

## TESTS BASED ON MAXIMUM LIKELIHOOD

### 1. The Basic Example.

To illustrate the properties of maximum likelihood estimates and tests, we consider the simplest possible case of estimating the *mean* of the normal distribution with known variance,  $\sigma^2$ . Given a random sample  $x = (x_1, \dots, x_n)$  from a normal distribution,  $N(\mu, \sigma^2)$ , we first derive the *maximum-likelihood estimate* of  $\mu$  with  $\sigma^2$  known:

$$\begin{aligned}
 (1.1) \quad L_n(\mu | x, \sigma^2) &= \ln \left( \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} \right) \\
 &= \sum_{i=1}^n \left[ \ln \left( \frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2} \left( \frac{x_i - \mu}{\sigma} \right)^2 \right] \\
 &= -n \ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2
 \end{aligned}$$

So by solving the *first-order condition* for  $\mu$  we obtain:

$$\begin{aligned}
 (1.2) \quad 0 &= \frac{dL_n}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = \frac{1}{\sigma^2} \sum_{i=1}^n x_i - \frac{n}{\sigma^2} \mu \\
 &\Rightarrow \sum_{i=1}^n x_i - n\mu = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}_n
 \end{aligned}$$

## 2. Covariance of Maximum Likelihood Estimators

For this simple case, it is well known that the *variance* of the sample mean,  $\hat{\mu}_n (= \bar{X}_n)$  is given by  $\text{var}(\hat{\mu}_n) = \sigma^2 / n$ . But by taking the second derivative of  $L_n$  with respect to  $\mu$ , and evaluating at the true mean value, say  $\mu_0$  we see that

$$(2.1) \quad \frac{d^2 L_n(\mu_0)}{d\mu^2} = -\frac{n}{\sigma^2} \Rightarrow \text{var}(\hat{\mu}_n) = \left( -\frac{d^2 L_n(\mu_0)}{d\mu^2} \right)^{-1}$$

Since in this case,  $d^2 L_n / d\mu^2$  is *independent* of the data,  $(x_1, \dots, x_n)$ , it follows that  $E[d^2 L_n(\mu_0) / d\mu^2] \equiv -n / \sigma^2$ , so that in fact

$$(2.2) \quad \text{var}(\hat{\mu}) = \left( -E \left[ \frac{d^2 L_n(\mu_0)}{d\mu^2} \right] \right)^{-1} \equiv \mathcal{I}_n(\mu_0)^{-1}$$

where  $\mathcal{I}_n(\mu_0)$  is designated as *Fisher Information* about  $\mu_0$  (and is seen to increase as the variance of the maximum-likelihood estimator,  $\hat{\mu}_n$ , decreases). Finally, since the Law of Large Numbers shows in this case that  $\hat{\mu}_n = \bar{X}_n \approx \mu_0$  for  $n$  sufficiently large, we may conclude from (2.2) that

$$(2.3) \quad \text{var}(\hat{\mu}_n) \approx \mathcal{I}_n(\hat{\mu}_n)^{-1}$$

More generally, for any parameter vector,  $\theta = (\theta_1, \dots, \theta_k)'$ , defining a “well behaved” distribution with likelihood function,  $L_n(\theta)$ , it can be shown that if  $\theta_0$  denotes the true value of  $\theta$ , then the *covariance matrix* of the maximum-likelihood estimator,  $\hat{\theta}_n$ , is well approximated for large  $n$  by

$$(2.3) \quad \text{cov}(\hat{\theta}_n) \approx \left(-E[\nabla_{\theta\theta} L_n(\theta_0)]\right)^{-1} \equiv \mathcal{I}_n(\theta_0)^{-1}$$

Moreover, as with the simple case of the normal mean above, it can also be shown that  $\hat{\theta}_n$  is always a *consistent* estimator of  $\theta_0$ , so that  $\hat{\theta}_n \approx \theta_0$  for large  $n$ . Thus as an extension of (2.3) we obtain the sample approximation:

$$(2.4) \quad \text{cov}(\hat{\theta}_n) \approx \mathcal{I}_n(\hat{\theta}_n)^{-1}$$

### 3. Wald Tests of Parameters

For the simple case of the normal mean above it follows at once (from the fact that linear combinations of independent normals are normal) that:

$$(3.1) \quad \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

This in turn implies from (1.2) and (2.2) that

$$(3.2) \quad \hat{\mu}_n \sim N\left(\mu, \mathcal{I}_n(\hat{\mu}_n)^{-1}\right)$$

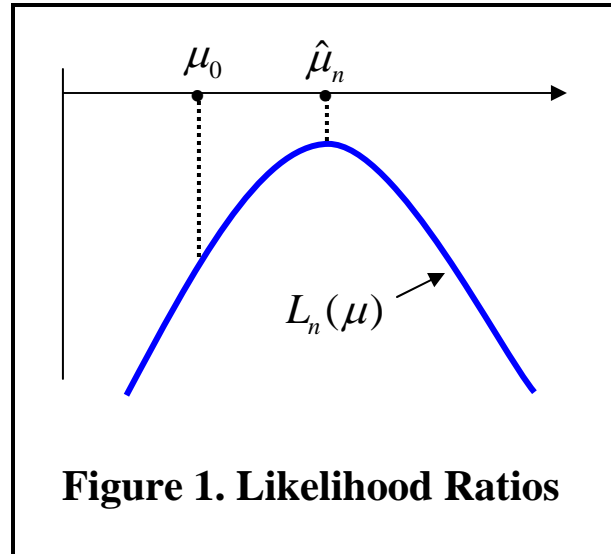
More generally it can be shown that for large  $n$  the distribution of the general maximum-likelihood estimator,  $\hat{\theta}_n$ , defined above is well approximated by

$$(3.3) \quad \hat{\theta}_n \sim N\left(\theta_0, \mathcal{I}_n(\hat{\theta}_n)^{-1}\right)$$

where  $\theta_0$  again denotes the true value of  $\theta$ . This forms the basis for all standard tests of hypotheses about the components of  $\theta$ .

#### 4. Likelihood-Ratio Tests of Parameters

Another way to test the hypothesis that, say  $\mu_0$ , is the true value of  $\mu$  is simply to compare the likelihood of  $\mu_0$  with that of the most likely value,  $\hat{\mu}_n$ . If the resulting *ratio* of likelihood values is close to one, then this implies that  $\mu_0$  is a good candidate for the true value of  $\mu$ . In terms of log likelihoods, this is in turn equivalent to a difference,  $L_n(\hat{\mu}_n) - L_n(\mu_0)$ , close to zero.



If we observe from (1.1) that

$$(4.1) \quad L_n(\mu_0 | x, \sigma^2) = -n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

then this difference can be evaluated as:

$$\begin{aligned}
 (4.2) \quad L_n(\hat{\mu}_n) - L_n(\mu_0) &= -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (x_i - \hat{\mu}_n)^2 - \sum_{i=1}^n (x_i - \mu_0)^2 \right] \\
 &= -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2) - \sum_{i=1}^n (x_i^2 - 2x_i\mu_0 + \mu_0^2) \right] \\
 &= -\frac{1}{2\sigma^2} \left[ -2\bar{x}_n \sum_{i=1}^n x_i + n\bar{x}_n^2 + 2\mu_0 \sum_{i=1}^n x_i - n\mu_0^2 \right] \\
 &= \frac{n}{2\sigma^2} \left[ \bar{x}_n^2 - 2\mu_0\bar{x}_n + \mu_0^2 \right] \\
 &= \frac{n}{2\sigma^2} (\bar{x}_n - \mu_0)^2
 \end{aligned}$$

Hence it follows that

$$(4.3) \quad 2[L_n(\hat{\mu}_n) - L_n(\mu_0)] = \left( \frac{\bar{x}_n - \mu_0}{\sigma / \sqrt{n}} \right)^2$$

But under the null hypothesis that  $H_0 : \mu = \mu_0$  it follows that

$$(4.4) \quad \frac{\bar{x}_n - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

and hence by definition that

$$(4.5) \quad \left( \frac{\bar{x}_n - \mu_0}{\sigma / \sqrt{n}} \right)^2 \sim \chi_1^2$$

Thus we see that under  $H_0$ :

$$(4.6) \quad 2[L_n(\hat{\mu}_n) - L_n(\mu_0)] \sim \chi_1^2$$

More generally, for any partition of parameters,  $\theta = (\theta^1, \theta^2)$ , the null hypothesis,  $H_0 : \theta^1 = \theta_0^1$ , can be tested in the same way by letting  $\hat{\theta}_0^2$  be defined by the condition,

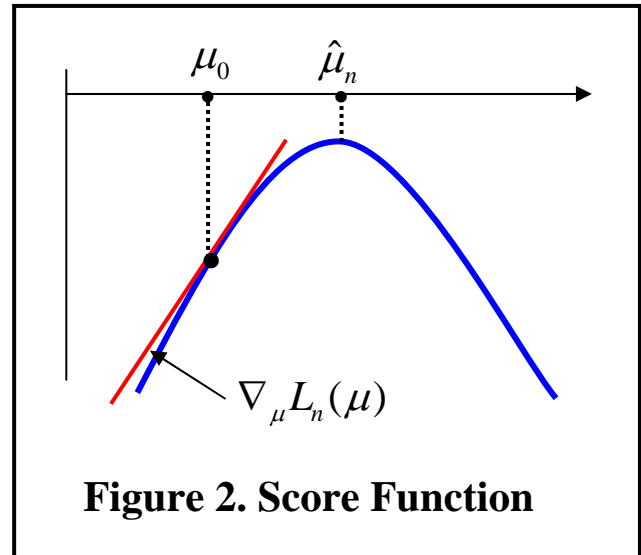
$$(4.7) \quad L_n(\theta_0^1, \hat{\theta}_0^2) = \max_{\theta^2} L_n(\theta_0^1, \theta^2)$$

and comparing  $L_n(\theta_0^1, \hat{\theta}_0^2)$  with the unconstrained maximum,  $L_n(\hat{\theta}_n)$ . If  $\theta_0^1$  has  $k$  components, then under  $H_0$  we now have

$$(4.8) \quad 2[L_n(\hat{\theta}_n) - L_n(\theta_0^1, \hat{\theta}_0^2)] \sim \chi_k^2$$

## 5. Lagrange Multiplier (Score) Tests of Parameters

A final way to test the hypothesis that  $\mu_0$  is the true value of  $\mu$  is to simply examine the *slope* of the likelihood function at  $\mu_0$ . If this slope is close to zero, then it follows (by continuity) that  $\mu_0$  must be “close” to,  $\hat{\mu}_n$ , and thus must again be a good candidate for the true value. Hence one may simply test whether the slope,  $\nabla_{\mu} L_n(\mu_0)$ , is significantly different from zero.



But by (1.1) it follows that the *score function*,  $s_n(\mu_0) = \nabla_{\mu} L_n(\mu_0)$  is given by

$$\begin{aligned}
 (5.1) \quad s_n(\mu_0) &= \nabla_{\mu} \left\{ -n \ln(\sigma \sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}_{\mu=\mu_0} \\
 &= \frac{2}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0) = \frac{1}{\sigma^2} \left( \sum_{i=1}^n x_i - n\mu_0 \right) \\
 &= \frac{n}{\sigma^2} \left( \frac{1}{n} \sum_{i=1}^n x_i - \mu_0 \right) = \frac{\bar{x}_n - \mu_0}{\sigma^2/n}
 \end{aligned}$$

Hence it follows that

$$(5.2) \quad \left( \sigma / \sqrt{n} \right) s_n(\mu_0) = \frac{\bar{x}_n - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

As in (4.4) and (4.5) above, this in turn implies that

$$(5.3) \quad (\sigma^2/n)s_n(\mu_0)^2 \sim \chi_1^2$$

Finally, recalling that  $\text{var}(\hat{\mu}_n) = \sigma^2/n$ , we obtain the following *score test statistic*:

$$(5.4) \quad \text{var}(\mu_0)s_n(\mu_0)^2 \sim \chi_1^2$$

where  $\text{var}(\mu_0)$  is formally the variance of  $\hat{\mu}_n$  under the null hypothesis (which is completely independent of  $\mu_0$  in this simple case).

It is important to note that this *score test* is also called a *Lagrange multiplier test*. To see the reason for this, observe that finding the maximum-likelihood estimate of  $\mu$  under the null hypothesis,  $\mu = \mu_0$ , is formally equivalent to the Lagrangian maximization problem:

$$(5.5) \quad \max_{\mu} \varphi(\mu) = L_n(\mu) + \lambda(\mu_0 - \mu)$$

with solution given by

$$(5.6) \quad 0 = \nabla_{\mu} \varphi = \nabla_{\mu} L_n(\mu) - \lambda$$

$$(5.7) \quad 0 = \nabla_{\lambda} \varphi = \mu_0 - \mu$$

By combining these two conditions, we see that the optimal value,  $\hat{\lambda}$ , of the Lagrange multiplier is given by,

$$(5.8) \quad \hat{\lambda} = \nabla_{\mu} L_n(\mu_0)$$

and thus that the score function,  $s_n(\mu_0) = \nabla_{\mu} L_n(\mu_0)$ , can also be interpreted as a Lagrange multiplier. However, the above slope interpretation seems to be simpler and more intuitive.

More generally, for any partition of parameters,  $\theta = (\theta^1, \theta^2)$ , as in section 4 above, the null hypothesis,  $H_0 : \theta^1 = \theta_0^1$ , can be tested in the same way by again letting  $\hat{\theta}_0^2$  be defined by (4.7) and setting  $\theta_0 = (\theta_0^1, \hat{\theta}_0^2)$ . If  $\theta^1$  is of dimension  $k$ , then the score vector,  $s_n(\theta_0)$ , also has dimension  $k$ , and the variance term  $\text{var}(\mu_0)$  in (5.4) is now replaced by the *covariance sub-matrix*,  $\text{cov}_{11}(\theta_0)$ , where

$$(5.9) \quad \text{cov}(\theta_0) = \begin{bmatrix} \text{cov}_{11}(\theta_0) & \text{cov}_{12}(\theta_0) \\ \text{cov}_{21}(\theta_0) & \text{cov}_{22}(\theta_0) \end{bmatrix}$$

If the *Fisher information matrix* is given by

$$(5.10) \quad \mathcal{I}_n(\theta_0) = \begin{bmatrix} \mathcal{I}_{n11}(\theta_0) & \mathcal{I}_{n12}(\theta_0) \\ \mathcal{I}_{n21}(\theta_0) & \mathcal{I}_{n22}(\theta_0) \end{bmatrix}$$

then it can easily be shown (by partitioned inverse identities) that

$$(5.11) \quad \text{cov}_{11}(\theta_0) \approx \left[ \mathcal{I}_n(\theta_0)^{-1} \right]_{11} = \left[ \mathcal{I}_{n11}(\theta_0) - \mathcal{I}_{n12}(\theta_0)\mathcal{I}_{n22}(\theta_0)\mathcal{I}_{n21}(\theta_0) \right]^{-1}$$

so that the following *score test statistic* is obtained as a direct generalization of (5.4):

$$(5.12) \quad s_n(\theta_0)' \left[ \mathcal{I}_n(\theta_0)^{-1} \right]_{11} s_n(\theta_0) \sim \chi_k^2$$

The single most important feature of this score test is that it only requires maximum-likelihood estimation *under the null hypothesis*, i.e.,

conditional maximum-likelihood estimation of  $\theta^2$  given  $\theta^1 = \theta_0^1$ . This is generally much easier to calculate. For example, if  $\theta^1 = \rho$  in the *SAR model*, or  $\theta^1 = \lambda$  in the *SL model*, then the remaining parameters  $\theta^2 = (\beta, \sigma^2)$  are directly obtainable from OLS.

## 6. Moran's $I$ as a Score Test for SAR

One key feature of score tests for our present purposes is that the score test statistic for  $\rho$  in the *SAR model* turns out to be precisely *Moran's  $I$*  statistic (up to a scale factor). To see this, observe first (from section 8.3 in the BULKPACK) that

$$(6.1) \quad \nabla_{\rho} L_n(\rho, \beta, \sigma^2) = \nabla_{\rho} \left\{ \text{const} + \ln |B_{\rho}| - \frac{1}{2\sigma^2} (y - X\beta)' B'_{\rho} B_{\rho} (y - X\beta) \right\}_{\rho=\rho_0}$$

$$= \nabla_{\rho} \left\{ \sum_{i=1}^n \ln(1 - \rho\omega_i) - \frac{1}{2\sigma^2} (y - X\beta)' B'_{\rho} B_{\rho} (y - X\beta) \right\}_{\rho=\rho_0}$$

where  $B_{\rho} = I_n - \rho W$  and where  $(\omega_i : i = 1, \dots, n)$  are the eigenvalues of  $W$ , so that

$$(6.2) \quad Wv_i = \omega_i v_i, \quad i = 1, \dots, n$$

To analyze (6.1) we first assume (for simplicity) that the eigenvectors in (6.2) are linearly independent so that the matrix,  $V = (v_i : i = 1, \dots, n)$ , is nonsingular. Hence if  $\Delta$  denotes the diagonal matrix of eigenvalues in (6.2) then by definition,

$$(6.3) \quad W[v_1, \dots, v_n] = [v_1, \dots, v_n]\Delta \Rightarrow WV = V\Delta \Rightarrow W = V\Delta V^{-1}$$

Next we observe that if the *trace* of a matrix  $A = (a_{ij} : i, j = 1, \dots, n)$  is defined by  $tr(A) = \sum_{i=1}^n a_{ii}$  then it follows that for any matrices,  $A = (a_1, \dots, a_n)'$  and  $B = (b_1, \dots, b_n)$ ,  $tr(AB) = \sum_{i=1}^n a_i' b_i = tr(BA)$ . Moreover, since  $tr(W) = 0$  for all *weight matrices*, it follows from (6.2) and (6.3) that

$$(6.4) \quad 0 = tr(W) = tr(V \Delta V^{-1}) = tr(\Delta V^{-1} V) = tr(\Delta) = \sum_{i=1}^n \omega_i$$

Hence for the single most important null hypothesis,  $\rho_0 = 0$ , we have

$$(6.5) \quad \nabla_{\rho} \left\{ \sum_{i=1}^n \ln(1 - \rho \omega_i) \right\}_{\rho=0} = - \left\{ \sum_{i=1}^n \frac{\omega_i}{1 - \rho \omega_i} \right\}_{\rho=0} = - \sum_{i=1}^n \omega_i = 0$$

and it follows that (6.1) reduces to

$$\begin{aligned} (6.6) \quad \nabla_{\rho} L_n(0, \beta, \sigma^2) &= - \frac{1}{2\sigma^2} \nabla_{\rho} \left\{ (y - X\beta)' B_{\rho}' B_{\rho} (y - X\beta) \right\}_{\rho=0} \\ &= - \frac{1}{2\sigma^2} \nabla_{\rho} \left\{ (y - X\beta)' (I_n - \rho W') (I_n - \rho W) (y - X\beta) \right\}_{\rho=0} \\ &= - \frac{1}{2\sigma^2} (y - X\beta)' \nabla_{\rho} \left\{ I_n - \rho(W' + W) + \rho^2 W' W \right\}_{\rho=0} (y - X\beta) \\ &= - \frac{1}{2\sigma^2} (y - X\beta)' \left\{ -(W' + W) + 2\rho W' W \right\}_{\rho=0} (y - X\beta) \\ &= \frac{1}{2\sigma^2} (y - X\beta)' (W' + W) (y - X\beta) \\ &= \frac{1}{\sigma^2} \left\{ \frac{1}{2} (y - X\beta)' W' (y - X\beta) + \frac{1}{2} (y - X\beta)' W (y - X\beta) \right\} \\ &= \frac{1}{\sigma^2} (y - X\beta)' W (y - X\beta). \end{aligned}$$

Hence if we let  $(\hat{\beta}_0, \hat{\sigma}_0^2)$  denote the OLS estimates of  $(\beta, \sigma^2)$  [i.e, the maximum-likelihood estimates under the hypothesis,  $\rho = 0$ ], so that in this case,  $\theta_0 = (0, \hat{\beta}_0, \hat{\sigma}_0^2)$ , then the score function is given by:

$$(6.7) \quad s_n(\theta_0) \equiv \nabla_{\rho} L_n(0, \hat{\beta}_0, \hat{\sigma}_0^2) = \frac{1}{\hat{\sigma}_0^2} (y - X \hat{\beta}_0)' W (y - X \hat{\beta}_0)$$

But if we now let  $\hat{\varepsilon}_0 = y - X \hat{\beta}_0$  denote the OLS residuals, so that

$$(6.8) \quad \hat{\sigma}_0^2 = \frac{1}{n} (y - X \hat{\beta}_0)' (y - X \hat{\beta}_0) = \frac{1}{n} \hat{\varepsilon}_0' \hat{\varepsilon}_0$$

then (6.7) takes the form;

$$(6.9) \quad s_n(\theta_0) = [n / \hat{\varepsilon}_0' \hat{\varepsilon}_0] (y - X \hat{\beta}_0)' W (y - X \hat{\beta}_0)$$

$$\Rightarrow \boxed{s_n(\theta_0) = n \left( \frac{\hat{\varepsilon}_0' W \hat{\varepsilon}_0}{\hat{\varepsilon}_0' \hat{\varepsilon}_0} \right)}$$

where the expression in brackets is precisely *Moran's I*.

To complete the argument, recall from expression (77) in the BULKPACK that the off-diagonal  $(\rho, \sigma^2)$  term in the Fisher information matrix,  $\mathcal{I}_n(\theta_0)$ , is given by

$$(6.10) \quad \begin{aligned} tr(G_{\rho}) &= tr[W(I_n - \rho W)^{-1}] = tr\left[W(I_n + \sum_{k=1}^{\infty} \rho^k W^k)\right] \\ &= tr\left(W + \rho \sum_{k=1}^{\infty} \rho^{k-1} W^k\right) \end{aligned}$$

Hence under the null hypothesis,  $\rho = 0$ , it follows that

$$(6.11) \quad tr(G_\rho)_{\rho=0} = tr(W) = 0$$

But this implies that  $\mathcal{I}_n(\theta_0)$  is *block diagonal*, so that in this simple case, (5.7) reduces to

$$(6.12) \quad cov_{11}(\theta_0) \approx [\mathcal{I}_n(\theta_0)^{-1}]_{11} = \mathcal{I}_{n11}(\theta_0)^{-1}$$

which is given from (77) in the BULKPACK by

$$(6.13) \quad \mathcal{I}_{n11}(\theta_0)^{-1} = \left( tr[G_\rho(G_\rho + G'_\rho)]_{\rho=0} \right)^{-1} = tr[W(W + W')]^{-1}$$

Hence the final *score statistic* is

$$(6.14) \quad s_n(\theta_0)' [\mathcal{I}_n(\theta_0)^{-1}]_{11} s_n(\theta_0) = \frac{n}{tr(WW + WW')} \left( \frac{\hat{\varepsilon}_0' W \hat{\varepsilon}_0}{\hat{\varepsilon}_0' \hat{\varepsilon}_0} \right)^2 \sim \chi_1^2$$

Note also that an equivalent test statistic, which is proportional to *Moran's I*, is obtained by taking the square root of (6.14):

$$(6.15) \quad \frac{\sqrt{n}}{\sqrt{tr(WW + WW')}} \left( \frac{\hat{\varepsilon}_0' W \hat{\varepsilon}_0}{\hat{\varepsilon}_0' \hat{\varepsilon}_0} \right) \sim N(0,1)$$