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, ... , 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, ... , 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) Σ_xf(x) = 1.
3) F(x) $\leq$ 1.

EX: X = # faced up (a die) with pdf: f(x) = 1/6, where x = 1, ... , 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 $\Pr(a \leq X \leq b)$: $Pr(a \leq X \leq b) = \int_a^b f(v)\,dv$.

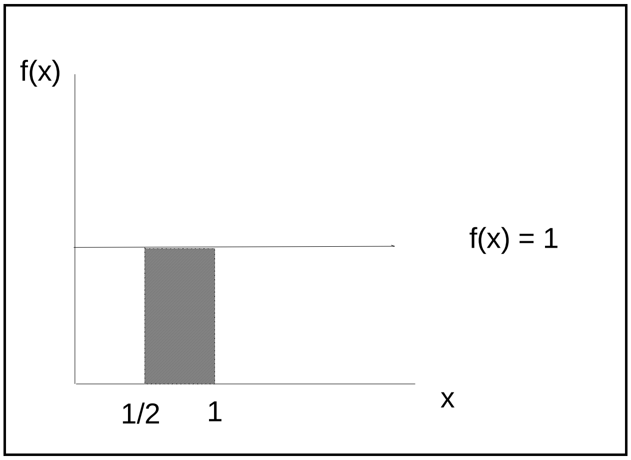
Note: In cases where X is continuous, $\Pr(a \leq X) = \Pr(a < X)$.

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.
  - $E[g(x)] = \Sigma_x g(x)f(x)$ (or $\int_\Omega g(x)f(x)\,dx$).

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

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

Population variance: $\sigma_x^2 = var(x) = E[(x-\mu_x)^2] = \Sigma_x(x - \mu_x)^2 f(x)$ [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 ($f(x) = 1/6$, $x = 1, 2, \ldots, 6$).
- Toss the die repeatedly billions and billions (b) times: $x^{(1)}, x^{(2)}, \ldots, x^{(b)}$ [a population].
- Mean of these = $(1/b)\Sigma_{j=1}^{b} x^{(j)} = \mu_x$, almost surely (a.s.).
- Mean dispersion of these = $(1/b)\Sigma_{j=1}^{b}(x^{(j)} - \mu_x)^2 = \sigma_x^2$, a.s.
- Similarly, $(1/b)\,\Sigma_{j=1}^{b}\, g[x^{(j)}] = E[g(x)]$, a.s.

Median:

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

$\rightarrow$ Order $x^{(1)}, \ldots, x^{(b)}$: $x^{[1]} \le x^{[2]} \le \ldots \le x^{[b]}$.

$\rightarrow$ 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.

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

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 ($x \ge \mu_x$) direction.

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

If X is normally distributed, KC = 0.

Exercise for E(x), var(x) and E[g(x)]:

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

  $E(x) = \Sigma_x xf(x) = 0 \times (1/2) + 1 \times (1/2) = 1/2$; $var(x) = (0-1/2)^2 \times (1/2) + (1-1/2)^2 \times (1/2) = 1/4$.
- $g(x) = (1/2)x^2 + (1/2)x + 2$.

  $E[g(x)] = [1/2+1/2+2] \times (1/2) + [0+0+2] \times (1/2) = 5/2$.
- 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, ... , 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 μ_x = $E[(x - \mu_x)^2] = var(x)$.
- 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: $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, \ldots$.
- $E(x) = var(x) = \lambda$.

(2) Normal distribution

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

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

- Pdf: $$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \quad -\infty < z < \infty .$$
- 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 , ... , Z_k are RVs iid with N(0,1).

  $y = \Sigma_{i=1}^{k} z_i^2 \sim \chi^2(k)$, $y > 0$, with degrees of freedom (df) = k.
- $E(y) = k$; $var(y) = 2k$.

(4) Student t distribution

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

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

- $E(t) = 0$, $k > 1$; $var(t) = k/(k-2)$, $k > 2$.
- As $k \rightarrow \infty$, $var(t) \rightarrow 1$: 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,

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

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

[3] Bivariate Distributions

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

Marginal (unconditional) pdf:

$f_x(x) = \Sigma_y f(x,y) = Pr(X=x)$ regardless of Y; $f_y(y) = \Sigma_x f(x,y) = Pr(Y=y)$ regardless of X.

Conditional pdf:

$f(x|y) = Pr(X = x$, 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:

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

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

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

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

- Conditional pdf:

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

→ $f(x|y=1) = 1/2$, $x = 0, 1$.

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

- Stochastic independence:

  $f_x(x) = f_y(y) = 1/2$; $f_X(x)f_Y(y) = 1/4 = f(x,y)$, for any x and y.

  → 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\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: $E[g(x,y)] = \Sigma_x\Sigma_y\, g(x,y)f(x,y)$ [or $\int\int_\Omega g(x,y)\, f(x,y)dxdy$].

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

Note: $\sigma_{xy} = cov(x,y) > 0 \Rightarrow$ positively linearly related; $\sigma_{xy} = cov(x,y) < 0 \Rightarrow$ negatively linearly related;

$\sigma_{xy} = cov(x,y) = 0 \Rightarrow$ no linear relation.

Correlation Coefficient:

The correlation coefficient between x and y is defined by:

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

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

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

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

Theorem: If X & Y are stoch. indep., 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.

$E(xy) = \Sigma_x\Sigma_y\, xyf(x,y) = 0\times0\times(1/4) + 0\times1\times(1/4) + 1\times0\times(1/4) + 1\times1\times(1/4) = 1/4$.

Conditioning in a Bivariate Distribution:

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

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

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

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

Conditional Mean and Variance:

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

$var(y|x) = E[(y-E(y|x))^2|x] = \Sigma_y(y-E(y|x))^2 f(y|x)$.

Regression model:

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

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

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

$\to$ ϵ = unexplained part of y (called disturbance term).

$\to$ $E(\epsilon|x) = 0$ and $var(\epsilon|x) = var(y|x)$.

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 $E_x(.)$ is the expectation over x = $\Sigma_x \bullet f_x(x)$ or $\int_x \bullet f_x(x)dx$.

Theorem: (Law of Iterative Expectations)

$E(y)$ [unconditional mean] = $E_x[E(y|x)]$.

Proof:

$E(y) = \Sigma_x\Sigma_y yf(x,y) = \Sigma_x\Sigma_y yf(y|x)f_x(x) = \Sigma_x[\Sigma_y yf(y|x)]f_x(x)$.

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

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

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 $var(y|x)$ is constant.

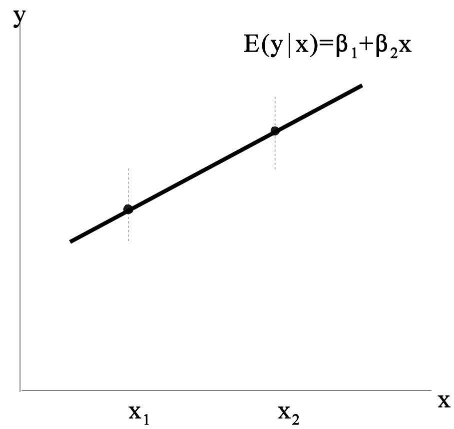
EX: $y = E(y|x) + \epsilon$ with $var(\epsilon|x) = \sigma^2$ (constant).

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

$\rightarrow$ 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 $x = x_1$.
- $var(y|x=x_1)$ measures the dispersion of y given $x = x_1$.
- If $var(y|x=x_1) = var(y|x=x_2) = ...$, we say that y is homoskedastic.
- Law of iterative expectation:

  $E(y) = \Sigma_x E(y|x)f_x(x) = E(y|x=x_1)Pr(x=x_1) + E(y|x=x_2)Pr(x=x_2) + ...$.

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

Theorem: (Decomposition of Variance)

$$var(y) = var_x[E(y|x)] + E_x[var(y|x)].$$

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

Coefficient of Determination:

$$R^2 = var_x[E(y|x)]/var(y).$$

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

$\rightarrow 0 \leq R^2 \leq 1$.

Note:

- $var(y)$ = total variation of y.
- $var_x[E(y|x)] \rightarrow$ 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.

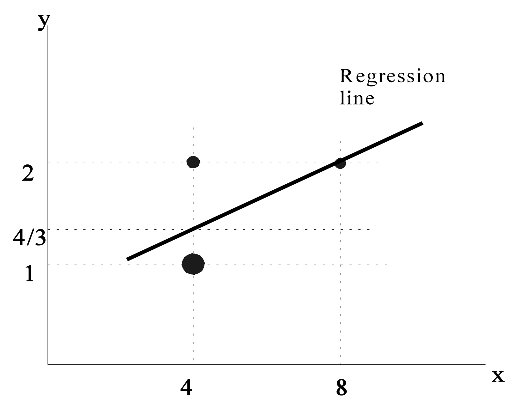
- Wish R^2 close to 1.

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 popuation:



- 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:
  - $E(x) \equiv \mu_x = \Sigma_x x f_x(x) = 4\times f_x(4) + 8\times f_x(8) = 4\times(3/4) + 8\times(1/4) = 5.$
  - $E(y) \equiv \mu_x = \Sigma_y y f_y(y) = 1.5$
- Variances of X and Y:
  - $\text{var}(x) \equiv \sigma_x^2 = \Sigma_x(x-\mu_x)^2 f_x(x) = (4-5)^2 f_x(4) + (8-5)^2 f_x(8) = 1\times(3/4) + 9\times(1/4) = 3.$
  - $\text{var}(y) \equiv \sigma_y^2 = 1/4.$

- Covariance between X and Y:
  - $cov(x,y) \equiv E[(x-\mu_x)(y-\mu_y)] = E(xy) - \mu_x \mu_y = \Sigma_x\Sigma_y xyf(x,y) - \mu_x\mu_y$

    $= 4\times1\times f(4,1) + 4\times2\times f(4,2) + 8\times1\times f(8,1) + 8\times2\times f(8,2) - 5\times1.5 = 0.5.$
  - $\rho_{xy} \equiv \dfrac{cov(x,y)}{\sigma_x\sigma_y} = \dfrac{0.5}{\sqrt{3}\sqrt{1/4}} \approx 0.58.$

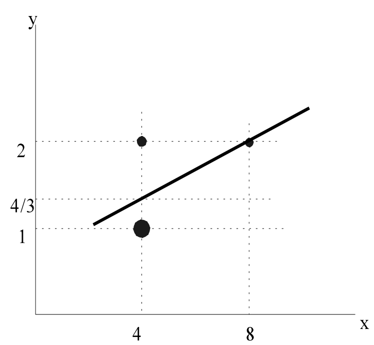
- Conditional Probabilities

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:
  - $E(y|x=4) = \Sigma_y yf(y|x=4) = 1\times f(y=1|x=4) + 2\times f(y=2|x=4) = 1\times(2/3) + 2\times(1/3) = 4/3$
  - $E(y|x=8) = 2.$



- Conditional variance of Y:
  - $var(y|x=4) = \Sigma_y[y-E(y|x=4)]^2 f(y|x=4) = 6/27.$
  - $var(y|x=8) = 0.$

- Law of iterative expectation:
  - $E_x[E(y|x)] = \Sigma_x E(y|x)f_x(x)$

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

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

[4] Bivariate Normal Distribution

Definition: (Bivariate Normal Distribution)

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

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

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

Facts:

1) $f_x(x) \sim N(\mu_x, \sigma_x^2)$ and $f_y(y) \sim N(\mu_y, \sigma_y^2)$.
2) $E(y|x) = \beta_1 + \beta_2 x$ and $var(y|x) = \sigma^2$ (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: $X_1, \ldots, X_n$: random variables.

Let $x = [x_1, \ldots, x_n]'$ ($n\times1$ vector). Then,

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

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

Proof: See Greene.

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

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

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

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

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

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

Definition:

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

$c'x = x'c = c_1x_1 + \ldots + c_nx_n = \Sigma_j c_jx_j$ (scalar).

Theorem:

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

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

Proof:

(1) $E(c'x) = E(\Sigma_j c_j x_j) = \Sigma_j E(c_1 x_1 + ... + c_n x_n) = c_1 E(x_1) + ... + c_n E(x_n) = \Sigma_j c_j E(x_j) = c'E(x)$.

(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 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:

- Cov(x) is symmetric and positive semidefinite.
- Usually, 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 $c = [c_1, ... , c_n]'$. Then, the scalar, $c'Bc$, is called a quadratic form of B.

Definition:

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

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

Theorem:

Let B be a symmetric and square matrix given by:

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

Define the principal minors by:

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

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

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×1 random vector and let A be a m×n matrix of constants (Ax is a m×1 random vector). Then,

$E(Ax) = AE(x); Cov(Ax) = ACov(x)A'$.

[6] Multivariate Normal distribution

Definition:

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

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

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

Pdf of x:

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

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

(2) $\Sigma = Cov(x) = diag(\sigma_x^2, ... , \sigma_x^2) = \sigma_x^2 I_n$.

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

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

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

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

$var(y|x^*) = \sigma^2$.

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

Proof: See Greene.

2. Distributions of linear functions of a normal vector

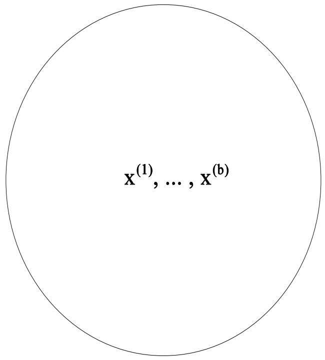
$x_{nx1} \sim N(\mu,\Sigma)$.

$y = Ax + b$, where $A_{m\times n}$ and $b_{m\times 1}$ are fixed.

→ $y \sim N(A\mu+b,A\Sigma A')$.

[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.
- $\{x_1, \ldots, x_T\}$: a sample of size T from the population.
- $\hat{\theta}$: an estimator of θ, which is a function of the sample.

  (e.g, $\hat{\theta} = \overline{x} = (1/T)\Sigma_{t=1}^{T} x_t$.)
- 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: $\{x_1^{[1]}, \ldots, x_t^{[1]}, \ldots, x_T^{[1]}\}$	$\rightarrow$	$\hat{\theta}^{[1]}$
SAM 2: $\{x_1^{[2]}, \ldots, x_t^{[2]}, \ldots, x_T^{[2]}\}$	$\rightarrow$	$\hat{\theta}^{[2]}$
: :		
SAM b': $\{x_1^{[b']}, \ldots, x_t^{[b']}, \ldots, x_T^{[b']}\}$	$\rightarrow$	$\hat{\theta}^{[b']}$

Since $\{x_1, \ldots, x_T\}$ is random, so is $\hat{\theta}$. $\rightarrow$ We can define $E(\hat{\theta})$ and $var(\hat{\theta})$.

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

$\rightarrow$ Means that $x_1, \ldots, x_T$ are iid.

$\rightarrow$ EX: $\{x_1, \ldots, x_T\}$ a RS from $N(\mu,\sigma^2)$.

$\rightarrow$ $x_t \sim N(\mu,\sigma^2)$ for any t = 1, ... , T.

$\rightarrow$ $E(x_t) = \mu$ and $var(x_t) = \sigma^2$, for any t = 1, ... , 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.

→ $x_1, \ldots, x_T$ are not iid.

(3) Criteria for a "good" estimator

1) Minimum Variance Unbiased Estimator

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

Implication: $\hat{\theta}^{[1]}, \ldots, \hat{\theta}^{[b']}$ → $(1/b')\Sigma_{j=1}^{b'}\hat{\theta}^{[j]} = \theta$, a.s.

EX:

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

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

→ $E(\bar{x}) = \mu$ 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.

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

Implication:

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

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

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

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

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

$\tilde{x} = x_1$.

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

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

→ But, $var(\bar{x}) = \sigma^2/T$.

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

Definition:

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

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

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

2) Minimum Mean Square Error (MMSE) Estimator

Definition:

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

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

Theorem:

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

Definition:

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

Note:

1) MMSE estimator could be biased.
2) MMSE is usually a function of θ.

→ To get MMSE, need to know θ.

→ If you know θ, why do you estimate?

→ 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 $x_1, \ldots, x_T = f(x_1, \ldots, x_T, \theta_o)$.
- $L_T(\theta) = f(x_1, \ldots, x_T, \theta)$ (likelihood function).

Remark:

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

Definition: (log-likelihood function)

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

EX:

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

→ $x_t \sim f(x_t,\theta_o)$.

→ $f(x_1, \ldots, x_T,\theta_o) = \Pi_{t=1}^{T} f(x_t, \theta_o)$.

→ $f(x_1, \ldots, x_T,\theta) = \Pi_{t=1}^{T} f(x_t, \theta)$.

→ $\ln[f(x_1, \ldots, x_T,\theta)] = \Sigma_{t=1}^{T} \ln[f(x_t, \theta)]$.

→ $l_T(\theta) = \Sigma_{t=1}^{T} \ln[f(x_t, \theta)]$.

Definition: (Maximum Likelihood Estimator)

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

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:

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

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

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

→ FOC (first order condition): $\partial l_T(\theta)/\partial\theta = \Sigma_t[-1+x_t/\theta] = 0$

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

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

→ $\hat{\theta}$ Efficient.

[8] Extention to the Estimation of Multiple Parameters

Definition:

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

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

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.

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

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

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

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

↔ $[Cov(\tilde{\theta}) - Cov(\hat{\theta})]$ 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, $\mathrm{var}(c'\tilde{\theta}) = \mathrm{var}(c_1\tilde{\theta}_1 + c_2\tilde{\theta}_2) > \mathrm{var}(c_1\hat{\theta}_1 + c_2\hat{\theta}_2) = \mathrm{var}(c'\hat{\theta})$, we can say that $\hat{\theta}$ is a better estimator.

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

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

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

But,

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

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

→ A is not positive definite.

→ Thus, $\hat{\theta}$ is not necessarily more efficient than $\tilde{\theta}$.

→ For example, you wish to estimate $\theta_{o,1} - \theta_{o,2} = c'\theta_o$ ($c' = (1,-1)$).

  → $\mathrm{var}(c'\hat{\theta}) = c'\mathrm{Cov}(\hat{\theta})c = 2$

  → $\mathrm{var}(c'\tilde{\theta}) = c'\mathrm{Cov}(\tilde{\theta})c = 1$

  → 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?

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

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

Theorem:

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

But, the reverse is not true.

Proof:

Let $c' = (1,0,...,0)$. 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)

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

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

Theorem:

$\text{MSE}(\hat{\theta}) = \text{Cov}(\hat{\theta}) + [\theta_o - E(\hat{\theta})][\theta_o - E(\hat{\theta})]'$,

where $[\theta_o - E(\hat{\theta})]$ is called the bias of $\hat{\theta}$.

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,

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

Definition: (MLE)

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

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

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:

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

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

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

For MLE, solve:

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

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

From (1):

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

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

Substituting (3) into (2):

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

Thus,

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

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 $E[(1/(T-1)\Sigma_t(x_t - \bar{x})^2] = \sigma_o^2$)

  $\rightarrow$ biased.
- Let $g(\hat{v}_{ML}) = [T/(T-1)]\hat{v}_{ML}$.

  $\rightarrow E[g(\hat{v}_{ML})] = \sigma_o^2$.

  $\rightarrow g(\hat{v}_{ML}) = [1/(T-1)]\Sigma_t(x_t - \bar{x})^2 = s_x^2$ is efficient.

[9] Large-Sample Theories

(1) Motivation:

- $\hat{\theta}_T$: an estimator from a sample of size T, $\{x_1, \ldots, x_T\}$
- 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:

  $\text{plim}_{T\rightarrow\infty}\hat{\theta}_T = \theta_o$ (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, $E(\hat{\theta}) = \theta_o + 1/T \neq \theta_o$ (biased)

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

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

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

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

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

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

- Law of Large Numbers (LLN)

A. Case of Scalar Random variables:

- Komogorov's Strong LLN:

  Suppose that $\{x_1, ... , x_T\}$ is a RS from a population with μ_o and σ_o^2.

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

- Generalized Weak LLN (GWLLN):
  - $\{x_1, ... , x_T\}$ is a sample (not necessarily RS)
  - Define $E(x_1) = \mu_{o,1}, ... , E(x_T) = \mu_{o,T}$.
  - Define $\text{var}(x_1) = \sigma_{o,1}^2, ... , \text{var}(x_T) = \sigma_{o,T}^2$.

    Assume that $\sigma_{o,1}^2, ... , \sigma_{o,T}^2 < \infty$.
  - Then, under suitable assumptions, $\text{plim}\ \bar{x} = \lim \frac{1}{T}\Sigma_t\mu_{o,t}$.

B. Case of Vector Random Variables:

- GWLLN
  - x_t: p×1 random vector.
  - $\{x_1, ... , x_T\}$ is a sample.
  - Let $E(x_1) = \mu_{o,1}$ (p×1), ... , $E(x_T) = \mu_{o,T}$.
  - Assume that $\text{Cov}(x_j)$ are well-defined and finite.
  - Then, under suitable assumptions.

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

- Central Limit Theorems (CLT)

A. Case of Scalar Random Variables:

- Motivation:
  - Suppose that $\{x_1, \ldots, x_T\}$ is a RS from a population with μ_o and σ_o^2.
  - We know $\bar{x} \to \mu_o$ as $T \to \infty$. 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 $\{x_1, \ldots, x_T\}$ is a RS from a population with μ_o and σ_o^2.
  - Then, $\sqrt{T}(\bar{x}-\mu_o) \to_d N(0,\sigma_o^2)$, or equivalently, $\sqrt{T}\frac{\bar{x}-\mu_o}{\sigma_o} \to_d N(0,1)$.
- Implication of CLT:
  - $\sqrt{T}(\bar{x}-\mu_o) \approx N(0,\sigma_o^2)$, if T is large.
  - $E[\sqrt{T}(\bar{x}-\mu_o)] = \sqrt{T}[E(\bar{x}) - \mu_o] \approx 0 \quad \to \quad E(\bar{x}) \approx \mu_o$.
  - $var[\sqrt{T}(\bar{x}-\mu_o)] = Tvar(\bar{x}-\mu_o) = Tvar(\bar{x}) \approx \sigma_o^2 \to var(\bar{x}) \approx \sigma_o^2/T$.
  - $\bar{x} \approx N(\mu_o, \sigma_o^2/T)$, if T is large.

B. Case of Random vectors:

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

$$\frac{1}{\sqrt{T}}\Sigma_{t=1}^T y_t \quad \to_d \quad N(0, \lim_{T\to\infty}\frac{1}{T}Cov(\Sigma_{t=1}^T y_t)) \; .$$

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

$$\frac{1}{\sqrt{T}}\Sigma_{t=1}^T y_t \quad \to_d \quad N(0, \lim_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^T Cov(y_t)) \; .$$

[Technical Details]

(3) Convergency in probability

Definition:

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

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

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

Definition: (Convergency in probability, Weak Convergency)

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

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

Or equivalently,

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

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

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

Choose $0 < \epsilon < 1$:

$\Pr(|\hat{\theta}_T - 0| > \epsilon) = \Pr(|\hat{\theta}_T| > \epsilon) = \Pr(\hat{\theta}_T > \epsilon) = 1/T$

$\Rightarrow \lim_{T\to\infty}\Pr(|\hat{\theta}_T - 0| > \epsilon) = 0$

$\Rightarrow \hat{\theta}_T \to_p 0$.

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

Choose $0 < \epsilon < 1$:

$\Pr(|\hat{\theta}_T - 0| > \epsilon) = \Pr(|\hat{\theta}_T| > \epsilon) = \Pr(\hat{\theta}_T > \epsilon) = 1/T$

$\Rightarrow \lim_{T\to\infty}\Pr(|\hat{\theta}_T - 0| > \epsilon) = 0$

$\Rightarrow \hat{\theta}_T \to_p 0$.

Digression to other stronger convergency:

Definition: (Convergence in mean square)

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

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

Theorem: m.s. $\Rightarrow$ p.

Proof:

Chebychev's inequality (see Greene) says:

For any $\epsilon > 0$, $\Pr(|\hat{\theta}_T - c| > \epsilon) \leq E(|\hat{\theta}_T - c|^2)/\epsilon^2$.

$\Rightarrow \lim_{T\to\infty}\Pr(|\hat{\theta}_T - c| > \epsilon) = \lim_{T\to\infty}E(|\hat{\theta}_T - c|^2)/\epsilon^2 = 0$.

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

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

- Observe $E[|\hat{\theta}_T - 0|^2] = E[\hat{\theta}_T{}^2] = 0^2 \times [1-(1/T)] + 1^2 \times (1/T) = 1/T$

  $\Rightarrow \lim_{T\to\infty}E[|\hat{\theta}_T - 0|^2] = 0$.

  $\Rightarrow \hat{\theta}_T \to 0$ m.s..

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

- $\hat{\theta}_T \to_p 0$.
- Observe $E[|\hat{\theta}_T - 0|^2] = E[\hat{\theta}_T{}^2] = 0^2 \times [1-(1/T)] + T^2 \times (1/T) = T$

  $\Rightarrow \lim_{T\to\infty}E[|\hat{\theta}_T - 0|^2] = \infty$.

  $\Rightarrow$ 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($\hat{\theta}_T$ deviates from c) should become increasingly small as $T \to \infty$. 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 $T \to \infty$. For example, if we assume $\Pr(\hat{\theta}_T = T^{1/4}) = 1/T$ instead, we can show that $\hat{\theta}_T \to c$, m.s.

Definition: (Almost sure convergency, Strong Convergency)

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

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

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

- $\hat{\theta}_T \to_p 0$ and $\hat{\theta}_T \to 0$, m.s..
- But, can't determine whether $\hat{\theta}_T \to 0$, a.s..

  (Observe that $\Sigma_{T=1}^{\infty}E(|\hat{\theta}_T - c|^2) = \Sigma_{T=1}^{\infty}(1/T) = \infty$.)

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

- $\hat{\theta}_T \to_p 0$.
- Observe $E[|\hat{\theta}_T - 0|^2] = E[\hat{\theta}_T{}^2] = 0^2 \times [1-(1/T^2)] + 1^2 \times (1/T^2) = 1/T^2$:
  - $\lim_{T\to\infty}E[|\hat{\theta}_T - 0|^2] = 0$.

    $\Rightarrow$ $\Sigma_{T=1}^{\infty}E(|\hat{\theta}_T - c|^2) = \Sigma_{T=1}^{\infty}(1/T^2) < \infty$

    $\Rightarrow$ $\hat{\theta}_T \to 0$, 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 $T \to \infty$.

End of Digression

Definition:

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

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

Question:

An example for a consistent estimator?

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

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

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

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

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

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

Note:

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

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

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

[Proof by GWLLN]

$(1/T)\Sigma_t E(x_t) = (1/T)\Sigma_t\mu_o = (1/T)T\mu_o = \mu_o$

$\rightarrow \lim_{T\rightarrow\infty}(1/T)\Sigma_t E(x_t) = \lim_{t\rightarrow\infty}\mu_o = \mu_o$

$\rightarrow$ By GWLLN, $\bar{x} \rightarrow_p \mu_o$.

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 $\hat{\theta}_T \rightarrow_p \theta_o$.

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

EX: $\{x_1, ... , x_T\}$: 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 an 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,

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

(4) Convergency in distribution

Definition: (Convergency in distribution)

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

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

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

⇒ $z_T \to_d z$.

Fact: d. differs from p.

EX: Two dice A and B.

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

B is unfair:

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

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

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

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

⇒ $z_T \to_d z$.

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

Theorem: (Mann and Wald)

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

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

Theorem:

A_T: a random matrix with plim $A_T = A$.

z_T: a random vector $\to_d z$.

⇒ $A_Tz_T \to_d Az$.

EX: (Central Limit Theorem, CLT)

$\{x_1, \ldots, x_T\}$: RS from a population with μ_o and σ_o^2.

$\Rightarrow$ Lindberg-Levy CLT says

$$\sqrt{T}(\bar{x} - \mu_o) \quad \to_d \quad N(0, \sigma_o^2) \ .$$

Theorem: (Generalized CLT, GCLT)

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

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

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

$$\frac{1}{\sqrt{T}}\Sigma_{t=1}^{T} y_t \quad \to_d \quad N(0, \lim_{T\to\infty}\frac{1}{T}Cov(\Sigma_{t=1}^{T} y_t)) \ .$$

Note:

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

EX: (Lindberg-Levy CLT)

$\{x_1, \ldots, x_T\}$: RS from a population with μ_o and σ_o^2.

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

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

$var(y_t) = var(x_t) = \sigma_o^2$.

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

$(1/T)var(\Sigma_t y_t) = (1/T)var(\Sigma_t x_t - T\mu) = (1/T)var(\Sigma_t x_t)$

$= (1/T)\Sigma_t var(x_t) = (1/T)T\sigma_o^2 = \sigma_o^2$

$\Rightarrow \lim (1/T)var(\Sigma_t y_t) = \sigma_o^2$.

$\Rightarrow \quad \sqrt{T}(\bar{x} - \mu_o) \to_d N(0, \sigma_o^2)$.

Corollary:

Assume the same things as GCLT.

Assume that the y_t's are linearly uncorrelated.

Then,

$$\frac{1}{\sqrt{T}}\Sigma_{t=1}^{T} y_t \quad \to_d \quad N(0, \lim_{T\to\infty}\frac{1}{T}\Sigma_{t=1}^{T} Cov(y_t)) \ .$$

Proof:

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

Lemma:

Let $E(y_t|y_{t-1}, y_{t-2}, ... , y_1) = 0$. [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.

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

⇒ By the law of iterative expectation,

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

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

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

Theorem: (GCLT for martingale difference sequences)

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

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

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

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

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

[10] Large-Sample Properties of MLE

A Short Digression to Matrix Algebra

Definition:

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

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

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

2) $w(\theta)$: a $m\times1$ vector:

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

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

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

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

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

$\Rightarrow$ 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:

Let $w(\theta) = \begin{bmatrix} \theta_1^2+\theta_2 \\ \theta_1+\theta_2^2 \end{bmatrix}$. 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: $p\times1$ ($c'\theta$ is a scalar)

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

2) R: m×p, θ: p×1 (Rθ is m×1)

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

3) A: p×p symmetric, θ: p×1 (θ′Aθ)

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

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

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

End of Digression

Definition: (Hessian matrix of log-likelihood function)

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

Definition: (Information matrix)

$\mathbf{I}_T(\theta_o) = E[-H_T(\theta_o)]$.

Note: To compute $\mathbf{I}_T(\theta_o)$, 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:

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

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

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

The first derivatives:

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

The second derivatives:

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

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

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

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

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

$$\rightarrow \quad E\left[-\frac{\partial^2 l_T(\theta_o)}{\partial v\partial v}\right] = E\left[-\frac{T}{2v_o^2} + \frac{1}{v_o^3}\Sigma_t(x_t-\mu_o)^2\right]$$

$$= -\frac{T}{2v_o^2} + \frac{1}{v_o^3}\Sigma_t E[(x_t-\mu_o)^2] = -\frac{T}{2v_o^2} + \frac{1}{v_o^3}\Sigma_t v_o = -\frac{T}{2v_o^2} + \frac{Tv_o}{v_o^3} = \frac{T}{2v_o^2}.$$

Therefore,

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

Hence,

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

[Sketchical Technical Notes For MLE]

Definition:

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

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

$E_\theta(g(x,\theta)) \equiv \int_\Omega g(x,\theta)f(x,\theta)dx$ (expected value of $g(x,\theta)$ 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 $p\times1$ 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) $\{x_1, ..., x_T\}$ is a random sample from a population satifying (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:

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

$$= E_\theta(s(x,\theta)),$$

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, $E(s(x,\theta_o)) = 0$.

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:

$$\int_\Omega \left[\frac{\partial^2 \ln f(x,\theta)}{\partial\theta\partial\theta} f(x,\theta) + \frac{\partial \ln f(x,\theta)}{\partial\theta}\frac{\partial f(x,\theta)}{\partial\theta}\right]dx = 0$$

$$\rightarrow \int_\Omega \left[\frac{\partial^2 \ln f(x,\theta)}{\partial\theta\partial\theta} f(x,\theta) + \frac{\partial \ln f(x,\theta)}{\partial\theta}\frac{\partial \ln f(x,\theta)}{\partial\theta} f(x,\theta)\right]dx = 0$$

$$\rightarrow E_\theta\left[\frac{\partial^2 \ln f(x,\theta)}{\partial\theta\partial\theta} + \frac{\partial \ln f(x,\theta)}{\partial\theta}\frac{\partial \ln f(x,\theta)}{\partial\theta}\right] = 0, \text{ for any } \theta$$

Corollary 2:

Under Assumption 2,

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

EX:

- $f(x,\theta) = \frac{1}{\sqrt{2\pi v}}\exp\left[-\frac{(x-\mu)^2}{2v}\right]$; $\theta = \begin{bmatrix}\mu\\ v\end{bmatrix}$; $\theta_o = \begin{bmatrix}\mu_o\\ v_o\end{bmatrix}$

- Assumption 1 holds?

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

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

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

- $\Rightarrow$ Clearly, $E[\ln f(x,\theta)]$ is maximized at $\mu = \mu_o$.
- $\Rightarrow$ 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 $T^{-1}l_T(\theta) = T^{-1}\Sigma_t \ln f(x_t,\theta)$. Since $\{x_1, \ldots, x_T\}$ ia a random sample, we can regard $\{\ln f(x_1,\theta), \ldots, \ln f(x_T,\theta)\}$ as a random sample from a population of the random variable $\ln f(x,\theta)$. Then, by LLN, $T^{-1}l_T(\theta) \to_p E[\ln f(x,\theta)]$. But Assumption 1 implies that θ_o uniquely maximize $E[\ln f(x,\theta)] = \text{plim}\, T^{-1}l_T(\theta_o)$. That is, θ_o maximizes 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 $s_t(\theta) = s(x_t,\theta) = \partial \ln f(x_t,\theta)/\partial\theta$. Under Assumptions 1-2 and other suitable assumptions,

$$\frac{1}{\sqrt{T}}\frac{\partial l_T(\theta)}{\partial\theta}\Big|_{\theta=\theta_o} \to_d N(0, \lim \frac{1}{T} Cov\left[\Sigma_t s_t(\theta_o)\right]) .$$

<Proof>

Note that:

$$\frac{1}{\sqrt{T}}\frac{\partial l_T(\theta)}{\partial\theta}\Big|_{\theta=\theta_o} = \frac{1}{\sqrt{T}}\Sigma_t\frac{\partial lnf(x_t,\theta)}{\partial\theta}\Big|_{\theta=\theta_o} = \frac{1}{\sqrt{T}}\Sigma_t s_t(\theta_o).$$

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

Lemma 4:

Under Assumptions 1-2 and other suitable assumptions,

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

<proof>

Note that $- \frac{1}{T}\frac{\partial^2 l_T(\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o} = - \frac{1}{T}\Sigma_t \frac{\partial^2 lnf(x_t,\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o}$.

Then, by GWLLN,

$$- \frac{1}{T}\frac{\partial^2 l_T(\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o} = - \frac{1}{T}\Sigma_t \frac{\partial^2 lnf(x_t,\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o}$$

$$\rightarrow_p \lim \frac{1}{T}\mathrm{E}\left[-\Sigma_t \frac{\partial^2 \mathrm{lnf}(x_t,\theta_o)}{\partial\theta\partial\theta'}|_{\theta=\theta_o}\right] = \lim \frac{1}{T}E\left[-\frac{\partial^2 l_T(\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o}\right] = \frac{1}{T}\mathrm{I}_T(\theta_o).$$

Lemma 5:

Under Assumptions 1-2, $Cov\left[\Sigma_t s_t(\theta_o)\right] = \mathrm{I}_T(\theta_o)$.

<proof>

Since $\{s_1(\theta_o), \ldots, s_T(\theta_o)\}$ is a RS,

$$\mathrm{Cov}\left[\Sigma_t s_t(\theta_o)\right] = \Sigma_t \mathrm{Cov}[s_t(\theta_o)] = \Sigma_t \mathrm{E}[s_t(\theta_o)s_t(\theta_o)'].$$

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

$$\mathrm{I}_T(\theta_o) = \Sigma_t \mathrm{E}\left[-\frac{\partial^2 \mathrm{lnf}(x_t,\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o}\right].$$

Thus, it is enought to show that

$$\mathrm{E}[s_t(\theta_o)s_t(\theta_o)'] = \mathrm{E}\left[-\frac{\partial^2 \mathrm{lnf}(x_t,\theta)}{\partial\theta\partial\theta'}|_{\theta=\theta_o}\right].$$

But this equality holds by Lemma 2.

Corollary 3:

Under Assumptions 1-2 and other suitable assumptions,

$$\frac{1}{\sqrt{T}}\frac{\partial l_T(\theta_o)}{\partial\theta} \rightarrow_d \mathrm{N}(0, \lim \frac{1}{\mathrm{T}}\mathrm{I}_T(\theta_o)).$$

Theorem 2:

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

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

<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:

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

$$\rightarrow \quad \sqrt{T}(\hat{\theta}-\theta_o) = [\lim\frac{1}{T}I_T(\theta_o)]^{-1}\frac{1}{\sqrt{T}}\frac{\partial l_T(\theta_o)}{\partial \theta} + o_p(1) \quad \text{(By Lemma 4)}$$

$$\rightarrow \quad \sqrt{T}(\hat{\theta}-\theta_o) \rightarrow_d N(0, [\lim\frac{1}{T}I_T(\theta_o)]^{-1}\lim\frac{1}{T}I_T(\theta_o)[\lim\frac{1}{T}I_T(\theta_o)]')$$
$$= N(0,[\lim\frac{1}{T}I_T(\theta_o)]^{-1}). \quad \text{(By Corollary 3)}$$

[11] Testing Hypotheses Based on MLE

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

General form of hypotheses:

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

Examples:

1) θ: a scalar

H_o: $\theta = 2$ → H_o: $\theta - 2 = 0$ → H_o: $w(\theta) = 0$, where $w(\theta) = \theta - 2$.

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

H_o: $\theta_1^2 = \theta_2 + 2$ and $\theta_3 = \theta_1 + \theta_2$

→ H_o: $\theta_1^2-\theta_2-2 = 0$ and $\theta_3-\theta_1-\theta_2 = 0$.

→ Let $w_1(\theta) = \theta_1^2-\theta_2-2$ and $w_2(\theta) = \theta_3-\theta_1-\theta_2$.

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

3) linear restrictions

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

H_o: $\theta_1 = \theta_2 + 2$ and $\theta_3 = \theta_1 + \theta_2$

→ H_o: $\theta_1 - \theta_2 - 2 = 0$ and $\theta_3 - \theta_1 - \theta_2 = 0$

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

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

Remark:

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

H_o: $R\theta - r = 0_{mx1}$,

where R and r are known mxp and mx1 matrices, respectively.

Definition:

$$W(\theta) \;=\; \frac{\partial w(\theta)}{\partial \theta'} = \begin{bmatrix} \frac{\partial w_1(\theta)}{\partial \theta_1} & \frac{\partial w_1(\theta)}{\partial \theta_2} & \cdots & \frac{\partial w_1(\theta)}{\partial \theta_p} \\ \frac{\partial w_2(\theta)}{\partial \theta_1} & \frac{\partial w_2(\theta)}{\partial \theta_2} & \cdots & \frac{\partial w_2(\theta)}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial w_m(\theta)}{\partial \theta_1} & \frac{\partial w_m(\theta)}{\partial \theta_2} & \cdots & \frac{\partial w_m(\theta)}{\partial \theta_p} \end{bmatrix}_{mxp}$$

Example:

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

H_o: $\theta_1^2 - \theta_2 = 0$ and $\theta_1 - \theta_2 - \theta_3^2 = 0$.

$$\rightarrow\; w(\theta) \;=\; \begin{bmatrix} \theta_1^2 - \theta_2 \\ \theta_1 - \theta_2 - \theta_3^2 \end{bmatrix} \;\rightarrow W(\theta) = \begin{bmatrix} 2\theta_1 & -1 & 0 \\ 1 & -1 & -2\theta_3 \end{bmatrix}_{2x3}$$

Example:

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

H_o: $\theta_1 = 0$ and $\theta_2 + \theta_3 = 1$.

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

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

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

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

Definition: (Restricted MLE)

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

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

Wald Test:

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

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

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

Likelihood Ratio Test: (LR)

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

Lagrangean 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_o: $w(\theta) = 0$,

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

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_o. Otherwise, do not reject H_o.

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_o: $\theta_1 = \theta_2$ may not be equal to W_T for H_o: $\theta_1/\theta_2 = 1$.

Example:

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

H_o: $\mu = 0$.

• $w(\mu) = \mu$

• $l_T(\mu) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t(x_t-\mu)^2$

• $s_T(\mu) = (1/v)\Sigma_t(x_t-\mu)$

• $\mathbf{I}_T(\mu_o) = E[-\partial^2 l_T(\mu)/\partial\mu^2|_{\theta=\theta o}] = T/v_o$

[Wald Test]

Unrestricted MLE:

• FOC: $\partial l_T(\mu)/\partial\mu = (1/v)\Sigma_t(x_t-\mu) = 0$

• $\hat{\mu} = \bar{x}$

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

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

[LR Test]

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

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

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

[LM Test]

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

With this information, can show:

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

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

H_o: $\mu = 0$.

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

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

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

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

$\Rightarrow I_T(\theta_o) = \mathrm{diag}[T/v_o, T/(2v_o^2)]$.

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

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

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

$\Rightarrow l_T(0,v) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - \{1/(2v)\}\Sigma_t x_t^2$

$\Rightarrow$ FOC: $\partial l_T(0,v)/\partial v = -T/(2v) + (1/(2v^2))/\Sigma_t x_t^2 = 0$

$\Rightarrow \tilde{v} = (1/T)\Sigma_t x_t^2$

[Wald Test]

$w(\hat{\theta}) = \hat{\mu} = \bar{x}$; $W(\hat{\theta}) = [1,0]$; $I_T(\hat{\theta}) = \text{diag}(T/\hat{v}, T/(2\hat{v}^2))$.

$\Rightarrow W_T = w(\hat{\theta})'[W(\hat{\theta})\{I_T(\hat{\theta})\}^{-1}W(\hat{\theta})]^{-1}w(\hat{\theta}) = T\bar{x}^2/\hat{v}$.

[LR Test]

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

$l_T(\tilde{\theta}) = -(T/2)\ln(2\pi) - (T/2)\ln(\tilde{v}) - \{1/(2\tilde{v})\}\Sigma_t x_t^2$

[LM Test]

$s_T(\tilde{\theta}) = [(1/\tilde{v})\Sigma_t x_t, -T/(2\tilde{v}) + (1/2\tilde{v}^2)\Sigma_t x_t^2]' = [T\bar{x}/\tilde{v}, - T/(2\tilde{v})+T/(2\tilde{v})]' = [T\bar{x}/\tilde{v}, 0]'$

$I_T(\tilde{\theta}) = \text{diag}(T/\tilde{v}, T/(2\tilde{v}^2))$

$\Rightarrow LM_T = s_T(\tilde{\theta})'[I_T(\tilde{\theta})]^{-1}s_T(\tilde{\theta}) = T\bar{x}^2/\tilde{v}$.