

# ECON 5340 Class Notes

## Review of Statistical Inference

### 1 Samples and Sampling Distributions

Definition. We say  $X_1, \dots, X_n$  is a random sample of size  $n$  if each  $X_i$  is drawn independently from the same pdf,  $f(x_i, \theta)$ .

Notes:

1.  $\{X_i\}_{i=1}^n$  is sometimes said to be an independent and identically distributed (i.i.d.) random sample.
2.  $\theta$  is a vector of parameters (e.g.,  $\theta = (\mu, \sigma^2)$ ).
3. Three data types: time series, cross sectional, and panel.

#### 1.1 Descriptive Statistics

Definition. A function of one or more random variables that does not depend on any unknown parameters is a statistic.

1. Measures of Central Tendency.

- Mean.  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ .
- Median. Let  $Y_1, \dots, Y_n$  be the reordering of  $X_1, \dots, X_n$  from smallest to largest.  $Y_i$  is called the  $i^{th}$  order statistic of  $X_1, \dots, X_n$ . The median is defined as  $Y_{(n+1)/2}$ .
- Mode. Most frequent  $X_i$ .

2. Measures of Dispersion.

- $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .
- $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ .

3. Measures of Association.

- Covariance.  $s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ .
- Correlation.  $r_{xy} = s_{xy} / (s_x s_y)$  where  $-1 \leq r_{xy} \leq 1$ .

## 1.2 Sampling Distribution

Definition. A statistic (e.g.,  $Y_1$ ,  $\bar{X}$  and  $s_{xy}$ ) is a random variable with a distribution called a sampling distribution.

Example. If  $X_1, \dots, X_n$  are a random sample with mean  $\mu$  and variance  $\sigma_x^2$ , then  $\bar{X}$  is a random variable with a sampling distribution that has mean  $\mu$  and variance  $\sigma_x^2/n$ .

Proof.

1.  $E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n}(n\mu) = \mu$ .
2.  $Var(\bar{X}) = \frac{1}{n^2} Var(\sum_{i=1}^n X_i) = \frac{1}{n^2}(n\sigma_x^2) = \frac{1}{n}\sigma_x^2$ .

See Gauss example S.1 for the sampling distributions of  $\bar{X}$  where  $X_i \sim N(0, 1)$  with  $n = 3, 10, 100$ .

## 2 Finite Sample Estimation

Definition. An estimator is a rule for using the sample data to form either a point (i.e., single value) or interval (i.e., range of values) estimate.

### 2.1 Estimation Criterion

1. Unbiasedness. An estimator is unbiased if  $E(\hat{\theta}) = \theta$ .

Examples.

- $\bar{X}$  is an unbiased estimator of  $\mu$ .
  - The statistic  $Z = X + 1000$  if coin is “heads”,  $Z = X - 1000$  if coin is “tails” is an unbiased estimator of  $\mu$ .
2. Efficient Unbiasedness. An unbiased estimator  $\hat{\theta}_1$  is efficient if there is no  $\hat{\theta}_i$  such that  $var(\hat{\theta}_i) < var(\hat{\theta}_1)$ ,  $i \neq 1$ .

Example continued.

- $var(\bar{X}) = \sigma_x^2/n$
- $var(Z) = 0.5E(X + 1000 - \mu)^2 + 0.5E(X - 1000 - \mu)^2$ .

3. Mean-Square Error. The mean-square error (MSE) of  $\hat{\theta}$  is

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

Notes:

1. Given some regularity conditions, the  $\text{var}(\hat{\theta})$  will never be smaller than the Cramer-Rao lower bound.
2. A minimum variance unbiased estimator (MVUE) is an efficient unbiased estimator among all linear and nonlinear estimators.
3. A minimum variance linear unbiased estimator (or sometimes called best linear unbiased estimator, BLUE) is an efficient estimator among all linear estimators.
4. Attaining the Cramer-Rao lower bound  $\implies$  efficiency. However, efficiency  $\nRightarrow$  attaining the Cramer-Rao lower bound.
5. A linear estimator is one that is a linear function of the data.

## 2.2 $s^2$ versus $\hat{\sigma}^2$ . Which is a better estimator?

- Is  $s^2$  unbiased?

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= \frac{1}{n-1} E\left(\sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2\right) \\ &= \frac{1}{n-1} \left[ E \sum_{i=1}^n (X_i - \mu)^2 - 2E \sum_{i=1}^n (X_i - \mu)(\bar{X} - \mu) + E \sum_{i=1}^n (\bar{X} - \mu)^2 \right] \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n E(X_i - \mu)^2 - 2nE(\bar{X} - \mu) \frac{1}{n} \sum_{i=1}^n (X_i - \mu) + nE(\bar{X} - \mu)^2 \right] \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \right] \\ &= \frac{1}{n-1} \left[ n\sigma^2 - n \frac{\sigma^2}{n} \right] = \sigma^2. \end{aligned}$$

Yes,  $s^2$  is an unbiased estimator of  $\sigma^2$ .

- Is  $\hat{\sigma}^2$  unbiased?

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) = \frac{n-1}{n} E(s^2) = \frac{n-1}{n} \sigma^2.$$

No,  $\hat{\sigma}^2$  is not an unbiased estimator of  $\sigma^2$ . However, the bias clearly shrinks as  $n$  grows.

- What is the variance of  $s^2$ ?

$$\text{var}(s^2) = \frac{2\sigma^4}{(n-1)} = \text{MSE}(s^2).$$

- What is the variance of  $\hat{\sigma}^2$ ?

$$\text{var}(\hat{\sigma}^2) = \text{var}\left(\frac{n-1}{n}s^2\right) = \left(\frac{n-1}{n}\right)^2 \text{var}(s^2) < \text{var}(s^2).$$

- Which estimator has a smaller MSE?

$$\begin{aligned} \text{MSE}(\hat{\sigma}^2) &= \text{var}(\hat{\sigma}^2) + \text{bias}(\hat{\sigma}^2)^2 \\ &= \left(\frac{n-1}{n}\right)^2 \text{var}(s^2) + \left(-\frac{1}{n}\sigma^2\right)^2 \\ &= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{(n-1)} + \frac{1}{n^2}\sigma^4 \\ &= \frac{(2n-1)\sigma^4}{n^2} = \frac{2\sigma^4}{n} - \frac{\sigma^4}{n^2} < \text{MSE}(s^2). \end{aligned}$$

Therefore,  $\hat{\sigma}^2$  has a smaller MSE than  $s^2$ .

### 3 Large-Sample Distribution Theory

Large-sample distribution theory is important because the small-sample distribution of random variables are often unknown.

#### 3.1 Convergence in Probability

Definition. Let  $X_n$  be a random variable whose distribution depends on  $n$ . We say  $X_n$  converges in probability to  $c$  (or  $\text{plim } X_n = c$ ) if  $\lim_{n \rightarrow \infty} \Pr(|X_n - c| > \epsilon) = 0$  for every  $\epsilon > 0$ . If  $X_n$  has mean  $\mu_n$  and variance  $\sigma_n^2$  with limits  $c$  and 0, then  $X_n$  convergence in mean square to  $c$ .

To calculate  $\text{plim } X_n$ , we will use Markov's inequality:

$$\Pr(u(X) \geq \delta) \leq E(u(X))/\delta$$

for all  $\delta > 0$  and  $u(X) \geq 0$ . If we let  $u(X) = (X - \mu)^2$  and let  $\delta = k^2\sigma^2$ , we get

$$\Pr((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{\sigma^2}{k^2\sigma^2} \implies \Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

which is Chebyshev's inequality.

Notes:

1.  $\hat{\theta}$  is a consistent estimator of  $\theta$  iff  $\text{plim}(\hat{\theta}) = \theta$ .
2. Convergence in mean square  $\implies$  convergence in probability. Convergence in probability  $\not\Rightarrow$  convergence in mean square.
3. Slutsky's Theorem. If  $g(X)$  is a continuous function not in  $n$ ,  $\text{plim}(g(x)) = g(\text{plim}(x))$ . For example,  $E(\bar{X}_n^2) = ?$  but  $\text{plim}(\bar{X}_n^2) = \text{plim}(\bar{X}_n)^2 = \mu^2$ .
4. Jensen's Inequality. If  $g(X_n)$  is concave in  $X_n$ ,  $g(E(X_n)) \geq E(g(X_n))$ .
5. Using Slutsky's theorem where  $\text{plim } X_n = c$  and  $\text{plim } Y_n = d$ ,
  - (a)  $\text{plim}(X_n + Y_n) = c + d$ .
  - (b)  $\text{plim}(X_n Y_n) = cd$ .
  - (c)  $\text{plim}(X_n/Y_n) = c/d, d \neq 0$ .

Example #1. Consider the pdf  $f_n(x)$

$$\begin{aligned} &= 1 - \frac{1}{n} \text{ if } x = 0 \\ &= \frac{1}{n} \text{ if } x = n. \end{aligned}$$

Find what, if anything,  $X_n$  converges to in probability and mean square.

Answer.

- Begin by finding the mean and variance of  $X_n$ .

$$\begin{aligned} E(X_n) &= 0 \left[1 - \frac{1}{n}\right] + n \left[\frac{1}{n}\right] = 1 = \mu \\ \text{var}(X_n) &= (-1)^2 \left[1 - \frac{1}{n}\right] + (n-1)^2 \left[\frac{1}{n}\right] = n - 1 = \sigma^2. \end{aligned}$$

- Convergence in mean square.

$$\lim_{n \rightarrow \infty} \mu = 1 \text{ and } \lim_{n \rightarrow \infty} \sigma^2 = \infty$$

Therefore,  $X_n \xrightarrow{ms} 1$ .

- Convergence in probability.

$$\lim_{n \rightarrow \infty} \Pr(|X_n - 1| > \epsilon) = 0$$

for any reasonably small  $\epsilon$ . Therefore,  $X_n \xrightarrow{p} 1$ . However,

$$\begin{aligned} \Pr(X_n = 0) &= 1 - \frac{1}{n} \\ \implies \lim_{n \rightarrow \infty} \Pr(X_n = 0) &= \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) = 1 \\ \implies \lim_{n \rightarrow \infty} \Pr(|X_n| \geq \epsilon) &= 0 \text{ for every } \epsilon > 0. \end{aligned}$$

Therefore,  $X_n \xrightarrow{p} 0$ .

### 3.2 Convergence in Distribution

Definition.  $X_n$  is said to converge in distribution to  $F(x)$  if  $\lim_{n \rightarrow \infty} F_n(x) = F(x)$  at all continuity points of  $F(x)$ .

Notes:

1. Converge in distribution:  $X_n \xrightarrow{d} X$ .
2.  $F(x)$  is the limiting distribution of  $X_n$ .
3. The mean and variance of  $F(x)$  are called the limiting mean and limiting variance.
4. Rules when  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{p} c$ .
  - (a)  $X_n Y_n \xrightarrow{d} cX$ ,  $X_n + Y_n \xrightarrow{d} X + c$  and  $X_n / Y_n \xrightarrow{d} X/c$ .
  - (b) If  $g(X_n)$  is a continuous function,  $g(X_n) \xrightarrow{d} g(X)$ .
  - (c) If  $\text{plim}(X_n - Y_n) = 0$ , then  $Y_n \xrightarrow{d} X$ , provided a limiting distribution for  $Y_n$  exists.

Example. The pdf of the  $n^{\text{th}}$  order statistic from the random sample  $X_1, \dots, X_n$ , where

$$f(x) = 1/\theta, \quad 0 < x \leq \theta; \quad 0 < \theta < \infty$$

(and zero elsewhere) is

$$g_n(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 < y \leq \theta$$

and zero elsewhere. Find the limiting distribution  $G(y)$ .

Answer. First, we need to find  $G_n(y)$ .

$$\begin{aligned} G_n(y) &= \int_0^y \frac{nz^{n-1}}{\theta^n} dz = \left(\frac{z}{\theta}\right)^n \Big|_{z=0}^y = \left(\frac{y}{\theta}\right)^n, \quad 0 < y < \theta \\ &= 1, \quad y \geq \theta. \end{aligned}$$

Now the limiting distribution is

$$\begin{aligned} G(y) &= \lim_{n \rightarrow \infty} G_n(y) = 0, \quad 0 < y < \theta \\ &= 1, \quad y \geq \theta. \end{aligned}$$

Therefore,  $G(y)$  is a degenerate pdf with all the mass at  $Y = \theta$ .

### 3.3 Central Limit Theorem

Question. What is the limiting distribution of  $\bar{X}_n$ ?

Answer. A spike at  $\mu$ .

Consider a stabilizing transformation of  $\bar{X}_n$ :

$$Y = \sqrt{n}(\bar{X}_n - \mu).$$

Definition. Let  $X_1, \dots, X_n$  denote a random sample from any distribution with finite mean  $\mu$  and finite variance  $\sigma^2$ . Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

Gauss example S.2 shows the CLT in action.

### 3.4 Asymptotic Distributions

Definition. An asymptotic distribution is used to approximate a true (and possibly unknown) finite-sample distribution.

Notes:

1. The mean and variance of an asymptotic distribution are called the asymptotic mean and asymptotic variance.
2.  $\hat{\theta}$  is said to be asymptotically efficient if  $asy.var.(\hat{\theta})$  is less than or equal to the asymptotic variance of any other consistent estimator.
3. Occasionally you will hear the term asymptotically unbiased:  $\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta$ .

Example #1. Consider the random variable

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

We say that

$$\bar{X}_n \overset{asy}{\sim} N(\mu, \sigma^2/n).$$

Example #2. Find the asymptotic distribution of  $Z_n = n(1 - Y_n)$ , where  $Y_n$  is the  $n^{th}$  order statistic from the *uniform*[0, 1] random sample  $X_1, \dots, X_n$ .

Answer. Start by finding the limiting distribution of  $Y_n$ :

$$\begin{aligned} G(y_n) &= 0, \quad 0 \leq y_n < 1 \\ &= 1, \quad y_n = 1. \end{aligned}$$

Therefore,  $Y_n$  has a degenerate limiting pdf with all the mass at  $Y_n = 1$ .

The pdf for  $Z_n$  can be found by the change of variable technique:

$$h_n(z_n) = (1 - z/n)^{n-1}, \quad 0 < z < n$$

and zero elsewhere. The cdf for  $Z_n$  is

$$\begin{aligned} H_n(z_n) &= 0, \quad z < 0 \\ &= \int_0^{z_n} (1 - w/n)^{n-1} dw = 1 - (1 - z_n/n)^n, \quad 0 \leq z < n \\ &= 1, \quad z \geq n \end{aligned}$$

and its limiting distribution  $H(z_n)$  is

$$\begin{aligned}\lim_{n \rightarrow \infty} H_n(z_n) &= 0, \quad z < 0 \\ &= 1 - e^{-z}, \quad 0 \leq z < \infty.\end{aligned}$$

Therefore,  $Z_n \stackrel{asy}{\sim} \text{exponential}(\lambda = 1)$ .

## 4 Maximum Likelihood Estimation

Example #1. Consider the random sample  $\{X_1 = 0.5, X_2 = 2.0, X_3 = 10.0, X_4 = 1.5, X_5 = 7.0\}$  generated from an exponential distribution. What is the maximum likelihood (ML) estimator of  $\beta$ ?

Answer. Begin by forming the likelihood function,  $L(\theta)$ :

$$\begin{aligned}L &= f(x_1, x_2, x_3, x_4, x_5; \beta) \\ &= \prod_{i=1}^5 f(x_i) = \prod_{i=1}^5 \frac{1}{\beta} \exp(-x_i/\beta) = \frac{1}{\beta^5} \exp\left(\sum_{i=1}^5 -x_i/\beta\right)\end{aligned}$$

where  $\theta = 1/\beta$ . It is often more convenient to work with the monotonic transformation:

$$\begin{aligned}\ln L(\theta) &= \ln(\theta^5) - \theta(x_1 + x_2 + x_3 + x_4 + x_5) \\ &= 5 \ln(\theta) - 21\theta.\end{aligned}$$

The ML estimator of  $\theta$ ,  $\hat{\theta}$ , is the value of  $\theta$  that maximizes  $L(\theta)$  or  $\ln L(\theta)$ . Now we calculate  $\hat{\theta}$ .

$$\frac{d \ln L(\theta)}{d\theta} = \frac{5}{\theta} - 21 = 0 \implies \hat{\theta} = 5/21 \implies \hat{\beta} = 4.2.$$

Next, we check the second-order condition to ensure that  $\hat{\theta} = 5/21$  is indeed a maximum.

$$\frac{d^2 \ln L(\theta)}{d\theta^2} = -5\theta^{-2} < 0.$$

Therefore,  $\hat{\beta} = 4.2$  is the maximum likelihood estimator of  $E(X) = \beta$ . See [Gauss example S.3](#) for further details.

Notes:

1. The information number is  $I(\theta) = -E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] = E \left[ \left( \frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right]$ .
2. The information matrix is  $I(\theta) = -E \left[ \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'} \right] = E \left[ \frac{\partial \ln L(\theta)}{\partial \theta} \frac{\partial \ln L(\theta)}{\partial \theta'} \right]$  where  $\theta = \{\theta_1, \dots, \theta_k\}'$  is a  $(k \times 1)$  column vector.
3. The Cramer-Rao lower bound,  $I(\theta)^{-1}$ , is the lowest value the variance of an unbiased estimator  $\hat{\theta}$  can attain, given certain regularity conditions are satisfied.

Example #2. Find the ML estimators for  $\mu$  and  $\sigma^2$  from a normal distribution. Let  $X_1, \dots, X_n$  be a random sample from  $N(\mu, \sigma^2)$ .

$$L(\mu, \sigma^2) = \prod_{i=1}^n \left[ (2\pi\sigma^2)^{-0.5} \exp \left\{ -\left(\frac{1}{2\sigma^2}\right)(x_i - \mu)^2 \right\} \right].$$

Taking natural logs:

$$\ln L(\mu, \sigma^2) = -0.5n \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

- First take partial derivatives with respect to  $\mu$  and  $\sigma^2$ :

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu); & \frac{\partial \ln L(\theta)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \\ \frac{\partial^2 \ln L(\theta)}{\partial \mu^2} &= -\frac{n}{\sigma^2}; & \frac{\partial^2 \ln L(\theta)}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu); & \frac{\partial^2 \ln L(\theta)}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

- Now set first derivatives equal to zero and solve for the ML estimators:

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \implies \hat{\mu} = \bar{X} \\ \frac{\partial \ln L(\theta)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2. \end{aligned}$$

- Cramer-Rao Lower Bound  $\theta = (\mu, \sigma^2)'$ . The information matrix is

$$I(\theta) = -E \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu) \\ -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2 \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

and the CRLB is

$$I(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}.$$

- Question. Are  $\bar{X}$ ,  $s^2$  and  $\hat{\sigma}^2$  efficient estimators?

Answer. Recall,  $E(\bar{X}) = \mu$ ,  $E(s^2) = \sigma^2$  and  $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$ .

- $var(\bar{X}) = \sigma^2/n \implies \bar{X}$  is a minimum variance linear unbiased estimator.
- $var(s^2) = 2\sigma^4/(n-1) \implies s^2$  may or may not be unbiased efficient.
- $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$  and  $asy.var.(\hat{\sigma}^2) = 2\sigma^4/n \implies \hat{\sigma}^2$  is asymptotically efficient.

Properties of ML Estimators (under regularity).

1.  $\hat{\theta}_{ML} \xrightarrow{p} \theta$ .
2.  $\hat{\theta}_{ML} \stackrel{asy}{\sim} N(\theta, I^{-1}(\theta))$ .
3.  $\hat{\theta}_{ML}$  achieves the CRLB and is therefore asymptotically efficient.
4. Invariance (i.e.,  $\gamma = g(\theta) \implies \hat{\gamma}_{ML} = g(\hat{\theta}_{ML})$ ).

Notes:

The asymptotic covariance matrix of  $\hat{\theta}_{ML}$  is often hard or impossible to estimate. Three possible (asymptotically equivalent) estimators are:

1.  $I^{-1}(\hat{\theta}_{ML})$ , which is often not feasible.
2.  $-\left(\frac{\partial^2 \ln L(\hat{\theta})}{\partial \theta \partial \theta'}\right)^{-1}$ , which is sometimes quite complicated.
3. BHHH estimator:

$$\left( \sum_{i=1}^n \frac{\partial \ln f(x_i, \hat{\theta})}{\partial \theta} \frac{\partial \ln f(x_i, \hat{\theta})'}{\partial \theta} \right)^{-1}.$$

## 5 Method of Moments

Definition. Let  $X_1, \dots, X_n$  be a random sample from  $f(x; \theta_1, \dots, \theta_r)$ . Let  $M_k = \sum_i X_i^k/n$  be the  $k^{th}$  sample moment and  $E(X^k)$  be the  $k^{th}$  population (uncentered) moment. The method of moments estimator for  $\theta = (\theta_1, \dots, \theta_r)'$  is obtained by solving the  $r$  equations  $E(X^i) = M_i$ ,  $i = 1, \dots, r$ .

Notes:

1. Method of moments may also use  $E((X - \mu)^k)$  or other functions  $\gamma_k(\theta)$  of the unknown parameters.
2. Method of moments estimators are NOT typically efficient.
3. Method of moments estimators are typically consistent by virtue of Slutsky's theorem.

Example. Suppose  $X_1, \dots, X_n$  is a random sample from a  $gamma(\alpha, \beta)$  distribution. The likelihood function is

$$L(\theta) = (\Gamma(\alpha)\beta^\alpha)^{-n} (x_1 x_2 \cdots x_n)^{\alpha-1} \exp\left[-\sum_{i=1}^n x_i/\beta\right]$$

is difficult to evaluate without using numerical methods. Consider instead the following two moments

$$\begin{aligned} M_1 &= \bar{X} = \alpha\beta \\ M_2 - M_1^2 &= \hat{\sigma}^2 = \alpha\beta^2. \end{aligned}$$

Using these two sample moment equations to solve for  $\hat{\alpha}$  and  $\hat{\beta}$  gives

$$\hat{\alpha} = \bar{X}^2/\hat{\sigma}^2 \text{ and } \hat{\beta} = \hat{\sigma}^2/\bar{X}.$$

## 5.1 Variance of the Method of Moments Estimator

Let the sample moments be

$$\bar{g}_k = \frac{1}{n} \sum_{i=1}^n g_k(X_i) \text{ for } k = 1, \dots, K$$

and  $g = (g_1, \dots, g_k)'$  have asymptotic covariance matrix

$$V = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n [(g_j(X_i) - \bar{g}_j))(g_k(X_i) - \bar{g}_k)']$$

for  $j, k = 1, \dots, K$ . Now let  $G$  be the matrix

$$\begin{bmatrix} \frac{\partial \bar{g}_1}{\partial \theta_1} & \frac{\partial \bar{g}_1}{\partial \theta_2} & \cdots & \frac{\partial \bar{g}_1}{\partial \theta_k} \\ \frac{\partial \bar{g}_2}{\partial \theta_1} & \frac{\partial \bar{g}_2}{\partial \theta_2} & & \frac{\partial \bar{g}_2}{\partial \theta_k} \\ \vdots & & \ddots & \vdots \\ \frac{\partial \bar{g}_k}{\partial \theta_1} & \frac{\partial \bar{g}_k}{\partial \theta_2} & \cdots & \frac{\partial \bar{g}_k}{\partial \theta_k} \end{bmatrix}_{k \times k}.$$

Consider the first-order Taylor approximation to  $\bar{g}_k = \gamma_k(\theta)$  around  $\theta$

$$\bar{g} \simeq \gamma(\theta) + G(\theta)(\hat{\theta} - \theta) \implies (\hat{\theta} - \theta) \simeq G^{-1}(\theta)(\bar{g} - \gamma(\theta)).$$

Using the CLT, we know

$$\widehat{asy.var.}(\hat{\theta}) = (\hat{G}^{-1})V(\hat{G}^{-1})'$$

Example continued.

Let  $\theta = (\alpha, \beta)'$ ,  $g_1 = X - \mu$ ,  $g_2 = (X - \mu)^2$ ,

$$\hat{G} = \frac{\partial \bar{g}}{\partial \theta'} = \begin{bmatrix} \hat{\beta} & \hat{\alpha} \\ \hat{\beta}^2 & 2\hat{\alpha}\hat{\beta} \end{bmatrix}$$

and

$$V = \frac{1}{n} \begin{bmatrix} \widehat{var}(g_1) & \widehat{cov}(g_1, g_2) \\ \widehat{cov}(g_1, g_2) & \widehat{var}(g_2) \end{bmatrix}.$$

Then the estimated  $asy.var.(\hat{\theta}) = (\hat{G}^{-1})V(\hat{G}^{-1})'$ .

## 6 Interval Estimation

Definition. An interval estimate is found by algebraic manipulation of a pivotal quantity – a quantity based on the point estimate and the parameter – subject to a desired confidence coefficient.

Example #1. Find the 90% interval estimate for  $\mu$  from a random  $N(\mu, \sigma^2)$  sample ( $n = 25$ ) with  $\bar{X} = 50$  and (i)  $\sigma^2 = 100$  and (ii)  $s^2 = 100$ .

Answer.

(i) We know that  $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$ . This implies that

$$\begin{aligned} \Pr(-z \leq \sqrt{n}(\bar{X} - \mu)/\sigma \leq z) = 0.90 &\implies \Pr(-z \leq 0.5(50 - \mu) \leq z) = 0.90 \\ &\implies \Pr(-z - 25 \leq -0.5\mu \leq z - 25) = 0.90 \\ &\implies \Pr(2z + 50 \geq \mu \geq 50 - 2z). \end{aligned}$$

From Table 1 in Greene, we know that  $z = 1.645$ . Therefore, the 90% confidence interval for  $\mu$  is

[46.71, 53.29].

(ii) We know that  $Z = \sqrt{n}(\bar{X} - \mu)/s \sim t(n - 1)$ . From Table 2 in Greene, we know that  $z = 1.711$ .

Therefore, the 90% confidence interval for  $\mu$  is [46.58, 53.42].

Example #2. Find the confidence coefficient for the  $\Pr(1 \leq \sigma^2 \leq 2)$  if  $s^2 = 1.5$  from a normally distributed random sample ( $n = 25$ ).

Answer. To solve this problem, we need to recognize that  $(n - 1)s^2/\sigma^2 \sim \chi^2(n - 1)$  and use Table 3 in Greene:

$$\begin{aligned}\Pr(1 \leq \sigma^2 \leq 2) &= \Pr\left(\frac{1}{(n-1)s^2} \leq \frac{\sigma^2}{(n-1)s^2} \leq \frac{2}{(n-1)s^2}\right) \\ &= \Pr\left(\frac{(n-1)s^2}{2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \frac{(n-1)s^2}{1}\right) \\ &= \Pr\left(18 \leq \frac{(n-1)s^2}{\sigma^2} \leq 36\right) \simeq 0.70.\end{aligned}$$

## 7 Hypothesis Testing

There are two principal areas of statistical inference:

1. parameter estimation (already covered) and
2. hypothesis testing.

General Methodology for Classical (Neyman-Pearson) Hypothesis Testing.

1. State the null ( $H_0: \theta = \theta_0$ ) and alternative hypotheses.
2. Determine the size of the critical region.
3. State the decision rule.
4. Calculate the statistic.
5. Make a decision (i.e., reject or fail to reject the null).
6. Consider possible errors.

## 7.1 Concepts

1. Type I Error. Reject true null hypothesis. The probability of a type I error is called the size of the test.
2. Type II Error. Fail to reject false hypothesis. One minus the probability of a type II error is called the power of the test.
3. Power Function. The power function yields the probability that the sample point falls in the critical region, given that the true value of  $\theta$  is not  $\theta_0$ .
4. Certain Best Tests. Assuming a simple alternative,  $C$  is the best critical region of size  $\alpha$  for testing  $H_0: \theta = \theta'$  versus  $H_1: \theta = \theta''$  if for every region  $A$  such that  $\Pr[u(X_1, \dots, X_n) \in A] = \alpha$ ,
  - $\Pr[u(X_1, \dots, X_n) \in C|H_0] = \alpha$
  - $\Pr[u(X_1, \dots, X_n) \in C|H_1] \geq \Pr[u(X_1, \dots, X_n) \in A|H_1]$ .
5. Uniformly Most Powerful Tests. Assuming a composite alternative, a test is uniformly most powerful if  $C$  is the best critical region of size  $\alpha$  for testing each simple hypothesis in  $H_1$ . In other words, the power function is no less than for any other test of equal size.

## 7.2 Tests Based on Confidence Intervals

Consider the following test:

- (1) Reject  $H_0: \theta = \theta_0$  if  $\theta_0$  falls outside  $[\theta_L, \theta_U]$ .
- (2) Accept  $H_1: \theta \neq \theta_0$  (i.e., fail to reject  $H_0$ ) if  $\theta_0$  falls inside  $[\theta_L, \theta_U]$ .

Example #1(i) continued from Section 6.

Consider the following test

$$H_0 : \mu = 48$$

$$H_1 : \mu \neq 48.$$

The 90% confidence interval is  $[46.71, 53.29]$ . The decision is to

- Reject  $H_0$  if  $46.71 \geq \mu_0 \geq 53.29$ .

- Fail to reject  $H_0$  if  $46.71 < \mu_0 < 53.29$ .

Therefore, we fail to reject the hypothesis  $H_0: \mu = 48$ .

### 7.3 Likelihood Ratio, Wald and Lagrange Multiplier Tests

The likelihood ratio (LR), Wald (W) and Lagrange multiplier (LM) tests are asymptotically equivalent tests that may produce different results in small samples. When no other information exists, you can choose the test that is the easiest to compute. See the attached figure for a graphical representation of each test.

#### 7.3.1 Likelihood Ratio Test

Let  $\hat{\theta}_R$  ( $\hat{\theta}_U$ ) and  $\hat{L}_R$  ( $\hat{L}_U$ ) be the restricted (unrestricted) estimate and likelihood value, respectively. Let the null and alternative hypotheses be

$$H_0 : c(\theta) = q$$

$$H_1 : c(\theta) \neq q.$$

The likelihood ratio is defined as

$$\lambda = \hat{L}_R / \hat{L}_U$$

where  $0 \leq \lambda \leq 1$ . The LR statistic is then

$$LR = -2 \ln \lambda \stackrel{asy}{\sim} \chi^2(r)$$

where  $r$  is the number of restrictions imposed.

#### 7.3.2 Wald Test

In the LR test, one needs to calculate  $\hat{L}_U$  and  $\hat{L}_R$ . An advantage of the Wald test is that  $\hat{\theta}_R$  does not need to be calculated. The Wald statistic is

$$W = (c(\hat{\theta}_U) - q)' var(c(\hat{\theta}_U) - q)^{-1} (c(\hat{\theta}_U) - q) \stackrel{asy}{\sim} \chi^2(r).$$

If  $c(\hat{\theta})$  is normally distributed, then  $W$  is a quadratic form in a normal vector and is distributed chi-square for all sample sizes.

Notes:

1. Because  $c(\hat{\theta})$  is often nonlinear,  $\text{var}(c(\hat{\theta}) - q)$  can be approximated by  $\text{var}(c(\hat{\theta}) - q) \simeq C \text{var}(\hat{\theta}) C'$  where  $C = \partial c(\hat{\theta}) / \partial \hat{\theta}'$ .
2. The power may be low because the alternative does not appear in computations.
3. Wald test is not invariant to the form of the restriction (e.g.,  $H_0: \theta_1/\theta_2 = c$  versus  $H_0: \theta_1 = c\theta_2$ ).
4. Wald test does not rely on strong distributional assumptions like the LR or LM.

### 7.3.3 Lagrange Multiplier Test

This test is based on the restricted model.

Derivation. Begin by forming the Lagrangian:

$$\ln L^*(\theta) = \ln L(\theta) + \lambda'(c(\theta) - q).$$

The first-order conditions are

$$\begin{aligned} \frac{\partial \ln L^*}{\partial \theta} &= \frac{\partial \ln L(\theta)}{\partial \theta} + \frac{\partial c(\theta)}{\partial \theta} \lambda = 0 \\ \frac{\partial \ln L^*}{\partial \lambda} &= c(\theta) - q = 0. \end{aligned}$$

At  $\hat{\theta}_R$ ,

$$\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R} = -\frac{\partial c(\hat{\theta}_R)}{\partial \hat{\theta}_R} \hat{\lambda} = \hat{g}_R.$$

If  $H_0: c(\theta) = q$  is correct,  $\hat{g}_R = 0$ . This fact is used as motivation for

$$LM = \hat{g}'_R I^{-1}(\hat{\theta}_R) \hat{g}_R \stackrel{asy}{\sim} \chi^2(r).$$

### 7.3.4 An Example Using the LR, W and LM Tests

Consider an artificial random sample ( $n = 100$ ) from an *exponential* ( $\beta = 0.1$ ) distribution. The log likelihood function is

$$\ln L(\theta) = n \ln(\theta) - \theta \sum_{i=1}^n x_i$$

where  $\theta = 1/\beta$ . The first-order condition and unrestricted ML estimator is

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 \implies \hat{\theta}_U = \bar{X}^{-1}.$$

The second-order condition is

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2} < 0$$

so  $\hat{\theta}_U$  is indeed a maximum.

Now consider testing the following hypothesis

$$H_0 : \theta = 7.5$$

$$H_1 : \theta \neq 7.5$$

so that  $\hat{\theta}_R = 7.5$

### 1. Likelihood Ratio Test

The likelihood values are

$$\begin{aligned} \hat{L}_U &= \hat{\theta}_U^{100} \exp(-\hat{\theta}_U \sum_{i=1}^n x_i) \\ \hat{L}_R &= \hat{\theta}_R^{100} \exp(-\hat{\theta}_R \sum_{i=1}^n x_i) \end{aligned}$$

and the LR statistic is

$$LR = -2 \ln(\hat{L}_R / \hat{L}_U).$$

### 2. Wald Test

The Wald statistic is

$$W = \frac{(\hat{\theta}_U - 7.5)^2}{\text{var}(\hat{\theta}_U - 7.5)} = \frac{(\hat{\theta}_U - 7.5)^2}{\text{var}(\hat{\theta}_U)}$$

where  $\text{var}(\hat{\theta}_U) = \hat{I}^{-1}(\hat{\theta}_U) = -\left(\frac{\partial^2 \ln L(\hat{\theta}_U)}{\partial \hat{\theta}_U^2}\right)^{-1} = \hat{\theta}_U^2/n$ .

### 3. Lagrange Multiplier Test

The LM statistic is

$$LM = \frac{\hat{g}_R^2}{I(\hat{\theta}_R)}$$

where

$$\hat{g}_R = \frac{n}{\hat{\theta}_R} - \sum_{i=1}^n x_i \text{ and } I(\hat{\theta}_R) = n/\hat{\theta}_R^2.$$

Finally, the critical region is defined by the chi-square critical value with  $r = 1$  degrees of freedom and a 95% confidence level. Using Table 3 in Greene, the critical value is 3.84. Therefore,

- If  $LR$ ,  $W$  or  $LM$  is greater than 3.84, we reject the null  $H_0: \theta = 7.5$  in favor of the alternative.
- If  $LR$ ,  $W$  or  $LM$  is less than or equal to 3.84, we fail to reject the null  $H_0: \theta = 7.5$ .

See Gauss example S.4 for further details.

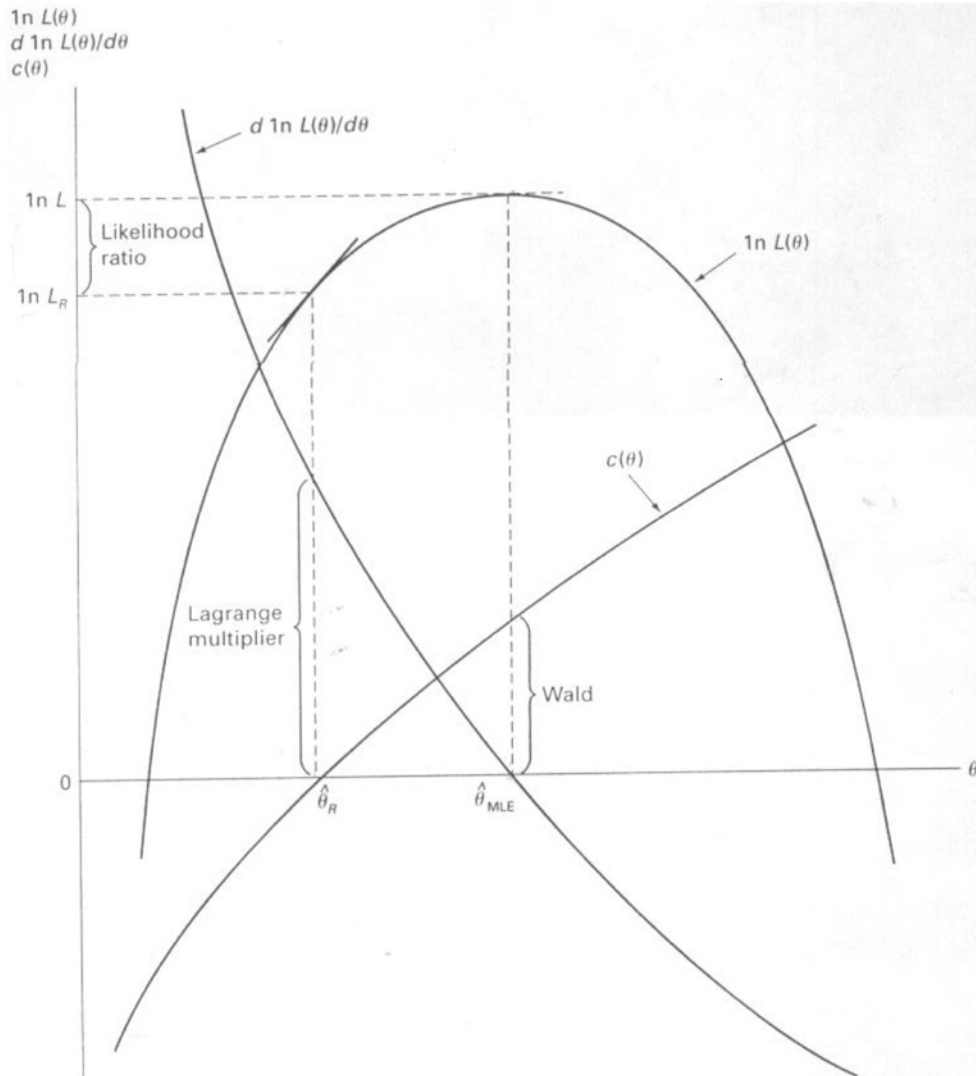


FIGURE 4.8 Three Bases for Hypothesis Tests.

function  $\ln L(\theta)$ , its derivative with respect to  $\theta$ ,  $d \ln L(\theta)/d\theta$ , and the constraint  $c(\theta)$ . There are three approaches to testing the hypothesis suggested in the figure:

- **Likelihood ratio test.** If the restriction  $c(\theta) = 0$  is valid, imposing it should not lead to a large reduction in the log-likelihood function. Therefore, we base the test on the difference,  $\ln L - \ln L_R$ , where  $L$  is the value of the likelihood function at the unconstrained value of  $\theta$  and  $L_R$  is the value of the likelihood function at the restricted estimate.