “o” from θ].

[Note $E(x) = \text{var}(x) = \theta_o$.]

$$\rightarrow l_T(\theta) = \sum_t \ln[f(x_t, \theta)] = \sum_t [-\theta + x_t \ln(\theta) - \ln(x_t!)]$$

$$\rightarrow \text{FOC (first order condition): } \partial l_T(\theta) / \partial \theta = \sum_t [-1 + x_t / \theta] = 0$$

$$\rightarrow -T + (1/\theta) \sum_t x_t = 0 \rightarrow -T\theta + \sum_t x_t = 0 \rightarrow \hat{\theta} = (1/T) \sum_t x_t = \bar{x}.$$

$$\rightarrow E(\hat{\theta}) = E(\bar{x}) = \theta.$$

$\rightarrow \hat{\theta}$ Efficient.

[8] Extention to the Estimation of Multiple Parameters

Definition:

$\theta_o = [\theta_{o,1}, \theta_{o,2}, \dots, \theta_{o,p}]'$: the unknown parameter vector.

$\hat{\theta} = [\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p]'$, where $\hat{\theta}_j$ is a function of $\{x_1, \dots, x_T\}$.

Definition: (Unbiasedness)

$\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta_o$:

$$E(\hat{\theta}) = \begin{bmatrix} E(\hat{\theta}_1) \\ E(\hat{\theta}_2) \\ \vdots \\ E(\hat{\theta}_p) \end{bmatrix} = \begin{bmatrix} \theta_{o,1} \\ \theta_{o,2} \\ \vdots \\ \theta_{o,p} \end{bmatrix} = \theta_o.$$

Definition: (Relative Efficiency)

$\tilde{\theta}, \hat{\theta}$: unbiased estimators.

$c = [c_1, \dots, c_p]'$ be any nonzero vector.

$\hat{\theta}$ is said to be efficient relative to $\tilde{\theta}$ iff $\text{var}(c' \tilde{\theta}) \geq \text{var}(c' \hat{\theta})$.

$$\leftrightarrow c' \text{Cov}(\tilde{\theta}) c - c' \text{Cov}(\hat{\theta}) c \geq 0$$

$$\leftrightarrow c' [\text{Cov}(\tilde{\theta}) - \text{Cov}(\hat{\theta})] c \geq 0$$

$$\leftrightarrow [\text{Cov}(\tilde{\theta}) - \text{Cov}(\hat{\theta})] \text{ is positive semidefinite.}$$

Note:

- Let $\theta = (\theta_1, \theta_2)'$ and $c = (c_1, c_2)'$.

- Suppose you wish to estimate $c'\theta = c_1\theta_1 + c_2\theta_2$.
- Suppose you have $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)'$ and $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)'$.
- If, for any c , $\text{var}(c'\tilde{\theta}) = \text{var}(c_1\tilde{\theta}_1 + c_2\tilde{\theta}_2) < \text{var}(c_1\hat{\theta}_1 + c_2\hat{\theta}_2) = \text{var}(c'\hat{\theta})$, we can say that $\tilde{\theta}$ is a better estimator.

EX: Let $\theta = (\theta_1, \theta_2)'$. Suppose:

$$\text{Cov}(\hat{\theta}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \text{Cov}(\tilde{\theta}) = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}$$

→ $\text{var}(\hat{\theta}_1) = 1 < 1.5 = \text{var}(\tilde{\theta}_1)$; $\text{var}(\hat{\theta}_2) = 1 < 1.5 = \text{var}(\tilde{\theta}_2)$.

But,

$$\text{Cov}(\tilde{\theta}) - \text{Cov}(\hat{\theta}) = \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix} \equiv A$$

$$|A_1| = 0.5 ; |A_2| = (0.5)^2 - 1 = -0.75 < 0.$$

- A is not positive definite.
- Thus, $\hat{\theta}$ is not necessarily more efficient than $\tilde{\theta}$.
- For example, you wish to estimate $\theta_{o,1} - \theta_{o,2} = c'\theta_o$ ($c' = (1, -1)$).
 - $\text{var}(c'\hat{\theta}) = c'\text{Cov}(\hat{\theta})c = 2$
 - $\text{var}(c'\tilde{\theta}) = c'\text{Cov}(\tilde{\theta})c = 1$
 - Thus, $c'\tilde{\theta}$ is a better estimator of $c'\theta$.
- Depending on c , a better estimator is determined.
 - Can't claim that one estimator is always superior.

Question:

How about the following rule?

$$\text{var}(\hat{\theta}_j) \leq \text{var}(\tilde{\theta}_j), \text{ for any } j = 1, \dots, p.$$

In fact, this rule is weaker than our relative efficiency rule.

Theorem:

If $\hat{\theta}$ is more efficient than $\tilde{\theta}$, $\text{var}(\hat{\theta}_j) \leq \text{var}(\tilde{\theta}_j)$, for any $j = 1, \dots, p$.

But, the reverse is not true.

Proof:

Let $c' = (1, 0, \dots, 0)$. Then, $\text{var}(\hat{\theta}_1) = \text{var}(c' \hat{\theta}) \leq \text{var}(c' \tilde{\theta}) = \text{var}(\tilde{\theta}_1)$.

Definition: (MVUE)

$\hat{\theta}$: a unbiased estimator.

$c = [c_1, \dots, c_p]'$ be any nonzero vector.

$\hat{\theta}$ is said to be efficient iff $\text{var}(c' \tilde{\theta}) \geq \text{var}(c' \hat{\theta})$ for any unbiased $\tilde{\theta}$.

Note:

$$\begin{aligned} \text{var}(c' \tilde{\theta}) \geq \text{var}(c' \hat{\theta}) &\rightarrow c' \text{Cov}(\tilde{\theta})c - c' \text{Cov}(\hat{\theta})c \geq 0 \\ &\rightarrow c' [\text{Cov}(\tilde{\theta}) - \text{Cov}(\hat{\theta})]c \geq 0 \\ &\rightarrow [\text{Cov}(\tilde{\theta}) - \text{Cov}(\hat{\theta})] \text{ is positive semidefinite.} \end{aligned}$$

Definition: (MSE)

$$\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)'] \quad (p \times p).$$

Note: If $E(\hat{\theta}) = \theta_0$, $\text{Cov}(\hat{\theta}) = \text{MSE}(\hat{\theta})$.

Theorem:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{Cov}(\hat{\theta}) + [\theta_0 - E(\hat{\theta})][\theta_0 - E(\hat{\theta})]', \\ &\text{where } [\theta_0 - E(\hat{\theta})] \text{ is called the bias of } \hat{\theta}. \end{aligned}$$

Definition: (Likelihood function)

$$L_T(\theta) = f(x_1, \dots, x_T, \theta) = f(x_1, \dots, x_T, \theta_1, \dots, \theta_p).$$

$$l_T(\theta) = \ln[f(x_1, \dots, x_T, \theta)] = \ln[f(x_1, \dots, x_T, \theta_1, \dots, \theta_p)].$$

Note: If $\{x_1, \dots, x_T\}$ is a RS,

$$l_T(\theta) = \sum_{t=1}^T \ln[f(x_t, \theta)] = \sum_{t=1}^T \ln[f(x_t, \theta_1, \dots, \theta_p)].$$

Definition: (MLE)

MLE $\hat{\theta}$ max. $l_T(\theta)$ given data points x_1, \dots, x_T :

$$\frac{\partial l_T(\hat{\theta})}{\partial \theta} = \begin{bmatrix} \partial l_T(\hat{\theta})/\partial \theta_1 \\ \partial l_T(\hat{\theta})/\partial \theta_2 \\ \vdots \\ \partial l_T/\partial \theta_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_{p \times 1} .$$

Theorem:

Let $\hat{\theta}$ be MLE. If $E(\hat{\theta}) = \theta_o$, it is efficient.

Theorem:

Let $\hat{\theta}$ be MLE. Suppose $E(\hat{\theta}) \neq \theta_o$. Suppose $\exists g(\hat{\theta})_{p \times 1} \ni E[g(\hat{\theta})] = \theta_o$. Then, $g(\hat{\theta})$ is efficient.

EX:

Let x_t be iid with $N(\mu, \sigma^2)$ [suppressing subscript "o" from μ and σ^2]. Let $\theta = (\mu, v)'$ where $v = \sigma^2$.

Note that:

$$f(x_t, \theta) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x_t - \mu)^2}{2v}\right] = (2\pi)^{-1/2} (v)^{-1/2} \exp\left[-\frac{(x_t - \mu)^2}{2v}\right] .$$

$$\ln[f(x_t, \theta)] = (-1/2)\ln(2\pi) - (1/2)\ln(v) - \frac{(x_t - \mu)^2}{2v} .$$

$$l_T(\theta) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln v - \frac{\sum_{t=1}^T (x_t - \mu)^2}{2v} .$$

For MLE, solve:

$$(1) : \frac{\partial l_T}{\partial \mu} = -\frac{1}{2v} \sum_{t=1}^T 2(x_t - \mu)(-1) = \frac{\sum_{t=1}^T (x_t - \mu)}{v} = 0 ,$$

$$(2) : \frac{\partial l_T}{\partial v} = -\frac{T}{2v} + \frac{\sum_{t=1}^T (x_t - \mu)^2}{2v^2} = 0 .$$

From (1):

$$(3) : \sum_t (x_t - \mu) = 0 \rightarrow \sum_t x_t - T\mu = 0$$

$$\rightarrow \hat{\mu}_{ML} = (1/T)\sum_t x_t = \bar{x}.$$

Substituting (3) into (2):

$$-Tv + \sum_t (x_t - \hat{\mu}_{ML})^2 = 0 \rightarrow \hat{v}_{ML} = (1/T)\sum_t (x_t - \hat{\mu}_{ML})^2 = (1/T)\sum_t (x_t - \bar{x})^2.$$

Thus,

$$\hat{\theta}_{ML} = \begin{pmatrix} \hat{\mu}_{ML} \\ \hat{v}_{ML} \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \frac{1}{T}\sum_{t=1}^T (x_t - \bar{x})^2 \end{pmatrix}.$$

Note:

- $E(\hat{\mu}_{ML}) = E(\bar{x}) = \mu_o \rightarrow$ unbiased \rightarrow efficient.
- $E(\hat{v}_{ML}) = \{(T-1)/T\}\sigma_o^2$ (by the fact that $E[(1/(T-1)\sum_t (x_t - \bar{x})^2)] = \sigma_o^2$)
 \rightarrow biased.
- Let $g(\hat{v}_{ML}) = [T/(T-1)]\hat{v}_{ML}$.
 $\rightarrow E[g(\hat{v}_{ML})] = \sigma_o^2$.
 $\rightarrow g(\hat{v}_{ML}) = [1/(T-1)]\sum_t (x_t - \bar{x})^2 = s_x^2$ is efficient.

[9] Large-Sample Theories

(1) Motivation:

- $\hat{\theta}_T$: an estimator from a sample of size T , $\{x_1, \dots, x_T\}$
- What would happen to $\hat{\theta}_T$ if $T \rightarrow \infty$?
- What do we wish?

[We wish $\hat{\theta}_T$ becomes closer to θ_o as T increases.]

(2) Main Points:

- Rough Definition of Consistency:

Suppose that distribution of $\hat{\theta}_T$ becomes condensed around θ_o more and more as T increase.

Then, we say that $\hat{\theta}_T$ is a consistent estimator. And we use the following notation:

$$\text{plim}_{T \rightarrow \infty} \hat{\theta}_T = \theta_o \text{ (or } \hat{\theta}_T \rightarrow_p \theta_o).$$

- Relation between unbiasedness and consistency:

- Biased estimators could be consistent.

EX: Suppose that $\tilde{\theta}$ is unbiased and consistent.

Define $\hat{\theta} = \tilde{\theta} + 1/T$.

Clearly, $E(\hat{\theta}) = \theta_0 + 1/T \neq \theta_0$ (biased)

But, $\text{plim}_{T \rightarrow \infty} \hat{\theta} = \text{plim}_{T \rightarrow \infty} \tilde{\theta} = \theta$ (consistent)

- A unbiased estimator $\hat{\theta}$ is consistent if $\text{var}(\hat{\theta}) \rightarrow 0$ as $T \rightarrow \infty$.

EX: Suppose that $\{x_1, \dots, x_T\}$ is a RS from $N(\mu_0, \sigma_0^2)$.

$$E(\bar{x}) = \mu_0.$$

$$\text{var}(\bar{x}) = \sigma_0^2/T \rightarrow 0 \text{ as } T \rightarrow \infty.$$

Thus, \bar{x} is a consistent estimator of μ_0 .

- Law of Large Numbers (LLN)

A. Case of Scalar Random variables:

- Komogorov's Strong LLN:

Suppose that $\{x_1, \dots, x_T\}$ is a RS from a population with μ_0 and σ_0^2 .

Then, $\text{plim } \bar{x} = \mu_0$.

- Generalized Weak LLN (GWLLN):

- $\{x_1, \dots, x_T\}$ is a sample (not necessarily RS)

- Define $E(x_1) = \mu_{0,1}, \dots, E(x_T) = \mu_{0,T}$.

- Define $\text{var}(x_1) = \sigma_{0,1}^2, \dots, \text{var}(x_T) = \sigma_{0,T}^2$.

Assume that $\sigma_{0,1}^2, \dots, \sigma_{0,T}^2 < \infty$.

- Then, under suitable assumptions, $\text{plim } \bar{x} = \lim \frac{1}{T} \sum_t \mu_{0,t}$.

B. Case of Vector Random Variables:

- GWLLN

- x_t : $p \times 1$ random vector.

- $\{x_1, \dots, x_T\}$ is a sample.

- Let $E(x_1) = \mu_{0,1}$ ($p \times 1$), $\dots, E(x_T) = \mu_{0,T}$.

- Assume that $\text{Cov}(x_j)$ are well-defined and finite.

- Then, under suitable assumptions.

$$\text{plim } \bar{x} = \lim \frac{1}{T} \sum_t \mu_{0,t}.$$

- Central Limit Theorems (CLT)

A. Case of Scalar Random Variables:

- Motivation:
 - Suppose that $\{x_1, \dots, x_T\}$ is a RS from a population with μ_o and σ_o^2 .
 - We know $\bar{x} \rightarrow \mu_o$ as $T \rightarrow \infty$. But we can never have an infinitely large sample!!!
 - For finite T, \bar{x} is still a random variable. What statistical distribution could approximate the true distribution of \bar{x} ?
- Lindberg-Levy CLT:
 - Suppose that $\{x_1, \dots, x_T\}$ is a RS from a population with μ_o and σ_o^2 .
 - Then, $\sqrt{T}(\bar{x} - \mu_o) \rightarrow_d N(0, \sigma_o^2)$, or equivalently, $\sqrt{T} \frac{\bar{x} - \mu_o}{\sigma_o} \rightarrow_d N(0, 1)$.
- Implication of CLT:
 - $\sqrt{T}(\bar{x} - \mu_o) \approx N(0, \sigma_o^2)$, if T is large.
 - $E[\sqrt{T}(\bar{x} - \mu_o)] = \sqrt{T}[E(\bar{x}) - \mu_o] \approx 0 \rightarrow E(\bar{x}) \approx \mu_o$.
 - $\text{var}[\sqrt{T}(\bar{x} - \mu_o)] = T\text{var}(\bar{x} - \mu_o) = T\text{var}(\bar{x}) \approx \sigma_o^2 \rightarrow \text{var}(\bar{x}) \approx \sigma_o^2/T$.
 - $\bar{x} \approx N(\mu_o, \sigma_o^2/T)$, if T is large.

B. Case of Random vectors:

- GCLT
 - $\{y_1, \dots, y_T\}$: a sequence of $p \times 1$ random vectors.
 - For any t, $E(y_t) = 0$ and $\text{Cov}(y_t)$ is well defined and finite.
 - Under some suitable conditions (acceptable for Econometrics I, II),

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T y_t \rightarrow_d N(0, \lim_{T \rightarrow \infty} \frac{1}{T} \text{Cov}(\sum_{t=1}^T y_t)) .$$
- Note:
 - $\text{Cov}(y_t)$ [$\text{var}(y_t)$ if y_t is a scalar] could differ across different t.
 - The y_t could be correlated as long as $\lim_{n \rightarrow \infty} \text{cov}(y_t, y_{t+n}) = 0$ (if the y_t are stationary).
 - If $E(y_t | y_{t-1}, y_{t-2}, \dots, y_1) = 0$ (Martingale Difference Sequence), the y_t 's are linearly uncorrelated. Then,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T y_t \rightarrow_d N(0, \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \text{Cov}(y_t)) .$$

[Technical Details]

(3) Convergency in probability

Definition:

When b and c are scalars, $|b - c|$ = absolute value of $(b-c)$.

When $b = [b_1, \dots, b_p]'$ and $c = [c_1, \dots, c_p]'$ be $p \times 1$ vectors,

$$|b - c| \text{ (norm)} = \sqrt{(b_1 - c_1)^2 + (b_2 - c_2)^2 + \dots + (b_p - c_p)^2} .$$

Definition: (Convergency in probability, Weak Convergency)

$\hat{\theta}_T$ converges in probability to c iff

$$\lim_{T \rightarrow \infty} \Pr[|\hat{\theta}_T - c| < \epsilon] = 1, \text{ for any small } \epsilon > 0.$$

Or equivalently,

$$\lim_{T \rightarrow \infty} \Pr[|\hat{\theta}_T - c| > \epsilon] = 0, \text{ for any small } \epsilon > 0.$$

If so, we say $\text{plim}_{T \rightarrow \infty} \hat{\theta}_T = c$ or $\hat{\theta}_T \rightarrow_p c$.

EX 1: $\hat{\theta}_T = 0$ with $\text{pr} = 1 - (1/T)$; $= 1$ with $\text{pr} = 1/T$.

Choose $0 < \epsilon < 1$:

$$\begin{aligned} \Pr(|\hat{\theta}_T - 0| > \epsilon) &= \Pr(|\hat{\theta}_T| > \epsilon) = \Pr(\hat{\theta}_T > \epsilon) = 1/T \\ &\Rightarrow \lim_{T \rightarrow \infty} \Pr(|\hat{\theta}_T - 0| > \epsilon) = 0 \\ &\Rightarrow \hat{\theta}_T \rightarrow_p 0. \end{aligned}$$

EX 2: $\hat{\theta}_T = 0$ with $\text{pr} = 1 - (1/T)$; $= T$ with $\text{pr} = 1/T$.

Choose $0 < \epsilon < 1$:

$$\begin{aligned} \Pr(|\hat{\theta}_T - 0| > \epsilon) &= \Pr(|\hat{\theta}_T| > \epsilon) = \Pr(\hat{\theta}_T > \epsilon) = 1/T \\ &\Rightarrow \lim_{T \rightarrow \infty} \Pr(|\hat{\theta}_T - 0| > \epsilon) = 0 \\ &\Rightarrow \hat{\theta}_T \rightarrow_p 0. \end{aligned}$$

Digression to other stronger convergency:

Definition: (Convergence in mean square)

$\hat{\theta}_T$ converges in mean square to c iff $\lim_{T \rightarrow \infty} E[|\hat{\theta}_T - c|^2] = 0$. For this case, we say

$$\hat{\theta}_T \rightarrow c, \text{ m.s..}$$

Theorem: $\text{m.s.} \Rightarrow \text{p.}$

Proof:

Chebychev's inequality (see Greene) says:

$$\begin{aligned} \text{For any } \epsilon > 0, \Pr(|\hat{\theta}_T - c| > \epsilon) &\leq E(|\hat{\theta}_T - c|^2)/\epsilon^2. \\ \Rightarrow \lim_{T \rightarrow \infty} \Pr(|\hat{\theta}_T - c| > \epsilon) &= \lim_{T \rightarrow \infty} E(|\hat{\theta}_T - c|^2)/\epsilon^2 = 0. \end{aligned}$$

Fact: p. does not necessarily imply m.s.

EX 1: $\hat{\theta}_T = 0$ with $\text{pr} = 1 - (1/T)$; $= 1$ with $\text{pr} = 1/T$: $\hat{\theta}_T \rightarrow_p 0$.

- Observe $E[|\hat{\theta}_T - 0|^2] = E[\hat{\theta}_T^2] = 0^2 \times [1 - (1/T)] + 1^2 \times (1/T) = 1/T$
 $\Rightarrow \lim_{T \rightarrow \infty} E[|\hat{\theta}_T - 0|^2] = 0.$
 $\Rightarrow \hat{\theta}_T \rightarrow 0$ m.s..

EX 2: $\hat{\theta}_T = 0$ with $\text{pr} = 1 - (1/T)$; $= T$ with $\text{pr} = 1/T$.

- $\hat{\theta}_T \rightarrow_p 0$.
- Observe $E[|\hat{\theta}_T - 0|^2] = E[\hat{\theta}_T^2] = 0^2 \times [1 - (1/T)] + T^2 \times (1/T) = T$
 $\Rightarrow \lim_{T \rightarrow \infty} E[|\hat{\theta}_T - 0|^2] = \infty.$
 \Rightarrow not m.s.

Implication:

- In EX 1 above, $\hat{\theta}_T$ is p and ms. But in EX 2 above, $\hat{\theta}_T$ is p., but not m.s.
- To be p., $\Pr(\hat{\theta}_T \text{ deviates from } c)$ should become increasingly small as $T \rightarrow \infty$. But this is not enough for m.s.. To be m.s., for any possible value of $\hat{\theta}_T$, the size of $|\hat{\theta}_T - c|$ should not grow too fast as $T \rightarrow \infty$. For example, if we assume $\Pr(\hat{\theta}_T = T^{1/4}) = 1/T$ instead, we can show that $\hat{\theta}_T \rightarrow c$, m.s.

Definition: (Almost sure convergency, Strong Convergency)

$\hat{\theta}_T$ converges *almost surely* to c , iff $\Pr[\lim_{T \rightarrow \infty} \hat{\theta}_T = c] = 1$. For this case, we say:

$$\hat{\theta}_T \rightarrow c, \text{ a.s..}$$

Theorem: a.s. \Rightarrow p. (See Rao (1973).)

Fact: 1) p. does not implies a.s.

2) No clear relation between a.s. and m.s. with few exceptions.

Theorem:

Suppose $\lim_{T \rightarrow \infty} E(|\hat{\theta}_T - c|^2) = 0$ and $\sum_{T=1}^{\infty} E(|\hat{\theta}_T - c|^2) < \infty$. Then, $\hat{\theta}_T \rightarrow c$, a.s.. (See Rao (1973).)

EX 1: $\hat{\theta}_T = 0$ with pr = $1 - (1/T)$; = 1 with pr = $1/T$.

- $\hat{\theta}_T \rightarrow_p 0$ and $\hat{\theta}_T \rightarrow 0$, m.s..
- But, can't determine whether $\hat{\theta}_T \rightarrow 0$, a.s..
(Observe that $\sum_{T=1}^{\infty} E(|\hat{\theta}_T - c|^2) = \sum_{T=1}^{\infty} (1/T) = \infty$.)

EX 2: $\hat{\theta}_T = 0$ with pr = $1 - (1/T^2)$; = 1 with pr = $1/T^2$.

- $\hat{\theta}_T \rightarrow_p 0$.
- Observe $E[|\hat{\theta}_T - 0|^2] = E[\hat{\theta}_T^2] = 0^2 \times [1 - (1/T^2)] + 1^2 \times (1/T^2) = 1/T^2$:
 - $\lim_{T \rightarrow \infty} E[|\hat{\theta}_T - 0|^2] = 0$.
 $\Rightarrow \sum_{T=1}^{\infty} E(|\hat{\theta}_T - c|^2) = \sum_{T=1}^{\infty} (1/T^2) < \infty$
 $\Rightarrow \hat{\theta}_T \rightarrow 0$, a.s..

Implication:

EX 1: $\Pr(\hat{\theta}_T = 1) = 1/T$.

EX 2: $\Pr(\hat{\theta}_T = 1) = 1/T^2$.

\Rightarrow To be a.s., $\Pr(\hat{\theta}_T$ deviates from c) should decrease rapidly as $T \rightarrow \infty$.

End of Digression

Definition:

$\hat{\theta}_T$: an estimator of θ_o .

We say that $\hat{\theta}_T$ is consistent, iff $\text{plim}_{T \rightarrow \infty} \hat{\theta}_T = \theta_o$.

Question:

An example for a consistent estimator?

Theorem: (Generalized Weak Law of Large Numbers, GWLLN)

$\{y_1, \dots, y_T\}$: a sequence of $p \times 1$ random vectors.

For any t , $E(y_t)$ and $\text{Cov}(y_t)$ are well defined and finite.

$\bar{y}_T = (1/T) \sum_{t=1}^T y_t$ (mean of the sequence).

Under some suitable conditions (acceptable for Econometrics I, II),

$$\bar{y}_T = (1/T)\sum_{t=1}^T y_t \rightarrow_p \lim_{T \rightarrow \infty} (1/T)\sum_{t=1}^T E(y_t).$$

Note:

- 1) Both $E(y_t)$ and $\text{Cov}(y_t)$ [$\text{var}(y_t)$ if y_t is a scalar] could differ across different t .
- 2) The y_t could be correlated as long as $\lim_{n \rightarrow \infty} \text{cov}(y_t, y_{t+n}) = 0$.

EX: $\{x_1, \dots, x_T\}$: RS from a population with $E(x) = \mu_o$ and $\text{var}(x) = \sigma_o^2$.

- By Kolmogorov's SLLN, $\bar{x} = (1/T)\sum_{t=1}^T x_t \rightarrow \mu_o$, a.s..
- $\bar{x} \rightarrow_p \mu_o$.

[Proof by GWLLN]

$$\begin{aligned} (1/T)\sum_{t=1}^T E(x_t) &= (1/T)\sum_{t=1}^T \mu_o = (1/T)T\mu_o = \mu_o \\ &\rightarrow \lim_{T \rightarrow \infty} (1/T)\sum_{t=1}^T E(x_t) = \lim_{T \rightarrow \infty} \mu_o = \mu_o \\ &\rightarrow \text{By GWLLN, } \bar{x} \rightarrow_p \mu_o. \end{aligned}$$

Theorem: (Slutzky)

$$\text{plim}_{T \rightarrow \infty} \hat{\theta}_T = \theta_o.$$

$g(\theta)$: a vector of continuous functions of θ .

$$\Rightarrow \text{plim}_{T \rightarrow \infty} g(\hat{\theta}_T) = g(\theta_o)$$

EX: θ is a scalar and $\hat{\theta}_T \rightarrow_p \theta_o$.

$$\text{plim}_{T \rightarrow \infty} \hat{\theta}_T^2 = \theta_o^2; \text{plim}_{T \rightarrow \infty} 1/\hat{\theta}_T = 1/\theta_o.$$

EX: $\{x_1, \dots, x_T\}$: Random sample from a population with μ_o and σ_o^2 .

$$\text{plim } \bar{x}/s_x^2 = [\text{plim } \bar{x}]/[\text{plim } s_x^2] = \mu_o/\sigma_o^2.$$

EX: $\text{plim } (\bar{x} + \bar{x}^2 + \bar{x}s_x^2 + s_x^2) = \mu_o + \mu_o^2 + \mu_o\sigma_o^2 + \sigma_o^2$.

Rules for Probability limits:

- 1) W_T is a square matrix of random variables and $\text{plim}W_T$ is invertible. Then,

$$\text{plim } [W_T]^{-1} = [\text{plim } W_T]^{-1}.$$

- 2) X_T and Y_T are conformable matrices of random variables Then,

$$\text{plim } X_T Y_T = [\text{plim } X_T][\text{plim } Y_T].$$

(4) Convergency in distribution

Definition: (Convergency in distribution)

$F(z)$: cdf of a random vector z .

z_T : a random vector with cdf $F_T(z_T)$.

\Rightarrow We say z_T converges in distribution to z , iff $\lim_{T \rightarrow \infty} F_T(z) = F(z)$ for a.

$\Rightarrow z_T \rightarrow_d z$.

Fact: d . differs from p .

EX: Two dice A and B.

A is fair one: $f(z) = 1/6, z = 1, 2, \dots, 6$.

B is unfair:

z_T be a possible outcome from the T 'th trial with

$$f_T(z_T) = 1/6 + 1/(T+100) \text{ for } z_T = 1, 2, 3,$$

$$f_T(z_T) = 1/6 - 1/(T+100) \text{ for } z_T = 4, 5, 6.$$

As $T \rightarrow \infty$, the unfairness of B decreases.

$\Rightarrow z_T \rightarrow_d z$.

But a realized value of z_T may not equal that of x at T 'th trial, even if $T \rightarrow \infty$.

Theorem: (Mann and Wald)

Suppose $g(z)$ is a continuous function. Then,

$$(z_T \rightarrow_d z) \Rightarrow (g(z_T) \rightarrow_d g(z)).$$

Theorem:

A_T : a random matrix with $\text{plim } A_T = A$.

z_T : a random vector $\rightarrow_d z$.

$$\Rightarrow A_T z_T \rightarrow_d A z.$$

EX: (Central Limit Theorem, CLT)

$\{x_1, \dots, x_T\}$: RS from a population with μ_0 and σ_0^2 .

\Rightarrow Lindberg-Levy CLT says

$$\sqrt{T}(\bar{x} - \mu_0) \rightarrow_d N(0, \sigma_0^2) .$$

Theorem: (Generalized CLT, GCLT)

$\{y_1, \dots, y_T\}$: a sequence of $p \times 1$ random vectors.

For any t , $E(y_t) = 0$ and $\text{Cov}(y_t)$ is well defined and finite.

Under some suitable conditions (acceptable for Econometrics I, II),

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T y_t \rightarrow_d N\left(0, \lim_{T \rightarrow \infty} \frac{1}{T} \text{Cov}\left(\sum_{t=1}^T y_t\right)\right) .$$

Note:

- 1) $\text{Cov}(y_t)$ [$\text{var}(y_t)$ if y_t is a scalar] could differ across different t .
- 2) The y_t could be correlated as long as $\lim_{n \rightarrow \infty} \text{cov}(y_t, y_{t+n}) = 0$.

EX: (Lindberg-Levy CLT)

$\{x_1, \dots, x_T\}$: RS from a population with μ_0 and σ_0^2 .

\Rightarrow Let $y_t = x_t - \mu_0$.

$$E(y_t) = E(x_t) - \mu_0 = 0;$$

$$\text{var}(y_t) = \text{var}(x_t) = \sigma_0^2.$$

$$[1/\sqrt{T}] \sum_t y_t = [1/\sqrt{T}] [\sum_t x_t - T\mu_0] = \sqrt{T}(\bar{x} - \mu_0)$$

$$(1/T) \text{var}(\sum_t y_t) = (1/T) \text{var}(\sum_t x_t - T\mu) = (1/T) \text{var}(\sum_t x_t)$$

$$= (1/T) \sum_t \text{var}(x_t) = (1/T) T \sigma_0^2 = \sigma_0^2$$

$$\Rightarrow \lim (1/T) \text{var}(\sum_t y_t) = \sigma_0^2.$$

$$\Rightarrow \sqrt{T}(\bar{x} - \mu_0) \rightarrow_d N(0, \sigma_0^2).$$

Corollary:

Assume the same things as GCLT.

Assume that the y_t 's are linearly uncorrelated.

Then,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T y_t \rightarrow_d N\left(0, \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \text{Cov}(y_t)\right) .$$

Proof:

When y_t is a scalar, $\text{var}(\sum_t y_t) = \sum_t \text{var}(y_t)$.

Lemma:

Let $E(y_t | y_{t-1}, y_{t-2}, \dots, y_1) = 0$. [Martingale Difference Sequence]

Then, the y_t 's are linearly uncorrelated.

Proof: [Assume y_t is a scalar.]

Consider the case in which y_t is a scalar.

\Rightarrow By the law of iterative expectation, $E(y_t) = 0$.

\Rightarrow By the law of iterative expectation,

$$E(y_{t+j} | y_t, y_{t-1}, \dots, y_1) = E_{y_{t+1}, \dots, y_{t+j-1}} [E(y_{t+j} | y_{t+j-1}, \dots, y_1)] = E_{y_{t+1}, \dots, y_{t+j-1}} (0) = 0 .$$

$\Rightarrow \text{cov}(y_t, y_{t+j}) = E[(y_t - E(y_t))(y_{t+j} - E(y_{t+j}))] = E(y_t y_{t+j})$

$$= E_{y_t} [E(y_t y_{t+j} | y_t)] = E_{y_t} [y_t E(y_{t+j} | y_t)] = E_{y_t} (0) = 0 .$$

Theorem: (GCLT for martingale difference sequences)

$\{y_1, \dots, y_T\}$: a sequence of $p \times 1$ random vectors.

$E(y_t | y_{t-1}, \dots, y_1) = 0$.

$\text{Cov}(y_t)$ is well defined and finite.

Under some suitable conditions (acceptable for Econometrics I, II),

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T y_t \rightarrow_d N(0, \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \text{Cov}(y_t)) .$$

[10] Large-Sample Properties of MLE

A Short Digression to Matrix Algebra

Definition:

1) $g(\theta) = g(\theta_1, \dots, \theta_p)$: a scalar function of θ .

$$g_j = \partial g / \partial \theta_j .$$

$$\frac{\partial g(\theta)}{\partial \theta} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_p \end{bmatrix} ; \frac{\partial g(\theta)}{\partial \theta'} = [g_1, g_2, \dots, g_p] ,$$

2) $w(\theta)$: a $m \times 1$ vector:

$$\Rightarrow w_{ij} = \partial w_i(\theta) / \partial \theta_j.$$

$$\frac{\partial w(\theta)}{\partial \theta'} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mp} \end{bmatrix}_{m \times p}$$

3) $g(\theta)$: a scalar function of θ

$$\Rightarrow \text{where } g_{ij} = \partial^2 g(\theta) / \partial \theta_i \partial \theta_j.$$

$$\frac{\partial^2 g(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1p} \\ g_{21} & g_{22} & \cdots & g_{2p} \\ \vdots & \vdots & & \vdots \\ g_{p1} & g_{p2} & \cdots & g_{pp} \end{bmatrix}_{p \times p}$$

\Rightarrow Called Hessian matrix of $g(\theta)$.

EX:

Let $g(\theta) = \theta_1^2 + \theta_2^2 + \theta_1 \theta_2$. Find $\partial g(\theta) / \partial \theta$.

$$\rightarrow (2\theta_1 + \theta_2, 2\theta_2 + \theta_1)'$$

EX:

$$\text{Let } w(\theta) = \begin{bmatrix} \theta_1^2 + \theta_2 \\ \theta_1 + \theta_2^2 \end{bmatrix}. \text{ Then, } \partial w(\theta) / \partial \theta' = \begin{bmatrix} 2\theta_1 & 1 \\ 1 & 2\theta_2 \end{bmatrix}.$$

EX:

Let $g(\theta) = \theta_1^2 + \theta_2^2 + \theta_1 \theta_2$. Find the Hessian matrix of $g(\theta)$.

$$\rightarrow \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Some useful results:

1) c' : $1 \times p$, θ : $p \times 1$ ($c'\theta$ is a scalar)

$$\Rightarrow \partial(c'\theta)/\partial\theta = c ; \partial(c'\theta)/\partial\theta' = c'.$$

2) $R: m \times p, \theta: p \times 1$ ($R\theta$ is $m \times 1$)

$$\Rightarrow \partial(R\theta)/\partial\theta = R$$

3) $A: p \times p$ symmetric, $\theta: p \times 1$ ($\theta' A \theta$)

$$\Rightarrow \partial(\theta' A \theta)/\partial\theta = 2A\theta.$$

$$\Rightarrow \partial(\theta' A \theta)/\partial\theta' = 2\theta' A$$

$$\Rightarrow \partial(\theta' A \theta)/\partial\theta\partial\theta' = 2A.$$

End of Digression

Definition: (Hessian matrix of log-likelihood function)

$$H_T(\theta) = \left[\frac{\partial^2 l_T(\theta)}{\partial\theta\partial\theta'} \right]; (i,j)\text{th ele. in } H_T = \left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right],$$

Definition: (Information matrix)

$$I_T(\theta_o) = E[-H_T(\theta_o)].$$

Note: To compute $I_T(\theta_o)$, compute $H_T(\theta)$ first, then, $H_T(\theta_o)$, and then, $E(-H_T(\theta_o))$.

Theorem:

Let $\hat{\theta}$ be MLE. Then, under suitable regularity conditions,

$\hat{\theta}$ is consistent, and

$$\sqrt{T}(\hat{\theta} - \theta_o) \rightarrow_d N(0, \lim[(1/T)I_T(\theta_o)]^{-1}).$$

Further, $\hat{\theta}$ is asymptotically efficient.

Implication:

$$\hat{\theta} \approx N(\theta_o, [I_T(\theta_o)]^{-1}) \Rightarrow \hat{\theta} \approx N(\theta_o, [I_T(\hat{\theta})]^{-1}).$$

EX:

$\{x_1, \dots, x_T\}$ iid with $N(\mu_o, \sigma_o^2)$.

$\theta = [\mu, v]'$ and $v = \sigma^2$.

$$l_T = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln v - \frac{1}{2v}\sum_t(x_t - \mu)^2.$$

The first derivatives:

$$\frac{\partial l_T}{\partial \mu} = \frac{\sum_t(x_t - \mu)}{v}; \quad \frac{\partial l_T}{\partial v} = -\frac{T}{2v} + \frac{1}{2v^2}\sum_t(x_t - \mu)^2.$$

The second derivatives:

$$\frac{\partial^2 l_T(\theta)}{\partial \mu \partial \mu} = \frac{1}{v}\sum_t(-1) = -\frac{T}{v} \rightarrow \frac{\partial^2 l_T(\theta_o)}{\partial \mu \partial \mu} = -\frac{T}{v_o} \rightarrow \mathbb{E}\left[-\frac{\partial^2 l_T(\theta_o)}{\partial \mu \partial \mu}\right] = \frac{T}{v_o}.$$

$$\frac{\partial^2 l_T(\theta)}{\partial \mu \partial v} = -\frac{\sum_t(x_t - \mu)}{v^2} \rightarrow \frac{\partial^2 l_T(\theta_o)}{\partial \mu \partial v} = -\frac{\sum_t(x_t - \mu_o)}{v_o^2}.$$

$$\rightarrow \mathbb{E}\left[-\frac{\partial^2 l_T(\theta_o)}{\partial \mu \partial v}\right] = \mathbb{E}\left[\frac{\sum_t(x_t - \mu_o)}{v_o^2}\right] = \frac{1}{v_o^2}\mathbb{E}[\sum_t(x_t - \mu_o)] = \frac{1}{v_o^2}\sum_t[\mathbb{E}(x_t) - \mu_o] = 0.$$

$$\frac{\partial^2 l_T(\theta)}{\partial v \partial v} = \frac{T}{2v^2} + \frac{0 \times 2v^2 - 1 \times 4v}{(2v^2)^2}\sum_t(x_t - \mu)^2 = \frac{T}{2v^2} - \frac{1}{v^3}\sum_t(x_t - \mu)^2$$

$$\rightarrow \frac{\partial^2 l_T(\theta_o)}{\partial v \partial v} = \frac{T}{2v_o^2} - \frac{1}{v_o^3}\sum_t(x_t - \mu_o)^2.$$

$$\begin{aligned} \rightarrow \mathbb{E}\left[-\frac{\partial^2 l_T(\theta_o)}{\partial v \partial v}\right] &= \mathbb{E}\left[-\frac{T}{2v_o^2} + \frac{1}{v_o^3}\sum_t(x_t - \mu_o)^2\right] \\ &= -\frac{T}{2v_o^2} + \frac{1}{v_o^3}\sum_t\mathbb{E}[(x_t - \mu_o)^2] = -\frac{T}{2v_o^2} + \frac{1}{v_o^3}\sum_tv_o = -\frac{T}{2v_o^2} + \frac{Tv_o}{v_o^3} = \frac{T}{2v_o^2}. \end{aligned}$$

Therefore,

$$I_T(\theta_o) = \begin{bmatrix} \frac{T}{\sigma_o^2} & 0 \\ 0 & \frac{T}{2\sigma_o^4} \end{bmatrix}; \quad [I_T(\theta_o)]^{-1} = \begin{bmatrix} \frac{\sigma_o^2}{T} & 0 \\ 0 & \frac{2\sigma_o^4}{T} \end{bmatrix}.$$

Hence,

$$\hat{\theta} = \begin{bmatrix} \hat{\mu}_{ML} \\ \hat{\sigma}_{ML}^2 \end{bmatrix} \approx N \left(\begin{bmatrix} \mu_o \\ \sigma_o^2 \end{bmatrix}, \begin{bmatrix} \frac{\hat{\sigma}_{ML}^2}{T} & 0 \\ 0 & \frac{2(\hat{\sigma}_{ML}^2)^2}{T} \end{bmatrix} \right).$$

[Sketchical Technical Notes For MLE]

Definition:

For any function $g(x, \theta)$ where x is a random variable (or vector) with probability density $f(x, \theta_o)$,

$$E(g(x, \theta)) \equiv \int_{\Omega} g(x, \theta) f(x, \theta_o) dx \text{ (true expected value of } g(x, \theta) \text{);}$$

$$E_{\theta}(g(x, \theta)) \equiv \int_{\Omega} g(x, \theta) f(x, \theta) dx \text{ (expected value of } g(x, \theta) \text{ assuming } f(x, \theta)),$$

where Ω denote the range of x .

Assumption 1:

(i) Let x is a random (vector or scalar) variable with pdf of a form $f(x, \theta)$, where θ is a $p \times 1$ vector of unknown parameters. Let θ_o be the true value of θ . Then, θ_o uniquely maximizes $E[\ln f(x, \theta)]$.

That is, $E[\ln f(x, \theta_o)] > E[\ln f(x, \theta)]$ for any $\theta \neq \theta_o$.

(ii) $\{x_1, \dots, x_T\}$ is a random sample from a population satisfying (i).

Assumption 2:

The range of x does not depend on θ .

Lemma 1:

Define $s(x, \theta) = \partial \ln f(x, \theta) / \partial \theta$. Then, under Assumption 2, $E_{\theta}(s(x, \theta)) = 0$, for all θ .

<Proof>

Since $f(x, \theta)$ is a probability density function, $1 = \int_{\Omega} f(x, \theta) dx$ for any θ . Differentiate both side of this equation with respect to θ . Then, we have:

$$\begin{aligned} 0 &= \frac{\partial \int_{\Omega} f(x, \theta) dx}{\partial \theta} = \int_{\Omega} \frac{\partial f(x, \theta)}{\partial \theta} dx = \int_{\Omega} \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) dx = \int_{\Omega} s(x, \theta) f(x, \theta) dx \\ &= E_{\theta}(s(x, \theta)), \end{aligned}$$

where Assumption 2 warrants the first equality, and the second equality results from the fact that $\partial \ln f(x, \theta) / \partial \theta = [\partial f(x, \theta) / \partial \theta] / f(x, \theta)$.

Corollary 1:

Under Assumption 2, $E(s(x, \theta_0)) = 0$.

Lemma 2:

Under Assumption 2,

$$E_{\theta}[s(x, \theta)s(x, \theta)'] = E_{\theta}\left[-\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta'}\right],$$

for all θ .

<Proof>

For simplicity, we only consider the cases where θ is a scalar. Lemma 1 implies:

$$\int_{\Omega} \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) dx = 0.$$

Differentiate both sides of this equation:

$$\begin{aligned} & \int_{\Omega} \left[\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta} f(x, \theta) + \frac{\partial \ln f(x, \theta)}{\partial \theta} \frac{\partial f(x, \theta)}{\partial \theta} \right] dx = 0 \\ \rightarrow & \int_{\Omega} \left[\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta} f(x, \theta) + \frac{\partial \ln f(x, \theta)}{\partial \theta} \frac{\partial \ln f(x, \theta)}{\partial \theta} f(x, \theta) \right] dx = 0 \\ \rightarrow & E_{\theta} \left[\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta} + \frac{\partial \ln f(x, \theta)}{\partial \theta} \frac{\partial \ln f(x, \theta)}{\partial \theta} \right] = 0, \text{ for any } \theta \end{aligned}$$

Corollary 2:

Under Assumption 2,

$$E[s(x, \theta_0)s(x, \theta_0)'] = E\left[-\frac{\partial^2 \ln f(x, \theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_0}\right].$$

EX:

- $f(x, \theta) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-\mu)^2}{2v}\right]; \theta = \begin{bmatrix} \mu \\ v \end{bmatrix}; \theta_0 = \begin{bmatrix} \mu_0 \\ v_0 \end{bmatrix}$
- Assumption 1 holds?
- $\ln f(x, \theta) = (-1/2)\ln(2\pi) - (1/2)\ln(v) - (x-\mu)^2/(2v) = (-1/2)\ln(2\pi) - (1/2)\ln(v) - [(x-\mu_0)-(\mu_0-\mu)]^2/(2v)$

$$= (-1/2)\ln(2\pi) - (1/2)\ln(v) - (x-\mu_o)^2/(2v) - 2(\mu_o-\mu)(x-\mu_o)/(2v) - (\mu-\mu_o)^2/(2v).$$

$$E[\ln f(x,\theta)] = (-1/2)\ln(2\pi) - (1/2)\ln(v) - v_o/(2v) - (\mu-\mu_o)^2/(2v).$$

⇒ Clearly, $E[\ln f(x,\theta)]$ is maximized at $\mu = \mu_o$.

⇒ Also, $E[\ln f(x,\theta)]$ is maximized at $v = v_o$, by FOC: $\partial E[\ln f(x,\theta)]/\partial v = - (1/2v) + v_o/(2v^2)$

$$= 0 \Rightarrow v = v_o.$$

- Assumption 2 holds?
 - Yes, since $-\infty < x < \infty$.

Theorem 1:

Under Assumption 1, the MLE $\hat{\theta}$ is consistent under some suitable assumptions. [See Amemiya.]

<An Intuition>

Observe that $T^{-1}l_T(\theta) = T^{-1}\sum_i \ln f(x_i, \theta)$. Since $\{x_1, \dots, x_T\}$ is a random sample, we can regard $\{\ln f(x_1, \theta), \dots, \ln f(x_T, \theta)\}$ as a random sample from a population of the random variable $\ln f(x, \theta)$.

Then, by LLN, $T^{-1}l_T(\theta) \rightarrow_p E[\ln f(x, \theta)]$. But Assumption 1 implies that θ_o uniquely maximizes $E[\ln f(x, \theta)] = \text{plim } T^{-1}l_T(\theta_o)$. That is, θ_o maximizes $\text{plim } T^{-1}l_T(\theta)$. Note that MLE $\hat{\theta}$ maximizes $T^{-1}l_T(\theta)$. But, when sample size T is large, searching for the maximizer $\hat{\theta}$ is similar to searching for θ_o . This provides an intuition for the consistency of MLE.

Lemma 3:

Define $s_i(\theta) = s(x_i, \theta) = \partial \ln f(x_i, \theta) / \partial \theta$. Under Assumptions 1-2 and other suitable assumptions,

$$\frac{1}{\sqrt{T}} \frac{\partial l_T(\theta)}{\partial \theta} \Big|_{\theta=\theta_o} \rightarrow_d N(0, \lim \frac{1}{T} \text{Cov}[\sum_i s_i(\theta_o)]).$$

<Proof>

Note that:

$$\frac{1}{\sqrt{T}} \frac{\partial l_T(\theta)}{\partial \theta} \Big|_{\theta=\theta_o} = \frac{1}{\sqrt{T}} \sum_i \frac{\partial \ln f(x_i, \theta)}{\partial \theta} \Big|_{\theta=\theta_o} = \frac{1}{\sqrt{T}} \sum_i s_i(\theta_o).$$

By Lemma 1, $E[s_i(\theta_o)] = 0$. Thus, by GWCLT, we obtain the desired result.

Lemma 4:

Under Assumptions 1-2 and other suitable assumptions,

$$-\frac{1}{T} \frac{\partial^2 l_T(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_o} \rightarrow_p \lim \frac{1}{T} I_T(\theta_o).$$

<proof>

$$\text{Note that } -\frac{1}{T} \frac{\partial^2 l_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = -\frac{1}{T} \sum_t \frac{\partial^2 \ln f(x_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o}.$$

Then, by GWLLN,

$$\begin{aligned} -\frac{1}{T} \frac{\partial^2 l_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} &= -\frac{1}{T} \sum_t \frac{\partial^2 \ln f(x_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \\ &\rightarrow_p \lim \frac{1}{T} E \left[-\sum_t \frac{\partial^2 \ln f(x_t, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right] = \lim \frac{1}{T} E \left[-\frac{\partial^2 l_T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right] = \frac{1}{T} I_T(\boldsymbol{\theta}_o). \end{aligned}$$

Lemma 5:

$$\text{Under Assumptions 1-2, } \text{Cov}[\sum_t s_t(\boldsymbol{\theta}_o)] = I_T(\boldsymbol{\theta}_o).$$

<proof>

Since $\{s_1(\boldsymbol{\theta}_o), \dots, s_T(\boldsymbol{\theta}_o)\}$ is a RS,

$$\text{Cov}[\sum_t s_t(\boldsymbol{\theta}_o)] = \sum_t \text{Cov}[s_t(\boldsymbol{\theta}_o)] = \sum_t E[s_t(\boldsymbol{\theta}_o) s_t(\boldsymbol{\theta}_o)'].$$

where the last equality results from Lemma 1. Note also that

$$I_T(\boldsymbol{\theta}_o) = \sum_t E \left[-\frac{\partial^2 \ln f(x_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right].$$

Thus, it is enough to show that

$$E[s_t(\boldsymbol{\theta}_o) s_t(\boldsymbol{\theta}_o)'] = E \left[-\frac{\partial^2 \ln f(x_t, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} \right].$$

But this equality holds by Lemma 2.

Corollary 3:

Under Assumptions 1-2 and other suitable assumptions,

$$\frac{1}{\sqrt{T}} \frac{\partial l_T(\boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \rightarrow_d N(0, \lim \frac{1}{T} I_T(\boldsymbol{\theta}_o)).$$

Theorem 2:

Let $\hat{\boldsymbol{\theta}}$ be MLE. Under Assumptions 1-2 and other suitable assumptions,

$$\sqrt{T}(\hat{\theta} - \theta_o) \rightarrow_d N(0, \lim[\frac{1}{T}I_T(\theta_o)]^{-1}).$$

<Proof>

Consider the first order condition for MLE:

$$\frac{\partial l_T(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = 0.$$

Use Taylor's expansion around θ_o :

$$\frac{\partial l_T(\theta)}{\partial \theta} \Big|_{\theta=\theta_o} + \frac{\partial^2 l_T(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\bar{\theta}} (\hat{\theta} - \theta_o) = 0,$$

where $\bar{\theta}$ is a vector between $\hat{\theta}$ and θ_o . Since $\hat{\theta}$ is consistent and $\bar{\theta}$ is between $\hat{\theta}$ and θ_o , $\bar{\theta}$ is also consistent. That is,

$$\frac{1}{\sqrt{T}} \frac{\partial l_T(\theta)}{\partial \theta} \Big|_{\theta=\theta_o} + \frac{1}{T} \frac{\partial^2 l_T(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_o} \sqrt{T}(\hat{\theta} - \theta_o) = o_p(1),$$

where $o_p(1)$ means "a term asymptotically negligible". Thus, we have:

$$\begin{aligned} \sqrt{T}(\hat{\theta} - \theta_o) &= \left[-\frac{1}{T} \frac{\partial^2 l_T(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_o} \right]^{-1} \frac{1}{\sqrt{T}} \frac{\partial l_T(\theta)}{\partial \theta} \Big|_{\theta=\theta_o} + o_p(1) \\ &\rightarrow \sqrt{T}(\hat{\theta} - \theta_o) = \left[\lim \frac{1}{T} I_T(\theta_o) \right]^{-1} \frac{1}{\sqrt{T}} \frac{\partial l_T(\theta_o)}{\partial \theta} + o_p(1) \quad (\text{By Lemma 4}) \\ &\rightarrow \sqrt{T}(\hat{\theta} - \theta_o) \rightarrow_d N(0, \left[\lim \frac{1}{T} I_T(\theta_o) \right]^{-1} \lim \frac{1}{T} I_T(\theta_o) \left[\lim \frac{1}{T} I_T(\theta_o) \right]^{-1}) \\ &= N(0, \left[\lim \frac{1}{T} I_T(\theta_o) \right]^{-1}). \quad (\text{By Corollary 3}) \end{aligned}$$

[11] Testing Hypotheses Based on MLE

Let $w(\theta) = [w_1(\theta), w_2(\theta), \dots, w_m(\theta)]'$, where $w_j(\theta) = w_j(\theta_1, \theta_2, \dots, \theta_p) = a^j$ of $\theta_1, \dots, \theta_p$.

General form of hypotheses:

H_o : The true θ (θ_o) satisfy the m restrictions, $w(\theta) = 0_{m \times 1}$ ($m \leq p$).

Examples:

1) θ : a scalar

$$H_0: \theta = 2 \rightarrow H_0: \theta - 2 = 0 \rightarrow H_0: w(\theta) = 0, \text{ where } w(\theta) = \theta - 2.$$

2) $\theta = [\theta_1, \theta_2, \theta_3]'$.

$$H_0: \theta_1^2 = \theta_2 + 2 \text{ and } \theta_3 = \theta_1 + \theta_2$$

$$\rightarrow H_0: \theta_1^2 - \theta_2 - 2 = 0 \text{ and } \theta_3 - \theta_1 - \theta_2 = 0.$$

$$\rightarrow \text{Let } w_1(\theta) = \theta_1^2 - \theta_2 - 2 \text{ and } w_2(\theta) = \theta_3 - \theta_1 - \theta_2.$$

$$\rightarrow H_0: w(\theta) = \begin{bmatrix} w_1(\theta) \\ w_2(\theta) \end{bmatrix} = \begin{bmatrix} \theta_1^2 - \theta_2 - 2 \\ \theta_3 - \theta_1 - \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

3) linear restrictions

$$\theta = [\theta_1, \theta_2, \theta_3]'$$

$$H_0: \theta_1 = \theta_2 + 2 \text{ and } \theta_3 = \theta_1 + \theta_2$$

$$\rightarrow H_0: \theta_1 - \theta_2 - 2 = 0 \text{ and } \theta_3 - \theta_1 - \theta_2 = 0$$

$$\rightarrow H_0: w(\theta) = \begin{bmatrix} w_1(\theta) \\ w_2(\theta) \end{bmatrix} = \begin{bmatrix} \theta_1 - \theta_2 - 2 \\ \theta_3 - \theta_1 - \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$$\rightarrow w(\theta) = \begin{bmatrix} 1 & -1 & 0 \\ -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} - \begin{bmatrix} 2 \\ 0 \end{bmatrix} = R\theta - r.$$

Remark:

If all restrictions are linear in θ , H_0 takes the following form:

$$H_0: R\theta - r = 0_{m \times 1},$$

where R and r are known $m \times p$ and $m \times 1$ matrices, respectively.

Definition:

$$W(\theta) = \frac{\partial w(\theta)}{\partial \theta'} = \begin{bmatrix} \frac{\partial w_1(\theta)}{\partial \theta_1} & \frac{\partial w_1(\theta)}{\partial \theta_2} & \dots & \frac{\partial w_1(\theta)}{\partial \theta_p} \\ \frac{\partial w_2(\theta)}{\partial \theta_1} & \frac{\partial w_2(\theta)}{\partial \theta_2} & \dots & \frac{\partial w_2(\theta)}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial w_m(\theta)}{\partial \theta_1} & \frac{\partial w_m(\theta)}{\partial \theta_2} & \dots & \frac{\partial w_m(\theta)}{\partial \theta_p} \end{bmatrix}_{m \times p}$$

Example:

$$\text{Let } \theta = [\theta_1, \theta_2, \theta_3]'$$

$$H_0: \theta_1^2 - \theta_2 = 0 \text{ and } \theta_1 - \theta_2 - \theta_3^2 = 0.$$

$$\rightarrow w(\theta) = \begin{bmatrix} \theta_1^2 - \theta_2 \\ \theta_1 - \theta_2 - \theta_3^2 \end{bmatrix} \rightarrow W(\theta) = \begin{bmatrix} 2\theta_1 & -1 & 0 \\ 1 & -1 & -2\theta_3 \end{bmatrix}_{2 \times 3}$$

Example:

$$\theta = [\theta_1, \theta_2, \theta_3]'$$

$$H_0: \theta_1 = 0 \text{ and } \theta_2 + \theta_3 = 1.$$

$$\rightarrow w(\theta) = \begin{bmatrix} \theta_1 \\ \theta_2 + \theta_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0 \rightarrow w(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0$$

$$\rightarrow R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}; r = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

$$\rightarrow w(\theta) = R\theta - r.$$

$$\rightarrow W(\theta) = R.$$

Definition: (Restricted MLE)

Let $\tilde{\theta}$ be the restricted ML estimator which maximizes

$$l_T(\theta) \text{ s.t. } w(\theta) = 0.$$

Wald Test:

$$W_T = w(\hat{\theta})' [W(\hat{\theta}) \text{Cov}(\hat{\theta}) W(\hat{\theta})']^{-1} w(\hat{\theta})$$

$$\Rightarrow W_T = w(\hat{\theta})' [W(\hat{\theta}) \{I_T(\hat{\theta})\}^{-1} W(\hat{\theta})']^{-1} w(\hat{\theta})$$

Note: Can be computed with any consistent estimator $\hat{\theta}$ and $\text{Cov}(\hat{\theta})$.

Likelihood Ratio Test: (LR)

$$LR_T = 2[l_T(\hat{\theta}) - l_T(\tilde{\theta})].$$

Lagrangian Multiplier (LM) test

Define: $s_T(\theta) = \partial l_T(\theta) / \partial \theta$.

$$LM_T = s_T(\tilde{\theta})' [I_T(\tilde{\theta})]^{-1} s_T(\tilde{\theta}).$$

Theorem:

Under H_0 : $w(\theta) = 0$,

$$W_T, LR_T, LM_T \rightarrow_d \chi^2(m).$$

Implication:

- Given confidence level $(1-\alpha)$ or significance level (α) , find a critical value such that
- Usually, $\alpha = 0.05$ or $\alpha = 0.01$.
- If $W_T > c$, reject H_0 . Otherwise, do not reject H_0 .

Comments:

- 1) Wald needs only $\hat{\theta}$; LR needs both $\hat{\theta}$ and $\tilde{\theta}$; and LM needs $\tilde{\theta}$ only.
- 2) In general, $W_T \geq LR_T \geq LM_T$.
- 3) W_T is not invariant to how to write restrictions. That is, W_T for $H_0: \theta_1 = \theta_2$ may not be equal to W_T for $H_0: \theta_1/\theta_2 = 1$.

Example:

(1) $\{x_1, \dots, x_T\}$: RS from $N(\mu_0, v_0)$ with v_0 known. So, $\theta = \mu$.

$H_0: \mu = 0$.

- $w(\mu) = \mu$

- $l_T(\mu) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\sum_t(x_t - \mu)^2$
- $s_T(\mu) = (1/v)\sum_t(x_t - \mu)$
- $\mathbf{I}_T(\mu_0) = E[-\partial^2 l_T(\mu)/\partial \mu^2 |_{\theta=0_0}] = T/v_0$

[Wald Test]

Unrestricted MLE:

- FOC: $\partial l_T(\mu)/\partial \mu = (1/v)\sum_t(x_t - \mu) = 0$
- $\hat{\mu} = \bar{x}$

$$W(\mu) = 1 \Rightarrow W(\hat{\mu}) = 1$$

$$\mathbf{I}_T(\hat{\mu}) = T/v_0$$

[LR Test]

Restricted MLE: $\tilde{\mu} = 0$

$$l_T(\hat{\mu}) = -(T/2)\ln(2\pi) - (T/2)\ln(v_0) - \{1/(2v_0)\}\sum_t(x_t - \bar{x})^2$$

$$l_T(\tilde{\mu}) = -(T/2)\ln(2\pi) - (T/2)\ln(v_0) - \{1/(2v_0)\}\sum_t x_t^2$$

[LM Test]

$$s_T(\tilde{\mu}) = (1/v_0)\sum_t x_t = (T/v_0)\bar{x}; \quad \mathbf{I}_T(\tilde{\mu}) = T/v_0$$

With this information, can show:

$$W = LR = LM = (T\bar{x}^2)/v_0.$$

(2) Both μ and v unknown: $\theta = (\mu, v)'$.

$$H_0: \mu = 0.$$

$$\Rightarrow w(\theta) = \mu$$

$$\Rightarrow W(\theta) = \partial w(\theta)/\partial \theta' = [\partial \mu/\partial \mu, \partial \mu/\partial v] = [1, 0]$$

$$\Rightarrow l_T(\theta) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\sum_t(x_t - \mu)^2$$

$$\Rightarrow s_T(\theta) = [(1/v)\sum_t(x_t - \mu), -T/(2v) + (1/(2v^2))\sum_t(x_t - \mu)^2]'$$

$$\Rightarrow \mathbf{I}_T(\theta_0) = \text{diag}[T/v_0, T/(2v_0^2)].$$

$$\Rightarrow \text{Unrest. MLE: } \hat{\mu} = \bar{x} \text{ and } \hat{v} = (1/T)\sum_t(x_t - \bar{x})^2$$

$$\Rightarrow \text{Restricted MLE: } \tilde{\mu} = 0, \text{ but need to compute } \tilde{v}$$

$$\begin{aligned}
&\Rightarrow l_T(\tilde{\mu}, v) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t(x_t - \tilde{\mu})^2 \\
&\Rightarrow l_T(0, v) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t x_t^2 \\
&\Rightarrow \text{FOC: } \partial l_T(0, v)/\partial v = -T/(2v) + (1/(2v^2))\Sigma_t x_t^2 = 0 \\
&\Rightarrow \tilde{v} = (1/T)\Sigma_t x_t^2
\end{aligned}$$

[Wald Test]

$$\begin{aligned}
&w(\hat{\theta}) = \hat{\mu} - \bar{x}; W(\hat{\theta}) = [1, 0]; I_T(\hat{\theta}) = \text{diag}(T/\hat{v}, T/(2\hat{v}^2)). \\
&\Rightarrow W_T = w(\hat{\theta})'[W(\hat{\theta})\{I_T(\hat{\theta})\}^{-1}W(\hat{\theta})]^{-1}w(\hat{\theta}) = T\bar{x}^2/\hat{v}.
\end{aligned}$$

[LR Test]

$$\begin{aligned}
&l_T(\hat{\theta}) = -(T/2)\ln(2\pi) - (T/2)\ln(\hat{v}) - \{1/(2\hat{v})\}\Sigma_t(x_t - \bar{x})^2 \\
&l_T(\tilde{\theta}) = -(T/2)\ln(2\pi) - (T/2)\ln(\tilde{v}) - \{1/(2\tilde{v})\}\Sigma_t x_t^2
\end{aligned}$$

[LM Test]

$$\begin{aligned}
&s_T(\tilde{\theta}) = [(1/\tilde{v})\Sigma_t x_t, -T/(2\tilde{v}) + (1/2\tilde{v}^2)\Sigma_t x_t^2]' = [T\bar{x}/\tilde{v}, -T/(2\tilde{v}) + T/(2\tilde{v})]' = [T\bar{x}/\tilde{v}, 0]' \\
&I_T(\tilde{\theta}) = \text{diag}(T/\tilde{v}, T/(2\tilde{v}^2)) \\
&\Rightarrow LM_T = s_T(\tilde{\theta})'[I_T(\tilde{\theta})]^{-1}s_T(\tilde{\theta}) = T\bar{x}^2/\tilde{v}.
\end{aligned}$$