Gregory Kordas
Last update: June 13, 2023

Lecture 8
Maximum Likelihood Estimation and Inference

Abstract: We discuss maximum likelihood estimation and testing. We show that the MLE is consistent, asymptotically normal and attains the Cramer-Rao lower bound. We then introduce the classical likelihood ratio, Lagrange multiplier and Wald tests and show their asymptotic equivalence.

Likelihood is the central concept in statistical modelling and inference. Fisher coined the term 'likelihood' in 1921 to distinguish the method of maximum likelihood from the Bayesian or inverse probability argument. Uncertainty is pervasive in the real world, and statistics is the only branch of science that puts systematic effort into dealing with uncertainty. Statistics is suited to problems with inherent uncertainty due to limited information; it does not aim to remove uncertainty, but in many cases it merely quantifies it; uncertainty can remain even after an analysis is finished.

But how do we go from observed data to statements about the parameter of interest? The degree of certainty in an inductive conclusion is typically stronger than the degree in the data constituent, and the truth quality of the conclusion improves as we use more and more data. However, a single new item of information can destroy a carefully crafted conclusion; this aspect of inductive inference is ideal for mystery novels or courtroom dramas, but it can be a bane for practicing statisticians.

Data can be collected from observational studies rather than controlled experiments.

Example 1. If we let X be the number of germinated seeds in n and $\theta \in (0,1)$ be the probability that a seed germinates, the probability of $X = x$ germinations in n trial is given by

$$f_X(x;\theta) = P(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{1}$$

(i) Suppose 100 seeds are planted and 10 seeds germinate. The information about θ is given by the likelihood function

$$L(\theta) = P(X = 10) = \binom{100}{x}\theta^x(1-\theta)^{100-x}. \tag{2}$$

(ii) Suppose 100 seeds were planted and it is known only that less than 10 seeds germinated. The exact number of germinating seeds is unknown. Then the information about

θ is given by the likelihood function

$$L(\theta) = P(X \leq 10) = \sum_{x=0}^{10} \binom{100}{x} \theta^x (1-\theta)^{100-x}. \tag{3}$$

■

Example 2. To estimate the number of carps (N) living in a small lake, the Department of Fisheries tags $N_1 = 25$ of them. Later on it captures $n = 60$ carps, and finds $n_1 = 5$ tagged and $n_2 = 55$ untagged ones. Assuming the carps were caught at random, the likelihood of N can be computed based on the hypergeometric probability:

$$P(n_1) = \frac{\binom{N_1}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N}{n}} \tag{4}$$

so that

$$L(N) = P(n_1 = 5) = \frac{\binom{25}{5}\binom{N-25}{55}}{\binom{N}{60}} \tag{5}$$

The mle is given by

$$\frac{N_1}{\hat{N}} = \frac{n_1}{n} = \frac{25}{\hat{N}} = \frac{5}{60} \tag{6}$$

or $\hat{N} = 300$.

■

Let $z = \{z_i, i = 1, ..., n\}$ be a random sample from a distribution with density function $f(z;\theta)$, where $\theta = (\theta_1, ..., \theta_p)'$ is a p vector of parameters. For example, f could be a normal density and $\theta = (\mu, \sigma^2)$. The assumed random sampling process implies that the observations are i.i.d., so the joint probability density of the $n \times 1$ vector z is given by

$$f(z;\theta) = \prod_{i=1}^{n} f(z_i;\theta).$$

This density may be interpreted as the probability that we observe a sample given a parameter vector θ. The idea of maximum likelihood estimation is to find θ that maximizes the probability that we have observed the sample at hand. The likelihood function is given by

$$L_n(\theta) = f(z;\theta) = \prod_{i=1}^{n} f(z_i;\theta).$$

The difference between the likelihood function $L_n(\theta)$ and the joint density function $f(z;\theta)$ is *purely conceptual*. The joint density $f(z;\theta)$ treats θ as known and assigns probabilities

as z varies. The likelihood function $L_n(\theta)$ on the other hand, treats z as known and assigns the probability of observing z as θ varies. It is natural then to seek the θ that maximizes $L_n(\theta)$, that is, the *maximum likelihood estimator* (MLE) is defined by

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\text{argmax}}\ L_n(\theta),$$

where Θ is a compact subset of $\mathbb{R}^p$. To simplify computations it is customary to work with the normalized (average) *log-likelihood function* given by

$$\ell_n(\theta) \equiv \frac{1}{n} \log L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(z_i; \theta).$$

Clearly, the maximizer of $L_n(\theta)$, also maximizes $\ell_n(\theta)$.

We also define the $(p \times 1)$ *score vector* by

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \log f(z_i; \theta) \ = \frac{\partial \ell_n(\theta)}{\partial \theta} = \Big[\frac{\partial \ell_n(\theta)}{\partial \theta_1}, ..., \frac{\partial \ell_n(\theta)}{\partial \theta_p} \Big]'$$

and the $(p \times p)$ *Hessian matrix* by

$$H_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta\theta'} \log f(z_i; \theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_1} & \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_1 \partial \theta_p} \\ \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_2^2} & \cdots & \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_p \partial \theta_1} & \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_p \partial \theta_2} & \cdots & \dfrac{\partial^2 \ell_n(\theta)}{\partial \theta_p^2} \end{bmatrix}.$$

Note that by the invariance of the order of partial differentiation, the Hessian matrix is symmetric about its diagonal.

The population analogues of these quantities are the *population log-likelihood*

$$\ell(\theta) = \text{plim}\ \ell_n(\theta) = E_z[\ell_n(\theta)] = E_z[\log f(z; \theta)]$$

the *population score*

$$s(\theta) = \nabla_\theta \ell(\theta) = \text{plim}\ s_n(\theta) = E_z[\nabla_\theta \ell_n(\theta)] = E_z[\nabla_\theta \log f(z; \theta)]$$

and the *population Hessian*,

$$H(\theta) = \nabla_{\theta\theta'} \ell(\theta) = \text{plim}\ H_n(\theta) = E_z[\nabla_{\theta\theta'} \ell_n(\theta)] = E_z[\nabla_\theta \nabla_{\theta'} \log f(z; \theta)]$$

We will show that $\ell(\theta)$ is maximized at θ_0, that $s(\theta_0) = 0$, and $H(\theta_0)$ is negative semi-definite.

If $\ell_n(\theta)$ is smooth in θ, the MLE $\hat{\theta}_n$ must satisfy the *first order condition*

$$s_n(\hat{\theta}_n) = 0,$$

as well as, the *second order condition* that

$$H_n(\hat{\theta}_n) \quad \text{is negative semi-definite.}$$

If, in addition, $\ell_n(\theta)$ is globally concave, we can compute the MLE by a *Newton-Raphson iteration*:

1. Specify an initial value $\theta_{[0]} \in \Theta \subset \mathbb{R}^p$.
2. Given $\theta_{[i]}$, compute $\theta_{[i+1]} = \theta_{[i]} - [H_n(\theta_{[i]})]^{-1} s_n(\theta_{[i]})$.
3. Iterate until convergence, i.e. until $||\theta_{[i+1]} - \theta_{[i]}|| < \varepsilon$.

The $(p \times 1)$ vector $[H_n(\theta_{[i]})]^{-1} s_n(\theta_{[i]})$ is called a *Newton step*; its sign specifies the *direction* in which we should move, while its absolute magnitude gives the *size of the step* to be taken towards this direction. If $\ell_n(\theta)$ is exactly quadratic in θ the Newton-Raphson iteration will converge in one step, which is the same as saying that the MLE has a closed-form solution.

Example 3. (Normal Likelihood). Let $z = \{z_i, i = 1, ..., n\}$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. Then $\theta = (\mu, \sigma^2)'$ and

$$\begin{aligned}
\log L_n(\mu, \sigma^2) &= \sum_{i=1}^{n} \log \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(z_i - \mu)^2}{2\sigma^2} \right\} \right\} \\
&= \sum_{i=1}^{n} \left\{ -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{(z_i - \mu)^2}{2\sigma^2} \right\} \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - \mu)^2 \\
&\overset{\mu,\sigma^2}{\propto} -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (z_i - \mu)^2.
\end{aligned}$$

The notation $\overset{\mu,\sigma^2}{\propto}$ means that the last line is "proportional-in-parameters" to the previous line, so the maximizer of the two functions is the same. In particular, adding/subtracting constants or multiplying/dividing by constants (where by the word "constants" we mean quantities that do not contain θ) leaves the maximizer unaffected. The score vector is given by

$$s_n(\mu, \sigma^2) = \begin{bmatrix} \dfrac{\partial \ell_n(\mu, \sigma^2)}{\partial \mu} \\ \dfrac{\partial \ell_n(\mu, \sigma^2)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (z_i - \mu) \\ -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4} \displaystyle\sum_{i=1}^{n} (z_i - \mu)^2 \end{bmatrix},$$

and the Hessian matrix is given by

$$\begin{aligned} H_n(\mu,\sigma^2;y) &= \begin{bmatrix} \dfrac{\partial^2\ell_n(\mu,\sigma^2)}{\partial\mu^2} & \dfrac{\partial^2\ell_n(\mu,\sigma^2)}{\partial\mu\partial\sigma^2} \\ \dfrac{\partial^2\ell_n(\mu,\sigma^2)}{\partial\sigma^2\partial\mu} & \dfrac{\partial^2\ell_n(\mu,\sigma^2)}{\partial(\sigma^2)^2} \end{bmatrix} \\ &= \begin{bmatrix} -\dfrac{n}{\sigma^2} & -\dfrac{1}{\sigma^4}\displaystyle\sum_{i=1}^n(z_i-\mu) \\ -\dfrac{1}{\sigma^4}\displaystyle\sum_{i=1}^n(z_i-\mu) & \dfrac{n}{2\sigma^4}-\dfrac{1}{\sigma^6}\displaystyle\sum_{i=1}^n(z_i-\mu)^2 \end{bmatrix}, \end{aligned}$$

and

$$EH_n(\mu,\sigma^2;y) = \begin{bmatrix} -\dfrac{n}{\sigma^2} & 0 \\ 0 & -\dfrac{n}{2\sigma^4} \end{bmatrix}.$$

The MLE is obtained by solving the 2×2 system of equations defined by the score

$$\begin{aligned} \frac{1}{\hat\sigma_n^2}\sum_{i=1}^n(z_i-\hat\mu_n) &= 0 \\ -\frac{n}{\hat\sigma_n^2}+\frac{1}{2\hat\sigma_n^4}\sum_{i=1}^n(z_i-\hat\mu_n)^2 &= 0. \end{aligned}$$

Since the log-likelihood is exactly quadratic in (μ,σ^2), the score is linear in them, and we get the closed-form solutions

$$\hat\mu_n = \frac{1}{n}\sum_{i=1}^n z_i, \qquad \hat\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n(z_i-\hat\mu_n)^2.$$

Note that while the MLE of μ is simply the sample mean, the MLE of σ^2 is not the usual unbiased estimator $\hat s_n^2 = (n-1)^{-1}\sum_{i=1}^n(z_i-\hat\mu_n)^2$. ∎

Example 4. (Bernoulli Likelihood). Let $z=\{z_i, i=1,...,n\}$ be a random sample from a Bernoulli distribution with parameter p. This is a single parameter family with $\theta=p$, and the log-likelihood is given by

$$\begin{aligned} \ell_n(p) &= \sum_{i=1}^n \log[p^{z_i}(1-p)^{1-z_i}] \\ &= \sum_{i=1}^n z_i\log(p)+(1-z_i)\log(1-p). \end{aligned}$$

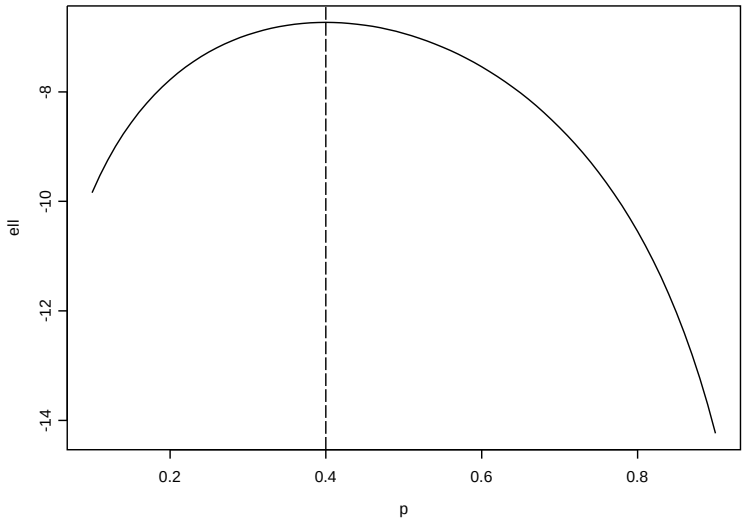

FIGURE 1

The likelihood for $p_0 = .5$ generated from the sample $(0, 0, 1, 0, 1, 1, 0, 0, 1, 0)$ is plotted in Figure 1. The score is given by

$$s_n(p) = \frac{d\ell_n(p)}{dp} = \sum_{i=1}^{n} \frac{z_i}{p} - \frac{1 - z_i}{1 - p} = \frac{1}{p(1-p)}\Big(\sum_{i=1}^{n} z_i - np\Big) = \frac{\overline{z}_n - p}{p(1-p)},$$

and the Hessian is given by

$$H_n(p) = \frac{d^2\ell_n(p)}{dp^2} = \frac{2\overline{z}_n p - \overline{z}_n - p^2}{p^2(1-p)^2}.$$

Setting the score equal to zero we obtain the MLE

$$s_n(\hat{p}_n) = 0 \Rightarrow \hat{p}_n = \overline{z}_n,$$

and evaluating the Hessian at $\hat{p}$ we get the curvature of the likelihood around $\hat{p}$.

$$H_n(\hat{p}_n) = -\frac{\hat{p}_n(1-\hat{p}_n)}{\hat{p}_n^2(1-\hat{p}_n)^2} = -\frac{1}{\hat{p}_n(1-\hat{p}_n)}.$$

For the example likelihood plotted above, $\hat{p}_n = .4$ and $H_n(\hat{p}_n) = -1/.16 = -6.25$. ■

Unfortunately, likelihoods are not always globally concave. If $\ell_n(\theta)$ is not globally concave the Newton-Raphson iteration will converge to a *local extremum* that may be a *local maximum* or even a *local minimum*! It is then prudent to repeat the procedure for several starting values and compare the values of the likelihood at the final estimates, in the hope that one of them will be the global maximum.

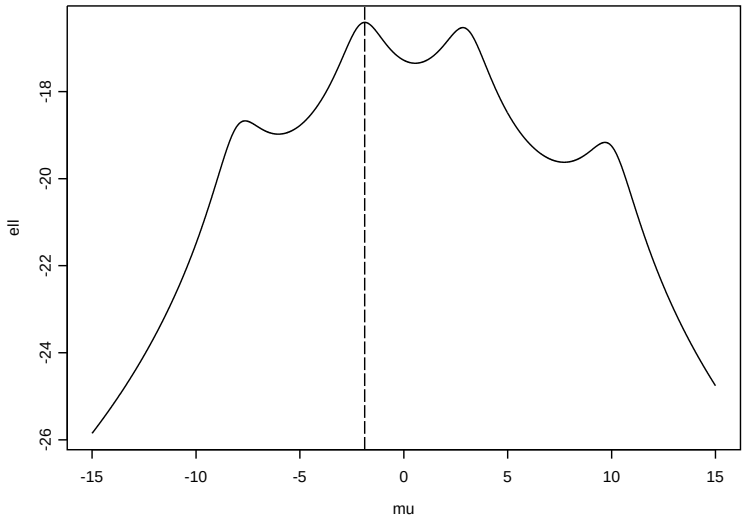

Figure 2

Example 5. (Cauchy Likelihood) Let $z = \{z_i, i = 1, ..., n\}$ be a random sample from a Cauchy distribution with parameter μ. The log-likelihood is given by

$$\ell_n(\mu) = \sum_{i=1}^{n} \log\left(\frac{1}{\pi[1 + (z_i - \mu)^2]}\right) \overset{\mu}{\propto} -\sum_{i=1}^{n} \log[1 + (z_i - \mu)^2].$$

The likelihood for $\mu_0 = 0$ generated from the sample $(-8, -2, 3, 10)$ is plotted in Figure 2. This likelihood has several local maxima and minima and our simple Newton-Raphson procedure will have trouble locating the global maximum at $\hat{\mu} = -1.88$. ■

1. Consistency and Asymptotic Normality of the MLE

Why is maximizing the sample likelihood a reasonable esimation strategy? Under minor regulatory conditions we will show that the MLE is consistent, asymptotically normal and asymptotically efficient (minimum variance). Our argument will be developed along the following lines:

1. First, we will show that the population likelihood $\ell(\theta) = \text{plim}\ell_n(\theta) = E[\ell_n(\theta)]$ is maximized at the true population parameter vector θ_0, i.e. that

$$\theta_0 = \underset{\theta \in \Theta}{\text{argmax}}\ \ell(\theta).$$

   We say that the likelihood *identifies* θ_0. This is a crucial result without which likelihoods would be useless.
2. We will next argue that, by the *Law of Large Numbers* (LLN), the sample likelihood $\ell_n(\theta)$ converges to the population likelihood $\ell(\theta)$ uniformly over $\theta \in \Theta$, that

is

$$\ell_n(\theta) \xrightarrow{p} \ell(\theta) \quad \text{for all } \theta \in \Theta, \quad \text{as } n \to \infty.$$

3. Our next step will be to use the *Continuous Mapping Theorem* (CMT) to assert that since the argmax functional is continuous, the maximizer of $\ell_n(\theta)$ converges in probability to the maximizer of $\ell(\theta)$, that is, our result in step 2 and the CMT imply that as $n \to \infty$

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\text{argmax}}\ \ell_n(\theta) \xrightarrow{p} \theta_0 = \underset{\theta \in \Theta}{\text{argmax}}\ \ell(\theta).$$

4. Having established the consistency of $\hat{\theta}_n$, we will then concentrate our attention to the asymptotic behavior of the sample likelihood $\ell_n(\theta)$ in small neighborhoods of θ_0. Using a Taylor series expansion of the score $s_n(\theta)$ around θ_0 we will arrive at an *asymptotic linearity expression* for the normalized random variable $\sqrt{n}(\hat{\theta}_n - \theta_0)$. The *Central Limit Theorem* (CLT) and the *Law of Large Numbers* (LLN) will then yield the asymptotic normality of the MLE.
5. Our last step will be to show that the MLE is "optimal" in the sense that it's variance attains the *Cramer-Rao Lower Bound* (CRLB). This will be done by showing that, for likelihoods (and likelihoods only), the *sandwich* variance term reduces to a *single* term.

The first theorem we present proves that, under minor regulatory conditions, the expected (population) likelihood $\ell(\theta)$ is maximized at the true parameter vector θ_0.

Theorem 1. (Identification of θ_0). *Suppose that $\{z_i, i = 1, ..., n\}$ is an i.i.d. sample from a distribution with p.d.f. $f(z|\theta_0)$ and*

(A.1) (Parameter Space). *$\theta_0 \in \Theta$ and Θ is a compact subset of $\mathbb{R}^p$.*
(A.2) (Density Identification). *For every $\theta \neq \theta_0$ in Θ, $f(z;\theta) \neq f(z;\theta_0)$; and*
(A.3) (Absolute Integrability). *For all $\theta \in \Theta$, $E[|\log f(z;\theta)|] < \infty$.*

Then $\ell(\theta) \equiv E[\log f(z;\theta)]$ has a unique maximum at θ_0.

Proof: For any $\theta \in \Theta$,

$$\begin{aligned}
\ell(\theta_0) - \ell(\theta) &= E\Big[-\log \frac{f(z;\theta)}{f(z;\theta_0)}\Big] \\
&> -\log E\Big[\frac{f(z;\theta)}{f(z;\theta_0)}\Big] \qquad \text{by Jensen's inequality} \\
&= -\log \int \frac{f(z;\theta)}{f(z;\theta_0)} f(z;\theta_0) dz \\
&= -\log 1 = 0.
\end{aligned}$$

■

The second theorem proves consistency of the MLE.

Theorem 2. (Consistency of the MLE). *In addition to (A.1)-(A.3), assume that*

(A.4) (Smoothness) $\ell(\theta)$ *is continuous at each* $\theta \in \Theta$ *with probability one.*

Then $\ell_n(\theta) \xrightarrow{p} \ell(\theta)$ *uniformly over* $\theta \in \Theta$*, and* $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Proof: We shall take as given that $\ell_n(\theta) \xrightarrow{p} \ell(\theta)$ uniformly in $\theta \in \Theta$, and prove only that this uniform convergence, Theorem 1, and our assumptions imply the consistency of $\hat{\theta}_n$. This is the (in)famous "3 epsilons" proof. Since $\hat{\theta}_n$ maximizes $\ell_n(\theta)$, for any $\epsilon > 0$ we have that with probability approaching 1 (w.p.a.1),

$$(A) \quad \ell_n(\hat{\theta}_n) > \ell_n(\theta_0) - \epsilon/3.$$

In addition, by the uniform convergence in probability of $\ell_n(\theta)$ to $\ell(\theta)$, we have that w.p.a.1

$$(B) \quad \ell(\hat{\theta}_n) > \ell_n(\hat{\theta}_n) - \epsilon/3,$$

and that w.p.a.1

$$(C) \quad \ell_n(\theta_0) > \ell(\theta_0) - \epsilon/3.$$

Therefore w.p.a.1,

$$\ell(\hat{\theta}_n) > \ell_n(\hat{\theta}_n) - \epsilon/3 > \ell_n(\theta_0) - 2\epsilon/3 > \ell(\theta_0) - \epsilon,$$

where the first inequality follows from (B), the second from (A), and the third from (C). Thus, for any $\epsilon > 0$, $\ell(\hat{\theta}_n) > \ell(\theta_0) - \epsilon$ w.p.a.1. Let $\mathcal{N}$ be any open subset of Θ containing θ_0. Since $\mathcal{N}$ is open, its complement $\mathcal{N}^c$ is closed, and since Θ is compact (and therefore, by the Heine-Borel Theorem, closed and bounded), $\Theta \cap \mathcal{N}^c$ is also compact. The compactness of $\Theta \cap \mathcal{N}^c$, Theorem 1, and (A.4), yield that $\sup_{\theta \in \Theta \cap \mathcal{N}^c} \ell(\theta) \equiv \ell(\theta^*) < \ell(\theta_0)$ for some $\theta^* \in \Theta \cap \mathcal{N}^c$, since a continuous real-valued function defined on a compact set is bounded and acheives maximum and minimum values[1]. Thus, choosing $\epsilon = \ell(\theta_0) - \ell(\theta^*)$, it follows that w.p.a.1 $\ell(\hat{\theta}_n) > \ell(\theta^*)$, and hence $\hat{\theta}_n \in \mathcal{N}$ w.p.a.1. ■

Consistency is a *global* property. To establish it, we needed to look at the asymptotic behavior of the sample criterion function (here the sample likelihood) over all $\theta \in \Theta$. Once we establish it, however, we need not worry any more about the behavior of the criterion function away from θ_0. Since $\hat{\theta}_n$ converges to θ_0 as n becomes large, the only thing about the likelihood that is relevant asymptotically is it's behavior around θ_0. Asymptotic normality is a *local* property, and we will prove it by studying the local behavior of the likelihood via a Taylor series expansion of the score around θ_0.

Expanding $s_n(\hat{\theta}_n) = \nabla_\theta \ell_n(\hat{\theta}_n)$ around θ_0 we obtain

$$0 = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(z_i;\theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta'} \log f(z_i;\overline{\theta}_n)\right](\hat{\theta}_n - \theta_0)$$

[1]See, for example, section 2.2 of Pugh C. C. (2002), *Real Mathematical Analysis*, Springer.

where $\overline{\theta}_n$ is a $1 \times p$ vector that is element-wise between $\hat{\theta}_n$ and θ_0. Solving for $(\hat{\theta}_n - \theta_0)$ we obtain

$$(\hat{\theta}_n - \theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta'} \log f(z_i; \overline{\theta}_n)\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(z_i; \theta_0),$$

and multiplying both sides with $\sqrt{n}$ we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta'} \log f(z_i; \overline{\theta}_n)\right]^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta \log f(z_i; \theta_0).$$

By the Weak Law of Large Numbers and the Continuous Mapping Theorem, the first term in the rhs converges in probability to $H(\theta_0)^{-1} = E_z[\nabla_{\theta\theta'}\ell_n(\theta_0)]^{-1}$. Also, by the Central Limit Theorem, the second term converges to a normal random variable with mean $E[\nabla_\theta \ell_n(\theta_0)] = 0$ (see Theorem 1) and variance $I(\theta_0) = E_z[\nabla_\theta \ell_n(\theta_0)\nabla_\theta \ell_n(\theta_0)']$. Therefore,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} -H(\theta_0)^{-1} N\Big(0, I(\theta_0)\Big) \sim N\Big(0, H(\theta_0)^{-1} I(\theta_0) H(\theta_0)^{-1}\Big).$$

The outer product of the population score

$$I(\theta) = E_z[\nabla_\theta \ell_n(\theta_0)\nabla_\theta \ell_n(\theta)'] = \nabla_\theta \ell(\theta)\nabla_\theta \ell(\theta)' = s(\theta)s(\theta)'$$

plays an important role and is call the *information matrix*. The last step is to show that the MLE attains the *Cramer-Rao Lower Bound*. But this is a direct result of the *information matrix equality*.

Theorem 3. (Information Matrix Equality)
Under the conditions stated in the next theorem,

$$I(\theta) = -H(\theta).$$

Proof: Since f is a density

$$\int f(z;\theta)dz = 1.$$

Differentiating with respect to θ we obtain

$$\begin{aligned} 0 &= \int \nabla_\theta f(z;\theta)dz \\ &= \int \frac{1}{f(z;\theta)}[\nabla_\theta f(z;\theta)]f(z;\theta)dz \\ &= \int [\nabla_\theta \log f(z;\theta)]f(z;\theta)dz. \end{aligned}$$

Differentiating again with respect to θ' we obtain

$$\begin{aligned} 0 &= \int [\nabla_{\theta\theta'} \log f(z;\theta)] f(z;\theta) + [\nabla_\theta \log f(z;\theta)][\nabla_\theta f(z;\theta)]' dz \\ &= \int [\nabla_{\theta\theta'} \log f(z;\theta)] f(z;\theta) dz + \int [\nabla_\theta \log f(z;\theta)] \frac{1}{f(z;\theta)} [\nabla_\theta f(z;\theta)]' f(z;\theta) dz \\ &= \int [\nabla_{\theta\theta'} \log f(z;\theta)] f(z;\theta) dz + \int [\nabla_\theta \log f(z;\theta)][\nabla_\theta \log f(z;\theta)]' f(z;\theta) dz \\ &\equiv E[\nabla_{\theta\theta'} \ell(\theta)] + E[\nabla_\theta \ell(\theta) \nabla_\theta \ell(\theta)']. \end{aligned}$$

Therefore, $E[\nabla_{\theta\theta'} \ell(\theta)] = -E[\nabla_\theta \ell(\theta) \nabla_\theta \ell(\theta)']$. ■

This result means that we can reduce the *sandwich* in the variance of the MLE to a *single* term. *Variances of the sandwich form are never optimal, while single-term variances are always optimal.* The following theorem states the final result.

Theorem 4. (Asymptotic Normality and Efficiency of the MLE).
In addition to A.1-A5, assume that

(A.6) (Parameter Space) θ_0 *belongs to the interior of* Θ.
(A.7) (Differentiability) $\ell_n(\theta)$ *is three times continuously differentiable in* $\sqrt{n}-$*neighborhoods of* θ_0.
(A.8) (Regularity) *For all* $\theta \in \Theta$, *the* $p \times p$ *matrix* $I(\theta) = E[\nabla_\theta \ell_n(\theta) \nabla_\theta \ell_n(\theta)']$ *is finite and positive definite.*

Then, as $n \to \infty$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p\Big(0, I(\theta_0)^{-1}\Big).$$

Proof: See Chapter 5, Theorems 5.3 and 5.4, of Knight Keith (1999), *Mathematical Statistics*, Crc Pr Inc. ■

This result can be used to construct asymptotically valid confidence intervals for θ_0. In particular, a $(1-\alpha)$ confidence interval for θ_0 is given by

$$\hat{\theta}_n \pm z_{\alpha/2} [I(\hat{\theta}_n)^{-1/2} / \sqrt{n}].$$

Example 6. (Bernoulli Likelihood, continued). Consider again the Bernoulli likelihood. Taking the expectation of the sample Hessian evaluated at p_0 we obtain

$$H(p_0) = E[H_n(p_0)] = E\Big[\frac{2\overline{z}p_0 - \overline{z} - p_0^2}{p_0^2(1-p_0)^2}\Big] = -\frac{p_0(1-p_0)}{p_0^2(1-p_0)^2} = -\frac{1}{p_0(1-p_0)}.$$

Checking the validity of the information matrix equality in this case

$$I(p_0) = E[s_n(p_0)^2] = E\left\{\Big[\frac{(\overline{z}-p_0)}{p(1-p_0)}\Big]^2\right\} = \frac{\mathrm{Var}(\overline{z})}{p_0^2(1-p_0)^2} = \frac{p_0(1-p_0)}{p_0^2(1-p_0)^2} = \frac{1}{p_0(1-p_0)},$$

we see that indeed $H(p_0) = -I(p_0)$. Theorem 2 yields that $\hat{p} = \overline{z} \xrightarrow{p} p_0$, while Theorem 4 gives

$$\sqrt{n}(\hat{p}_n - p_0) \xrightarrow{d} N\Big(0, p_0(1-p_0)\Big).$$

The asymptotic variance may be estimated by $\hat{p}(1-\hat{p})$, so an asymptotic $(1-\alpha)$ confidence interval for p_0 is given by $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. In terms of our sample, $n = 10$, $\hat{p} = .4$, $I(\hat{p}_n)^{-1} = .16$, and a 95% asymptotic confidence interval for p is given by $.4 \pm 1.96(\sqrt{.16/10})$. ■

Example 7. (Cauchy Likelihood, continued).

■

The information matrix equality is also useful for computational purposes. It suggests that we could replace the Hessian in the Newton-Raphson iteration by the outer product of the score. This idea yields the *BHHH algorithm*. Let

$$S_n(\theta) = \frac{1}{n}\begin{bmatrix} \dfrac{\partial \ell_n(\theta; z_1)}{\partial \theta_1} & \dfrac{\partial \ell_n(\theta; z_1)}{\partial \theta_2} & \cdots & \dfrac{\partial \ell_n(\theta; z_1)}{\partial \theta_p} \\ \dfrac{\partial \ell_n(\theta; z_2)}{\partial \theta_1} & \dfrac{\partial \ell_n(\theta; z_2)}{\partial \theta_2} & \cdots & \dfrac{\partial \ell_n(\theta; z_2)}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial \ell_n(\theta; z_n)}{\partial \theta_1} & \dfrac{\partial \ell_n(\theta; z_n)}{\partial \theta_2} & \cdots & \dfrac{\partial \ell_n(\theta; z_n)}{\partial \theta_p} \end{bmatrix}.$$

be the $n \times p$ matrix of partial derivatives evaluated at each observation $\{z_i, i = 1, ..., n\}$. Clearly then, the sample score $s_n(\theta)$ is the sum of each column of $S_n(\theta)$, i.e., if we let $1 = (1, ..., 1)'$ be an $n \times 1$ vector of ones, and $S_n^{(j)}(\theta)$ be jth column of $S_n(\theta)$, $j = 1, ..., p$, then

$$s_n(\theta) = \left(1'S_n^{(1)}(\theta), ..., 1'S_n^{(p)}(\theta)\right)'.$$

The sample outer product of the score is the $p \times p$ matrix given by

$$I_n(\theta) = S_n(\theta)'S_n(\theta).$$

The *BHHH algorithm* is:

1. Specify an initial value $\theta_{[0]} \in \Theta \subset \mathbb{R}^p$.
2. Given $\theta_{[i]}$, compute $\theta_{[i+1]} = \theta_{[i]} + I_n(\theta_{[i]})^{-1} s_n(\theta_{[i]})$.
3. Iterate until convergence, i.e. until $||\theta_{[i+1]} - \theta_{[i]}|| < \varepsilon$.

Note the plus sign in front of the outer product of the score in step 2 (recall that $H(\theta) = -J(\theta)$). The BH^3 algorithm saves us the trouble of deriving the Hessian, *but it is only applicable with likelihoods.*

Finally, we present a result, called the *delta method*, that allows us to derive the asymptotic distribution of any smooth function $g(\theta)$ of θ. For instance, in the ellipticity of the earth example, we were interested in the parameter $\eta = \beta_1/\beta_2$ ($\beta_2 \neq 0$). This parameter is a smooth but non-linear function of the β's, so our discussion regarding linear transformations of the parameters in a previous lecture does not apply.

Theorem 5. (Delta Method) *If* $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, \Sigma)$*, and* $g(\cdot)$ *is a smooth* $m \times 1$ *vector function of* θ*, then*

$$\sqrt{n}\Big(g(\hat{\theta}) - g(\theta_0)\Big) \xrightarrow{d} N_m\Big(0, G(\theta_0)\Sigma G(\theta_0)'\Big),$$

where $G(\theta) = \nabla_\theta g(\theta)$.

Proof: Expanding $g(\hat{\theta}_n)$ around θ_0 we obtain

$$g(\hat{\theta}_n) = g(\theta_0) + \nabla_\theta g(\overline{\theta}_n)(\hat{\theta}_n - \theta_0)$$

where $\overline{\theta}_n$ is between $\hat{\theta}_n$ and θ_0. Rearranging and rescaling we obtain

$$\sqrt{n}\Big(g(\hat{\theta}_n) - g(\theta_0)\Big) = \nabla_\theta g(\overline{\theta}_n) \quad \sqrt{n}(\hat{\theta}_n - \theta_0).$$

The result now follows from (a) the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$, (b) the fact that since $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\overline{\theta}_n$ is stuck between them, $\overline{\theta}_n \xrightarrow{p} \theta_0$ also, and (c) the continuous mapping theorem that implies that $\nabla_\theta g(\overline{\theta}_n) \xrightarrow{p} \nabla_\theta g(\theta_0) \equiv G(\theta_0)$. ∎

This result is not particular to MLE's, but applies to any consistent estimator with an asymptotically normal distribution. It says that even when the function g is nonlinear in θ, asymptotically only the linear part of it's expansion around θ_0 is important for inference! This is a direct result of the consistency of $\hat{\theta}_n$ which implies that, for large n, $\hat{\theta}_n$ is close to θ_0, and the fact that, at least locally, smooth functions behave like linear ones.

2. Maximum Likelihood Estimation Under Equality Restrictions

Now consider the problem of estimating θ under the restriction

$$g(\theta) = 0$$

where $g(\theta)$ is a $m \times 1$ vector function of the $p \times 1$ vector θ, $m \leq p$, such that the $m \times p$ matrix of first partial derivatives of g, $G(\theta) = \nabla_\theta g(\theta)$, is of full rank. For example, if $g(\theta)$ is linear in θ, we can write $g(\theta) = R\theta - r$, where R is a $m \times p$ matrix of rank m, and r is an m vector. In this case $G(\theta) = R$.

The *restricted maximum likelihood estimator* (RMLE) given by

$$\widetilde{\theta}_n = \underset{\theta,\lambda}{\operatorname{argmax}}\ \ell_n(\theta) - \lambda' g(\theta)$$

where λ is an $m \times 1$ vector of Lagrange multipliers.

Theorem 6. *Let assumptions A1-A8 hold and write*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p\left(0, I(\theta_0)^{-1}\right)$$

for the asymptotic distribution of the unrestricted MLE $\hat{\theta}_n$. Then, if $g(\theta_0) = 0$ and $G(\theta) = \nabla_\theta g(\theta)$ is of full rank, $\widetilde{\theta}_n \xrightarrow{p} \theta_0$ and

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \xrightarrow{d} N_p\left(0, \widetilde{I}(\theta_0)^{-1}\right)$$

where

$$\widetilde{I}(\theta_0)^{-1} = I(\theta_0)^{-1} - I(\theta_0)^{-1}G(\theta_0)'[G(\theta_0)I(\theta_0)^{-1}G(\theta_0)]^{-1}G(\theta_0)I(\theta_0)^{-1}.$$

Furthermore, $\widetilde{\lambda}_n \xrightarrow{p} 0$, and

$$\sqrt{n}\,\widetilde{\lambda}_n \xrightarrow{d} N_m\left(0, G(\theta_0)I(\theta_0)^{-1}G(\theta_0)'\right).$$

Proof: We will prove a stronger result, namely the *joint* asymptotic distribution $\widetilde{\theta}_n$ and $\widetilde{\lambda}_n$. The marginals given in the Theorem will then follow directly. Differentiating we obtain the FOC's

$$\begin{aligned} \nabla_\theta \ell_n(\widetilde{\theta}_n) - \widetilde{\lambda}_n' \nabla_\theta g(\widetilde{\theta}_n) &= 0 \\ g(\widetilde{\theta}_n) &= 0. \end{aligned}$$

Expanding $\nabla_\theta \ell_n(\widetilde{\theta}_n)$ and $g(\widetilde{\theta}_n)$ around θ_0 we obtain

$$\begin{aligned} \nabla_\theta \ell_n(\theta_0) + \left[\nabla_{\theta\theta'}\ell(\overline{\theta}_n)\right](\widetilde{\theta}_n - \theta_0) - \widetilde{\lambda}_n' \nabla_\theta g(\widetilde{\theta}_n) &= 0 \\ g(\theta_0) + \nabla_\theta g(\overline{\theta}_n)(\widetilde{\theta}_n - \theta_0) &= 0 \end{aligned}$$

Setting $g(\theta_0) = 0$, we rewrite this system as

$$\begin{aligned} s_n(\theta_0) + H_n(\overline{\theta}_n)(\widetilde{\theta}_n - \theta_0) - \widetilde{\lambda}_n' G(\widetilde{\theta}_n) &= 0 \\ G(\overline{\theta}_n)(\widetilde{\theta}_n - \theta_0) &= 0 \end{aligned}$$

In matrix notation the system is,

$$\begin{bmatrix} H_n(\overline{\theta}_n) & G(\widetilde{\theta}_n)' \\ G(\overline{\theta}_n) & 0 \end{bmatrix} \begin{bmatrix} (\widetilde{\theta}_n - \theta_0) \\ \widetilde{\lambda}_n \end{bmatrix} = \begin{bmatrix} -s_n(\theta_0) \\ 0 \end{bmatrix}.$$

It follows that

$$\sqrt{n}\begin{bmatrix} (\widetilde{\theta}_n - \theta_0) \\ \widetilde{\lambda}_n \end{bmatrix} \xrightarrow{p} \begin{bmatrix} H(\theta_0) & G(\theta_0)' \\ G(\theta_0) & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\sqrt{n}s_n(\theta_0) \\ 0 \end{bmatrix}.$$

The result now follows from observing that $\sqrt{n}s_n(\theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$, and Lemma 5.1 regarding the inversion of partitioned matrices. ∎

The result in Theorem 6 applies to any asymptotically normal estimator. If $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, V)$, then $\sqrt{n}(\widetilde{\theta}_n - \theta_0) \xrightarrow{d} N_p(0, \widetilde{V})$, where $\widetilde{V} = V - VG'[GVG]^{-1}GV$, and all quantities that depend on θ are evaluated at $\theta = \theta_0$. See the appendix of Manski and MacFadden (1981).

Example 8. (Restricted Least Squares) Consider least squares estimation under linear restrictions. Let $\widetilde{\beta}$ denote the Restricted Least Squares (RLS) estimator. Then

$$\begin{aligned}
\widetilde{\beta} - \beta &= (X'X)^{-1}X'u + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\beta - R(X'X)^{-1}X'u) \\
&= (X'X)^{-1}X'u - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'u \\
&= \Big[I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\Big](X'X)^{-1}X'u.
\end{aligned}$$

The variance of $\widetilde{\beta}$ is now given by

$$\begin{aligned}
V(\widetilde{\beta}) &= \sigma_u^2\Big[I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\Big](X'X)^{-1}\Big[I - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\Big] \\
&= \sigma_u^2(X'X)^{-1} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}.
\end{aligned}$$

This is of the form of $\widetilde{J}(\theta_0)^{-1}$ in Theorem 6, with $J(\theta_0)^{-1} = \sigma_u^2(X'X)^{-1}$ and $G(\theta_0) = R$. In fact, we can write

$$V(\widetilde{\beta}) = \widetilde{\sigma}_u^2(X'X)^{-1}$$

where

$$\widetilde{\sigma}_u^2 = \sigma_u^2 - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R,$$

is the error variance under the restriction. Clearly, $\widetilde{\sigma}_u^2 < \sigma_u^2$. ∎

3. Testing Using the Likelihood

Finally, consider the problem of testing the hypothesis

$$H_0 : g(\theta_0) = 0$$

against the alternative

$$H_1 : g(\theta_0) \neq 0$$

where $g(\cdot)$ is a $m \times 1$ vector function of the $p \times 1$ vector θ, $m \leq p$. For example, if $g(\theta)$ is linear in θ we can write $g(\theta) = R\theta - r$, where R is a $m \times p (m < p)$ matrix of known constants and r is a $m \times 1$ known vector.

The test statistics are defined as

$$\begin{aligned}
\xi_{LR} &= 2n[\ell_n(\hat{\theta}_n) - \ell_n(\widetilde{\theta}_n)] \\
\xi_W &= ng(\hat{\theta}_n)'[G(\hat{\theta}_n)'I_n(\hat{\theta}_n)^{-1}G(\hat{\theta}_n)]^{-1}g(\hat{\theta}_n) \\
\xi_{LM} &= ns_n(\widetilde{\theta}_n)'I_n(\widetilde{\theta}_n)^{-1}s_n(\widetilde{\theta}_n)
\end{aligned}$$

where, as before, $\hat{\theta}_n$ and $\widetilde{\theta}_n$ are the unrestricted and restricted estimates, respectively.

The Likelihood Ratio test statistic ξ_{LR} compares the values of the likelihood at the two estimates. Since imposing a restriction in estimation can only decrease the likelihood, it is clear that $\ell(\hat{\theta}_n) \geq \ell(\widetilde{\theta}_n)$. If imposing the restriction $g(\theta) = 0$ does not affect the likelihood very much, we say that the data are compatible with the null and we accept it. On the other hand, if $\ell(\widetilde{\theta}_n)$ is much smaller than $\ell(\hat{\theta}_n)$, then the restriction puts an undue strain on the data, and we should reject the null.

The Wald test statistic ξ_W, on the other hand, is based on the distance between $g(\hat{\theta}_n)$, the value of the restriction evaluated at the unrestricted estimate, from zero, which is the value of $g(\theta_0)$ under the null. If we find that $g(\hat{\theta}_n)$ differs from the zero vector by a lot, we could interpret this as evidence against the null, while if it happens to be close to zero, then the null seems reasonable and we should accept it.

Finally, the Lagrange Multiplier statistic ξ_{LM} is based on the slope of the likelihood (as measured by the score) at the restricted estimate $\widetilde{\theta}_n$. The slope of the likelihood at the unrestricted estimate $\hat{\theta}_n$ is by definition zero, so if we find that the slope at $\widetilde{\theta}_n$ is much bigger than zero, it would cast doubt on the null. If, on the other hand, we find that the slope of the likelihood at $\widetilde{\theta}_n$ is close to zero, we would be inclined to accept the null.

Example 9. Assume there is only one parameter in the model, so that $p = 1$, and consider the null hypothesis $H_0 : \theta = \theta_0$. Figure 3 plots $\ell_n(\theta)$ against θ. It is clear that ξ_{LR} is based on the distance AB, ξ_W on the distance CD and ξ_{LM} on the slope of the line EF. If the null hypothesis is correct, all three quantities should be close to zero for sufficiently large samples. For similar expositions see Buse (1982) and Engle (1984).

Another geometrical interpretation, which provides even more insight, is suggested by Pagan (1981). Consider Figure 4, where $s_n(\theta)$ is plotted against θ. The unrestricted MLE, $\hat{\theta}_n$, is obtained by setting $s_n(\theta) = 0$, i.e., at point B. It is easily seen that

$$\ell_n(\hat{\theta}) - \ell_n(\theta_0) = \int_{\theta_0}^{\hat{\theta}} s_n(\theta) d\theta = \text{area}(ABC).$$

Hence, $\xi_{LR} = 2[\ell_n(\hat{\theta}) - \ell(\theta_0)] = 2\text{area}(ABC)$. Furthermore, as $g(\theta) = \theta - \theta_0$ and $G(\theta) = 1$, we have $\xi_W = (\hat{\theta}_n - \theta_0) I(\hat{\theta}_n)$. An estimate of $I(\hat{\theta}_n)$ is given by $-\nabla_{\theta\theta'} \ell_n(\hat{\theta}_n)$, i.e., AD/AB. Hence, $\xi_W = (AB)^2 AD/AB = 2\text{area}(ABD)$. Finally, the LM statistic depends on $s_n(\theta_0)$, which is AC. Using $-\nabla_{\theta\theta'} \ell_n(\theta_0)$ as an estimate of $I_n(\theta_0)$, we obtain $\xi_{LM} = (AC)^2 AE/AC = 2\text{area}(AEC)$.

Three comments are worth making:

(i) Since the three tests are based on three different quantities, they may yield conflicting inferences if the same critical value is used.

(ii) The LM test depends only on $s_n(\theta_0)$ and the slope of $s_n(\theta)$ at θ_0. We can draw many lines through C with the same slope at C, the dotted line being an example.

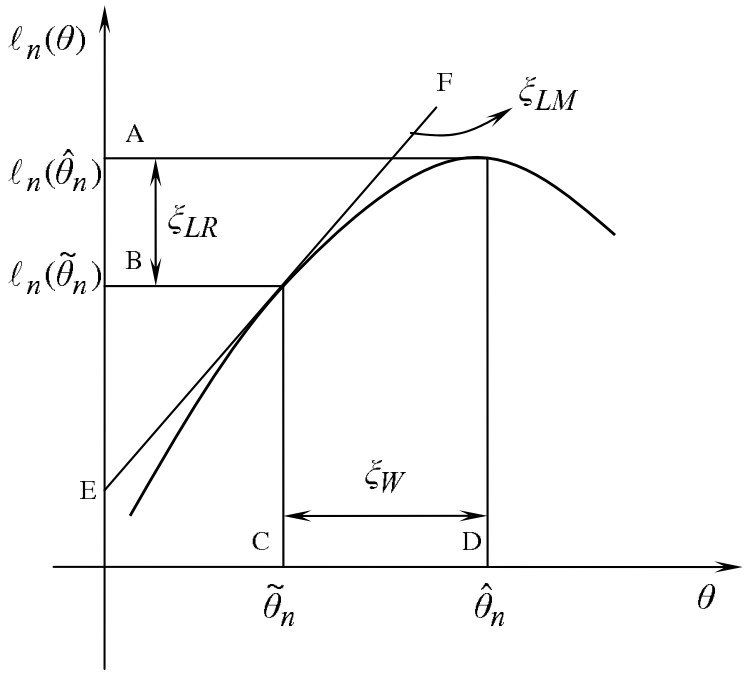

Figure 3. A graphical interpretation of the three classical test statistics due to Buse (1982).

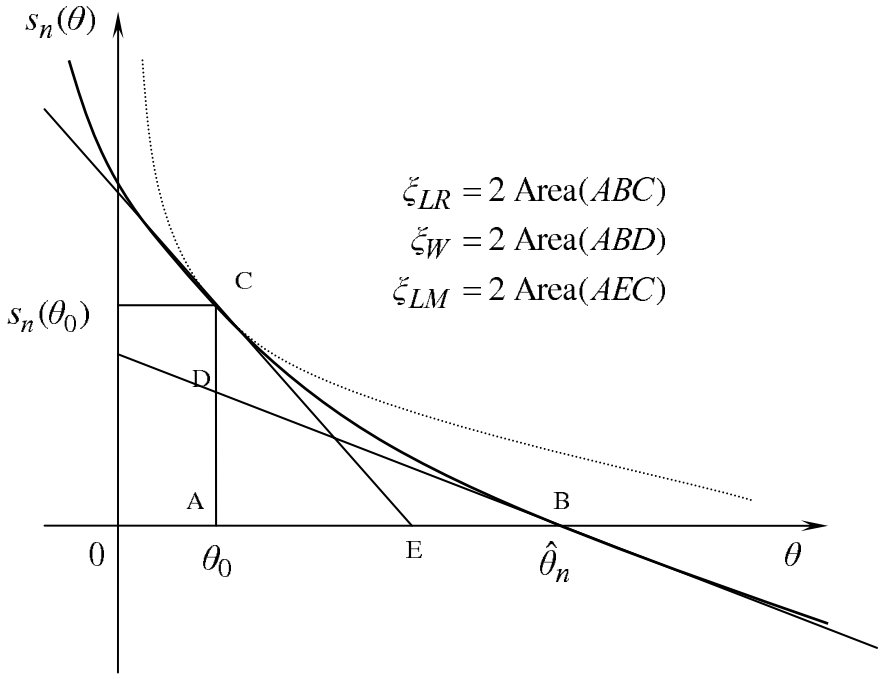

Figure 4. Another graphical interpretation of the three classical test statistics based on the score $s_n(\theta)$ due to Pagan (1981).

This implies there may be other likelihood functions (alternative hypotheses) with the same slope at θ_0 giving rise to the same LM statistic. We call this the

invariance property of the LM test. Alternative hypotheses giving rise to the same LM test are said to be equivalent.

(iii) The variance estimate of the LM statistic can be estimated in a number of ways which are asymptotically equivalent. This leads to different versions of test statistics with different properties in small samples.

■

But what do we mean by "close" and "almost zero"? In order to implement a test we need exact critical values that if our statistics exceed should lead us to reject the null. The following theorem tells us that $\chi^2_{m,\alpha}$ is the correct critical value for all three likelihood tests.

Theorem 7. *Under the null, the three statistics are asymptotically equivalent and distributed as* χ^2_m.

Thus our decision rule is to reject the null if the ξ_{LR}, ξ_W, or ξ_{LM} statistics exceed $\chi^2_{m,\alpha}$. Asymptotically, the decision of all three test statistics will be the same, so at least in large samples it shouldn't matter which one we employ. In finite samples, however, they may disagree.

So, how do we decide which one to use? Since all three tests are asymptotically equivalent, and since our justification for using them is only valid in large samples, there is no theoretical reason to prefer one over the others. Therefore, people usually employ the statistic that is the most convenient to estimate. To implement the Wald test we only need $\hat{\theta}_n$, i.e. it is enough to have estimates of the unrestricted model (estimation under the alternative). On the other hand, the LM test requires the restricted estimates $\widetilde{\theta}_n$, so to implement it we will need to first compute the likelihood under the null. The LR test, finally, needs both the restricted and unrestricted estimates, and it is thus the hardest to implement. In applications, unrestricted estimation is usually the simplest, so the Wald test is a favorite among practitioners.

Example 10. (Multinomial Distribution and Pearson's Goodness-of-Fit Test) Consider a multinomial distribution with p classes and let the probability that an observation belongs to the jth class be θ_j, such that $\sum_{j=1}^p \theta_j = 1$. Given a random sample of n observations, we denote the frequency of the jth class by n_j, so that $\sum_{j=1}^p n_j = n$. The likelihood is given by

$$L_n(y;\theta) = \frac{n!}{n_1!n_2!\cdots n_p!}\theta_1^{n_1}\theta_2^{n_2}\cdots\theta_p^{n_p} = n!\prod_{j=1}^p \frac{\theta_j^{n_j}}{n_j!},$$

where, $y = (n_1, n_2, ..., n_p)'$ and $\theta = (\theta_1, \theta_2, ..., \theta_p)'$. The LR, W, and LM statistics for testing the hypothesis

$$H_0 : \theta_j = \theta_{j0},\ j = 1, ..., p,\ \ [\theta_{j0} > 0, \sum_{j=1}^{p} \theta_{j0} = 1]$$

are given by

$$\xi_{LR} = 2 \sum_{j=1}^{p} n_j \log \left(\frac{n_j}{n\theta_{j0}} \right),$$

$$\xi_W = \sum_{j=1}^{p} \frac{(n_j - n\theta_{j0})^2}{n_j},$$

and

$$\xi_{LM} = \sum_{j=1}^{p} \frac{(n_j - n\theta_{j0})^2}{n\theta_{j0}}.$$

Note that the LM statistic is simply the classical *Pearson goodness-of-fit test statistic.*

As an application, consider the digits of π and assume we wish to test the hypothesis that they all occur with the same frequency. Let $\theta_j, j = 1, ..., 10$ be the frequency of each of the digits 0 through 9, and write the null as

$$H_0 : \theta_j = \frac{1}{10}, \quad j = 1, ..., 10.$$

The first 10,000 digits of π yield the following table,

j	:	0	1	2	3	4	5	6	7	8	9
n_j	:	968	1026	1021	975	1012	1046	1021	969	948	1014

The test statistics along with their p-values from the χ^2_{10} distribution are:

	ξ_{LR}	ξ_W	ξ_{LM}
statistic	9.357	9.424	9.328
p-value	.501	.508	.499

The null is accepted by all three tests. ■

Example 11. (Least Squares Estimation) See Johnston and DiNardo (pp. 142-151). Consider the linear regression model $y = X\beta + u$ with normal errors, and assume we wish to test the linear in β restriction

$$H_0 : g(\beta) \equiv R\beta - r = 0,$$

where R is a $m \times p\,(m < p)$ matrix of known constants and r is a $m \times 1$ known vector. The likelihood ratio test statistic is given by

$$\begin{aligned}\xi_{LR} &= 2[\ell(\hat{\theta}) - \ell(\widetilde{\theta})] \\ &= n[\log(\widetilde{u}'\widetilde{u}) - \log(\hat{u}'\hat{u})] \\ &= n\log\Big(1 + \frac{\widetilde{u}'\widetilde{u} - \hat{u}'\hat{u}}{\hat{u}'\hat{u}}\Big).\end{aligned}$$

The Wald test statistic is given by

$$\begin{aligned}\xi_W &= (R\hat{\beta} - r)'[RJ(\beta_0)^{-1}R']^{-1}(R\hat{\beta} - r) \\ &= \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{\hat{\sigma}^2} \\ &= \frac{n(\widetilde{u}'\widetilde{u} - \hat{u}'\hat{u})}{\hat{u}'\hat{u}},\end{aligned}$$

and the Lagrange multiplier test statistic is

$$\begin{aligned}\xi_{LM} &= \frac{n(\widetilde{u}'\widetilde{u} - \hat{u}'\hat{u})}{\widetilde{u}'\widetilde{u}} \\ &= nR_a^2.\end{aligned}$$

where R_a^2 is the R-squared of the *auxiliary regression* of $\widetilde{u}$ on X. All of these statistics are asymptotically distributed as χ_m^2. See Johnston and DiNardo (pp. 142-151) for details. ■

4. Profile Likelihood

It often happens that the parameter vector θ can be partitioned into two subsets $\theta = (\theta_1, \theta_2)$ Given the joint likelihood $L(\theta, \eta)$ the *profile likelihood* of θ is

$$L(\theta) = \max_{\eta} L(\theta, \eta),$$

where the maximization over η is performed at *each* fixed value of θ. Note that this is not the same as the *estimated likelihood* $L(\theta, \hat{\eta})$, the likelihood evaluated at the mle of η. In general, at each fixed θ, the maximizer of the above problem is a function of θ, say $\hat{\eta}(\theta)$ and the profile likelihood is $L(\theta) = L(\theta, \hat{\eta}(\theta))$. If η is well estimated (i.e., $\hat{\eta}$ has a small s.e.), the estimated likelihood $L(\theta, \hat{\eta})$ and the profile likelihood $L(\theta, \hat{\eta}(\theta))$ will be close. Otherwise, there will be significant differences. In any case, inference should be based only on the profile likelihood, since it takes into account the uncertainty about η, as opposed to the estimated likelihood that treats η as 'fixed' to its mle value without uncertainty.

Therefore, the set

$$\left\{\theta, \frac{L(\theta)}{L(\hat{\theta})} > e^{-\frac{1}{2}\chi^2_{p,(1-\alpha)}}\right\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

Example 12. Suppose $x_1, ..., x_n$ is and iid sample from $N(\mu, \sigma^2)$ with both parameters μ and σ^2 unknown. The likelihood function of (μ, σ^2) is given by

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right\}.$$

At each fixed μ the mle of σ^2 is given by

$$\hat{\sigma}^2(\mu) = \frac{1}{n}\sum_i (x_i - \mu)^2,$$

so the profile likelihood of μ is given by

$$L(\mu) = \text{constant} \times [\hat{\sigma}^2(\mu)]^{-n/2}.$$

This is not the same as the estimated likelihood

$$L(\mu, \hat{\sigma}^2) = \text{constant} \times \exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_i (x_i - \mu)^2\right\},$$

the slice of $L(\mu, \sigma^2)$ at $\sigma^2 = \hat{\sigma}^2$. The two will be close if σ^2 is estimated with a small s.e.. In any case, the profile likelihood is always wider than the estimated one, as it accounts for the uncertaintly over σ^2 too.

Fow example, suppose we observe

■

4.1. **Low-Dose Aspirin Consumption and the Risk of a Heart Attack.** In a landmark study of the preventive benefits of low-dose daily aspirin consumption for healthy individuals (Steering Committee of the Physicians Health Study Research Group 1989), a total of 22,071 healthy physicians were randomized to either aspirin or placebo groups, and were followed for an average of 5 years. The number of heart attacks and strokes during follow-up are shown in Table 1.

The main medical question is statistical: is aspirin beneficial? Obviously, there were fewer heart attacks in the aspirin group that the placebo group, 139 versun 239, but we face question: is the evidence in favor of low-dose daily aspirin usage strong enough? The

TABLE 1. The number of heart attacks and strokes during follow-up in the Physician's Health Study.

Group	Heart Attacks	Strokes	Total
Aspirin	139	119	11,037
Placebo	239	98	11,034
Total	378	217	22,071

side effects, as measured by the number of strokes, were greater in the aspirin group, although 119 versus 98 are not as convincing a s the benefit.

Suppose we express the benefit of aspirine as a relative risk

$$\frac{139/11,037}{239/11,034} = 0.581.$$

A relative risk of 1 indicates that aspirin is not beneficial, while a value much less than 1 indicates a benefit. Is 0.581 'far enough' from 1? Answering such a question requires a stocastic model that describes the data we observe.

Assume that the number numer of heart attacks in the aspirin group, x_a, follows binomial(n_a, θ_a), while the number of heart attacks in the placebo group, x_p, follows binomial(n_p, θ_p), and that x_a and x_p are independent. The likelihood in terms of the parameters (θ_a, θ_p) is given by

$$\begin{aligned} L(\theta_a, \theta_p) &= e^{-n_a\theta_a}\frac{(n_a\theta_a)^{x_a}}{x_a!}e^{-n_p\theta_p}\frac{(n_p\theta_p)^{x_p}}{x_p!} \\ &= \text{constant} \times e^{-(n_a\theta_a + n_p\theta_p)}\theta_a^{x_a}\theta_p^{x_p}. \end{aligned}$$

Changing variables to $\theta = \theta_a/\theta_p$ and θ_p the likelihood becomes

$$L(\theta, \theta_p) = \text{constant} \times e^{-\theta_p(n_a\theta + n_p)}\theta^{x_a}\theta_p^{x_a+x_p}.$$

Differenting w.r.t. θ_p we get ,

$$\begin{aligned} \frac{\partial L(\theta, \theta_p)}{\partial \theta_p} &= -(n_a\theta + n_p)e^{-\theta_p(n_a\theta+n_p)}\theta^{x_a}\theta_p^{x_a+x_p} + e^{-\theta_p(n_a\theta+n_p)}\theta^{x_a}(x_a + x_p)\theta_p^{x_a+x_p-1} \\ &= e^{-\theta_p(n_a\theta+n_p)}\theta^{x_a}[(x_a + x_p)\theta_p^{x_a+x_p-1} - (n_a\theta + n_p)\theta_p^{x_a+x_p}]. \end{aligned}$$

Setting this equal to 0 we obtain the mle of θ_p as a function of θ,

$$\begin{aligned} e^{-\hat{\theta}_p(n_a\theta+n_p)}\theta^{x_a}[(x_a + x_p)\hat{\theta}_p^{x_a+x_p-1} - (n_a\theta + n_p)\hat{\theta}_p^{x_a+x_p}] &= 0 \\ \Rightarrow (x_a + x_p)\hat{\theta}_p^{x_a+x_p-1} &= (n_a\theta + n_p)\hat{\theta}_p^{x_a+x_p} \\ \Rightarrow \hat{\theta}_p(\theta) &= \frac{x_a + x_p}{n_a\theta + n_p}. \end{aligned}$$

Substituting this back to the likelihood, we obtain the profile likelihood of θ,

$$\begin{aligned} L(\theta) &= \text{constant} \times e^{-(x_a+x_p)}\theta^{x_a}\left(\frac{x_a+x_p}{n_a\theta+n_p}\right)^{x_a+x_p} \\ &= \text{constant} \times \frac{\theta^{x_a}}{(n_a\theta+n_p)^{x_a+x_p}}. \end{aligned}$$

Differenting w.r.t. θ we get

$$\begin{aligned} \frac{dL(\theta)}{d\theta} &= \frac{d}{d\theta}\left[\theta^{x_a}(n_a\theta+n_p)^{-(x_a+x_p)}\right] \\ &= x_a\theta^{x_a-1}(n_a\theta+n_p)^{-(x_a+x_p)} - \theta^{x_a}(x_a+x_p)(n_a\theta+n_p)^{-(x_a+x_p)-1}n_a. \end{aligned}$$

Setting the derivative equal to 0, we obtain the mle of the relative risk θ,

$$\begin{aligned} \frac{x_a}{\hat{\theta}} &= \frac{n_a(x_a+x_p)}{n_a\hat{\theta}+n_p} \\ \Rightarrow \frac{\hat{\theta}}{x_a} &= \frac{\hat{\theta}+n_p/n_a}{x_a+x_p} \\ \Rightarrow \hat{\theta}\left[\frac{1}{x_a}-\frac{1}{x_a+x_p}\right] &= \frac{n_p/n_a}{x_a+x_p} \\ \Rightarrow \hat{\theta} &= \frac{n_p/n_a}{x_p/x_a} \\ \Rightarrow \hat{\theta} &= \frac{x_a/n_a}{x_p/n_p}, \end{aligned}$$

as expected. We have already computed this to be $\hat{\theta} = 0.581$. Evaluating the profile likelihood at the mle, we obtain

$$L_{max} = L(\hat{\theta}) = \left(\frac{x_a+x_p}{x_a}\right)^{x_a}\left(\frac{x_a+x_p}{x_p/n_p}\right)^{x_p}.$$

Setting the constant in the profile likelihood equal to the inverse of this quantity we obtain the normalized profile likelihood that ranges from 0 to 1,

$$\begin{aligned} L(\theta) &= \frac{1}{L_{max}} \times \frac{\theta^{x_a}}{(n_a\theta+n_p)^{x_a+x_p}} \\ &= \left(\frac{n_a\theta}{x_a}\right)^{x_a}\left(\frac{n_p}{x_p}\right)^{x_p}\left(\frac{x_a+x_p}{n_a\theta+n_p}\right)^{x_a+x_p}, \end{aligned}$$

which we can now use to do inference about θ.

Figure 5(a) graphs the profile likelihood for heart attacks. Using a 15% cut-off value we find that a 95% approximate CI for θ is given by $(0.471, 0.714)$. We see that the relative risk is significantly less than 1 at the 5% level, and conclude that low-dose daily aspirin consumption significantly reduces the risk of a heart attack.

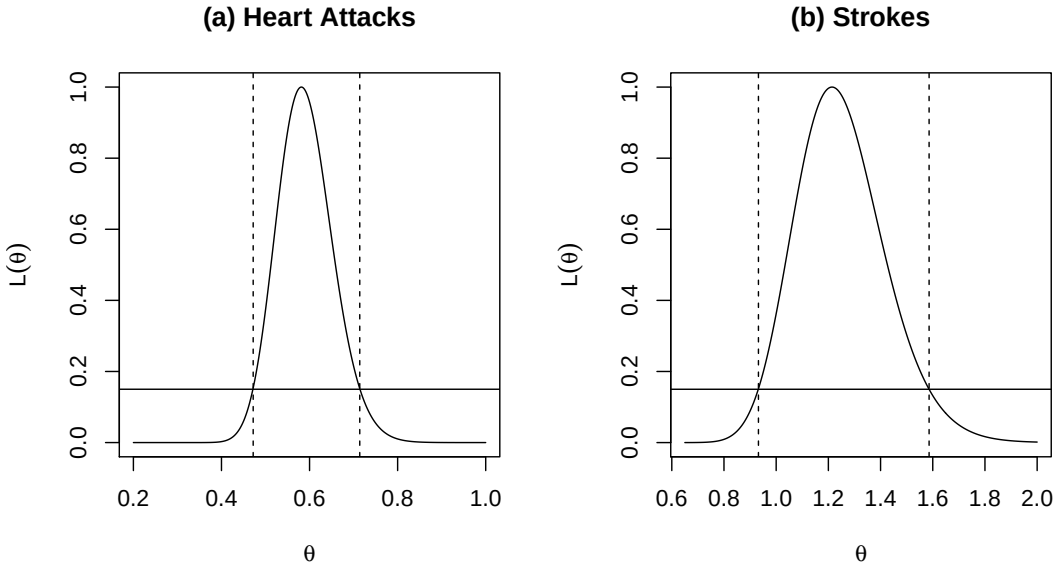

FIGURE 5

The exact same analysis yieds that the relative risk of a stroke is

$$\frac{119/11,037}{98/11,034} = 1.214.$$

Although this is greater than 1, the approximate 95% CI obtained from the profile likelihood for strokes, presented in Figure 5(b), is $(0.932, 1.586)$, so the relative risk of strokes is not significantly different from 1 at the 5% level. We conclude that daily low-dose apirin consumption significantly reduces the relative risk of a heart attack, without affecting significantly the relative risk of a stroke.

The R code used to produce Figure 5 is given below.

```
par(mfrow=c(1,2))

# Heart Attacks
Like <- function(theta){
      L <- 0
L <- (11037*theta/(11037*theta+11034))^139 *
              (1-(11037*theta/(11037*theta+11034)))^239
      return(L)
     }

thetahat <- (139/11037)/(239/11034)
thetahat
Like(thetahat)
thetaval <- seq(0.2,1,.001)
```

```
Lval <- Like(thetaval)/Like(thetahat)
plot(thetaval, Lval,type="l",xlab=expression(theta),ylab=expression(L(theta)),
     main="(a) Heart Attacks")
abline(h=0.15)
abline(v=0.472,lty=2)
abline(v=0.714,lty=2)
cbind(thetaval,round(Lval,3))

# Strokes
Like <- function(theta){
      L <- 0
L <- (11037*theta/(11037*theta+11034))^119 *
             (1-(11037*theta/(11037*theta+11034)))^98
      return(L)
     }
thetahat <- (119/11037)/(98/11034)
thetahat
Like(thetahat)
thetaval <- seq(0.65,2,.001)
Lval <- Like(thetaval)/Like(thetahat)
plot(thetaval, Lval,type="l",xlab=expression(theta),ylab=expression(L(theta)),
     main="(b) Strokes")
abline(h=0.15)
abline(v=0.932,lty=2)
abline(v=1.586,lty=2)
cbind(thetaval,round(Lval,3))
```

5. Numerical Optimization of the Likelihood

To compute the mle by Newton-Raphson iterations we need to be able to compute the sample score $s_n(\theta)$ and the sample Hessian $H_n(\theta)$ at any given $\theta \in \Theta$. Often, analytic expression of these quantities can easily be derived by differentiating the likelihood. If, however, analytic expressions are too hard to derive, or we are just too lazy to do the work, we can always resort to numerical derivatives.

Numerical derivatives exploit the definition of a derivative. The derivative of a function $f(x)$ at a point x_0 is given by

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h/2) - f(x_0 - h/2)}{h},$$

so for h small,

$$f'(x_0) \approx \frac{f(x_0 + h/2) - f(x_0 - h/2)}{h}.$$

Similarly, the second derivative is approximated by

$$\begin{aligned} f''(x_0) &\approx \frac{\dfrac{f(x_0+h)-f(x_0)}{h} - \dfrac{f(x_0)-f(x_0-h)}{h}}{h} \\ &= \frac{f(x_0+h)+f(x_0-h)-2f(x_0)}{h^2}. \end{aligned}$$

Similar arguments produce the following approximate formulae for partial derivatives,

$$\frac{\partial f(x_0, y_0)}{\partial x} \approx \frac{f(x_0 + h/2, y_0) - f(x_0 - h/2, y_0)}{h},$$

$$\frac{\partial f(x_0, y_0)}{\partial y} \approx \frac{f(x_0, y_0 + h/2) - f(x_0, y_0 - h/2)}{h}$$

and

$$\begin{aligned} \frac{\partial f(x_0, y_0)}{\partial x \partial y} \approx & \frac{[f(x_0 + h/2, y_0 + h/2) - f(x_0 + h/2, y_0 - h/2)]}{h^2} \\ & - \frac{[f(x_0 - h/2, y_0 + h/2) - f(x_0 - h/2, y_0 - h/2)]}{h^2}. \end{aligned}$$

To achieve accurate approximations, function evaluations should be done in double precision (16 significant digits), and h should be chosen carefully. Letting δ be the machine precision, Press et. al. (1997) suggest choosing h according to the rule

$$h(x_0) \sim \delta^{1/3}|x_0|.$$

This formula adjusts h for each x_0. This, however, does not seem necessary, so we will only use a constant h given by $h \sim \delta^{1/3}$. For example, when evaluations of f are done in double precision we have $\delta = 10^{-16}$, and $h = \delta^{1/3} = 10^{-5} = .00001$ approximately.

Intuition suggests that the second derivative should be approximable with only half the accuracy of the first one, so a larger h should be used. In theory, the same h can be used for both first and second derivatives: the numerator of $f''(x_0)$ can be written as $(f(x_0 + h) - f(x_0)) + (f(x_0 - h) - f(x_0))$, and since each of these quantities can be approximated with the same accuracy as $f'(x_0)$, it follows that the same h can be used in both approximations. In practise, however, second derivatives are indeed harder to estimate and it is advisable that a larger h be used (e.g. $h = 10^{-4}$). Of course, nothing prevents us from picking different h's for different derivatives: although the formulas above are written in terms of a single h, we are free to vary it as we please (e.g. choose $h = 10^{-5}$ for first derivatives and $h = 10^{-4}$ for second ones). The following simple example will demonstrate the issues at hand.

Example 13. Suppose we wish to numerically approximate the derivative of $f(x) = \exp(x)$ at $x_0 = 2$, which is equal to $f'(2) = \exp(2) \approx 7.389056098930650$ (up to 16 digits). Using double precision in our calculations (16 significant digits), we obtain the following numerical estimates of $f'(x_0)$ for various choices of h :

h	Numerical$f'(2)$	Absolute Error
10^{-1}	7.392135257174780	0.003079158244129
10^{-2}	7.389086886702770	0.000030787772117
10^{-3}	7.389056406808870	0.000000307878219
10^{-4}	7.389056102002910	0.000000003072261
10^{-5}	7.389056098805470	0.000000000125183
10^{-6}	7.389056100315370	0.000000001384721
10^{-7}	7.389056104756260	0.000000005825612
10^{-8}	7.389056122519830	0.000000023589180
10^{-9}	7.389056477791200	0.000000378860548
10^{-10}	7.389058254148040	0.000002155217388
10^{-11}	7.389022727011250	0.000033371919400
10^{-12}	7.389644451905040	0.000588352974390
10^{-13}	7.398526236102040	0.009470137171392
10^{-14}	7.371880883511040	0.017175215419612
10^{-15}	6.217248937900880	1.171807161029770
10^{-16}	0.000000000000000	7.389056098930650
10^{-17}	0.000000000000000	7.389056098930650

The absolute error decreases up to $h = 10^{-5}$ and then increases again. Happily, however, h's between 10^{-4} and 10^{-8} produce reasonably accurate results (accurate up to the 7th decimal place), so any choice of h in this range would work acceptably here. As we would expect, the numerical derivative vanishes for $h \leq 10^{-16}$, since these h's exceed the machine precision.

Turning now to the numerical approximation of the second derivative of $f(x)$ at $x_0 = 2$, we again have $f''(2) = \exp(2) \approx 7.389056098930650$, and we obtain the following

approximation results for various choices of h :

h	Numerical $f''(2)$	Absolute Error
10^{-1}	7.395215698561940	0.006159599631289
10^{-2}	7.389117674598820	0.000061575668171
10^{-3}	7.389056715823020	0.000000616892367
10^{-4}	7.389056122519830	0.000000023589182
10^{-5}	7.389040490579640	0.000015608351005
10^{-6}	7.389644451905040	0.000588352974389
10^{-7}	7.105427357601000	0.283628741329652
10^{-8}	0.000000000000000	7.389056098930650
10^{-9}	0.000000000000000	7.389056098930650

We see that the absolute error is now minimized at $h = 10^{-4}$, and although $h = 10^{-5}$ also produces reasonable results, the performance of the approximation deteriorates rapidly thereafter. Indeed, the approximation vanishes at $h \geq 10^{-8}$, which is in keeping with our intuition that the overall approximation has only half the accuracy of the first derivative approximation which vanishes at $h \geq 10^{-16}$. We conclude that, although theory is not incorrect in allowing the same h for both first and second derivatives, prudence requires a larger h in the approximation of the second derivative. ■

To apply numerical derivatives to the likelihood function, write the sample score as the p vector $s_n(\theta) = [s_n^i(\theta)]_{i=1,...,p}$, and approximate its elements by

$$s_n^i(\theta) \approx \frac{\ell_n(\theta_1, ..., \theta_i + h/2, ..., \theta_p) - \ell_n(\theta_1, ..., \theta_i - h/2, ..., \theta_p)}{h}.$$

Similarly, let $H_n(\theta) = [h_n^{i,j}(\theta)]_{i,j=1,...,p}$, be the $p \times p$ sample Hessian matrix, and approximate its elements by

$$\begin{aligned} h_n^{i,j}(\theta) \approx \frac{1}{h^2}\{&[\ell_n(\theta_1, ..., \theta_i + h/2, ..., \theta_j + h/2, ..., \theta_p) - \ell_n(\theta_1, ..., \theta_i + h/2, ..., \theta_j - h/2, ..., \theta_p)] \\ &-[\ell_n(\theta_1, ..., \theta_i - h/2, ..., \theta_j + h/2, ..., \theta_p) - \ell_n(\theta_1, ..., \theta_i - h/2, ..., \theta_j - h/2, ..., \theta_p)]\}. \end{aligned}$$

Assuming that the log-likelihood is globally concave, the mle $\hat{\theta}_n$ can now be computed by Newton-Raphson iterations, and the variance-covariance of $\hat{\theta}_n$ may be computed by $\text{Var}(\hat{\theta}_n) = [-H_n(\hat{\theta}_n)]^{-1}$.

6. Application: Testing the Uniformity of the Lottery

Johnson and Klotz (1993) considered the problem of testing the uniformity of the draws of Lotto America using a sample of $n = 200$ lotteries contacted form February 8, 1989 to January 5, 1991. The data for these lotteries are given in Table 1 of their paper. In each lottery $m = 6$ numbers are drawn, without replacement, from an urn containing $p = 54$

numbers. Letting $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ be the p vector of probabilities for each of the p numbers, we wish to test the hypothesis that $\theta_j = 1/p,\ j = 1, ..., p$.

The fact that the m numbers of each lottery are drawn without replacement makes the binomial distribution inappropriate in this application. Instead, the probability of a draw $X = (x_{i1}, x_{i2}, ..., x_{im})$ is given by

$$\begin{aligned} P[X = (x_{i1}, x_{i2}, ..., x_{im}); \theta] &= \theta_{x_{i1}} \times \frac{\theta_{x_{i2}}}{1 - \theta_{x_{i1}}} \times \frac{\theta_{x_{i3}}}{1 - \theta_{x_{i1}} - \theta_{x_{i2}}} \times \\ &\quad \cdots \times \frac{\theta_{x_{im}}}{1 - \theta_{x_{i1}} - \cdots - \theta_{x_{i(m-1)}}}. \end{aligned}$$

The likelihood of n lotteries is, therefore, given by

$$L_n(\theta; x) = \prod_{i=1}^{n} P[X = (x_{i1}, x_{i2}, ..., x_{im}); \theta],$$

and the log-likelihood is

$$\ell_n(\theta; x) = \sum_{i=1}^{n} \log P[X = (x_{i1}, x_{i2}, ..., x_{im}); \theta].$$

The mle is defined as

$$\hat{\theta}_n = \underset{\theta \in (0,1)^p}{\operatorname{argmax}}\ \ell_n(\theta; x), \qquad \text{s.t.} \sum_{j=1}^{p} \theta_j = 1.$$

We impose the summing-up constraint by maximizing with respect to the first $p - 1$ elements of θ and then setting $\theta_p = 1 - \sum_{j=1}^{p-1} \theta_j$.

Note that we use the un-normalized log-likelihood (we have not divided by n), and likewise, we will use the un-normalized score and Hessian. The reason is that according to Theorem 4 the variance-covariance of $\hat{\theta}_n$ can be estimated by $[H_n(\hat{\theta}_n)]^{-1}/n = [nH_n(\hat{\theta}_n)]^{-1}$, so if we do not normalize we can use the second derivative directly. In what follows, therefore, s_n and H_n will be the first and second derivatives of the un-normalized likelihood, i.e. they will be n times the quantities we worked with above. With this new definitions, the test statistics become

$$\begin{aligned} \xi_{LR} &= 2[\ell_n(\hat{\theta}_n) - \ell_n(\widetilde{\theta}_n)] \\ \xi_W &= g(\hat{\theta}_n)'[G(\hat{\theta}_n)' I_n(\hat{\theta}_n)^{-1} G(\hat{\theta}_n)]^{-1} g(\hat{\theta}_n) \\ \xi_{LM} &= s_n(\widetilde{\theta}_n)' I_n(\widetilde{\theta}_n)^{-1} s_n(\widetilde{\theta}_n) \end{aligned}$$

where, as before, $\hat{\theta}_n$ and $\widetilde{\theta}_n$ are the unrestricted and restricted estimates, respectively.

Although this log-likelihood is globally concave, analytical expressions for the score vector and the Hessian matrix are very complicated. We, therefore, approximate numerically s_n and H_n and then apply Newton-Raphson iterations to compute the mle. The `S` code,

initialized at $\theta_{[j0]} = 1/p = 1/54$, converged in 6 steps and approximately 42 minutes (7 minutes/iteration) on a 2.50 GHz `Dell Latitude` laptop running on `Windows XP`:

iteration	time started	log-likelihood	
0	18 : 32	−4729.238	(at $\theta_j = 1/54$)
1	18 : 38	−4702.306	
2	18 : 45	−4696.547	
3	18 : 52	−4695.497	
4	18 : 59	−4695.382	
5	19 : 05	−4695.380	
6	19 : 12	−4695.380	(at the mle)

Execution time can be reduced considerably if `Fortran` or `C` were used. `S`, like most higher-level languages, is slow on iterations, and since θ is a large vector here, many iterations are needed to approximate the score and especially the Hessian. Also, `S for Windows` is configured to use only half of the available CPU, so as other applications can run simultaneously. Accordingly, execution is slowed down to half speed.

Standard errors of the first 53 θ's are computed as the square root of the diagonal of the 53×53 matrix $[-H_n(\hat{\theta}_n)]^{-1}$. The standard error of $\hat{\theta}_{54}$ was computed by

$$\hat{\sigma}(\hat{\theta}_{54}) = \sqrt{e'[-H_n(\hat{\theta}_n)]^{-1}e}\ ,$$

where e is a 53×1 vector of ones.The estimates along with their standard errors are given in Table 1.

In order to test the uniformity hypothesis

$$H_0 : \theta_{j0} = 1/p,\ \ j = 1, ..., p,$$

Johnson and Klotz (1993) use the LR test. From our results above we get

$$\xi_{LR} = 2[\ell(\hat{\theta}_n) - \ell(\theta_0)] = 2(4,729.238 - 4,695.380) = 67.715.$$

For this null, $\xi_{LR} \overset{a}{\sim} \chi^2_{53}$, so we find that the p-value of the test is equal to .08403. Thus, we accept the null at the 5% level, but we reject it at the 10%. Johnson and Klotz (1993) report only the LR test. We can, however, easily implement all three. The test statistics and their p-values from the χ^2_{53} distribution are:

	ξ_{LR}	ξ_W	ξ_{LM}
statistic	67.715	76.531	69.484
p-value	.08403	.01889	.06386

We see that the LM and Wald tests yield even less support for the null. Notably, the Wald test rejects the null even at the 5% level.

Johnson and Klotz (1993) conclude that "the moderate evidence for non-uniformity is a result of the mechanical mixing process, in which balls enter the urn in sequence (always

Table 2

j	$\hat{\theta}_{jn}$	$\hat{\sigma}(\hat{\theta}_{jn})$	j	$\hat{\theta}_{jn}$	$\hat{\sigma}(\hat{\theta}_{jn})$	j	$\hat{\theta}_{jn}$	$\hat{\sigma}(\hat{\theta}_{jn})$
1	.02615	.00464	19	.02368	.00443	37	.01840	.00389
2	.02018	.00408	20	.01124	.00299	38	.01992	.00403
3	.03106	.00510	21	.00975	.00280	39	.01976	.00400
4	.02701	.00471	22	.01136	.00302	40	.02069	.00410
5	.01480	.00346	23	.01576	.00359	41	.01400	.00337
6	.01410	.00340	24	.01917	.00396	42	.02175	.00422
7	.02371	.00443	25	.02267	.00432	43	.02264	.00431
8	.01974	.00399	26	.01923	.00397	44	.01658	.00368
9	.02099	.00416	27	.01801	.00381	45	.02489	.00457
10	.02095	.00415	28	.01835	.00388	46	.01999	.00404
11	.01922	.00397	29	.01308	.00325	47	.01147	.00305
12	.01308	.00325	30	.02348	.00439	48	.01722	.00373
13	.02280	.00434	31	.01742	.00377	49	.01566	.00357
14	.01400	.00337	32	.01912	.00395	50	.01733	.00375
15	.01409	.00339	33	.01733	.00375	51	.01392	.00335
16	.02227	.00432	34	.01645	.00365	52	.01760	.00381
17	.02552	.00460	35	.01983	.00401	53	.01823	.00385
18	.01302	.00323	36	.01572	.00358	54	.01560	.00355

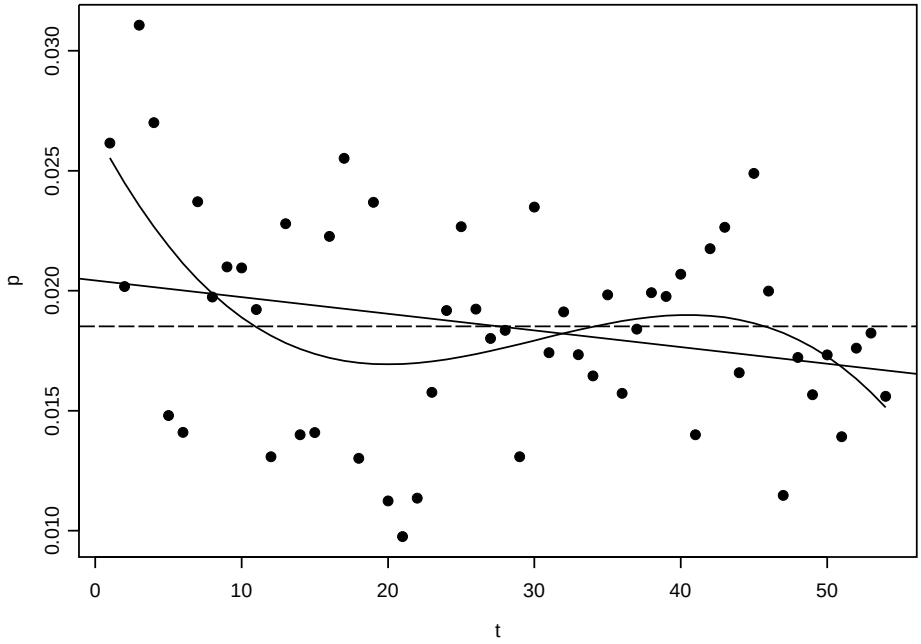

Figure 6

the same), are mixed, and then are drawn out at the bottom. Roughly, smaller-numbered

balls appear to have slightly higher odds of selection because they fall to the bottom first." Figure 5 plots θ_j against j. The dotted line is the expectation $\theta_j = 1/54$, while the two solid lines are the linear regressions of θ_j (a) on j and (b) on j, j^2 and j^3 (i.e. a spline). There is substantial evidence that small numbers are picked with probability greater than $1/54$, middle numbers are picked with probability equal to $1/54$, and large numbers are picked with probability less than $1/54$. We conclude that Lottery America at the time (early 1990's) needed to improve the mixing process inside the urn.

The `R` code used here is given below.

```
# Ref: Johnson R., and Klotz J., (1993) "Estimating Hot Numbers and
# Testing Uniformity for the Lottery", JASA, vol. 88, pp. 662-668.

Like <- function(p){
 L <- 0
 for (i in 1:n){
     L <- L + log(
        p[X[i,1]] *
        p[X[i,2]]/(1-p[X[i,1]]) *
        p[X[i,3]]/(1-p[X[i,1]]-p[X[i,2]])  *
        p[X[i,4]]/(1-p[X[i,1]]-p[X[i,2]]-p[X[i,3]]) *
        p[X[i,5]]/(1-p[X[i,1]]-p[X[i,2]]-p[X[i,3]]-p[X[i,4]]) *
        p[X[i,6]]/(1-p[X[i,1]]-p[X[i,2]]-p[X[i,3]]-p[X[i,4]]-p[X[i,5]]) )
    }
    return(L)
}

X <- t(matrix(c(
  06, 01, 38, 05, 11, 30,
  40, 37, 51, 04, 19, 30,
  40, 49, 09, 17, 41, 33,
  ......................
  26, 20, 12, 32, 39, 35
  ),6,200))

n <- length(X[,1])
p <- rep(1/54,54)
print(Like(p))

diff <- 100
while(diff > 0.0001){
    h <- 0.00001
    S <- rep(0,53)
    for(j in 1:53){
        pv1 <- p
        pv2 <- p
        pv1[j] <- pv1[j] + h/2
        pv1[54] <- 1-sum(pv1[1:53])
        pv2[j] <- pv2[j] - h/2
```

```
40          pv2[54] <- 1-sum(pv2[1:53])
41          S[j] <- (Like(pv1)-Like(pv2))/h
42      }
43      h <- 0.0001
44      H <- matrix(0,53,53)
45      for(i in 1:53){
46          for(j in 1:53){
47              pv1 <- p
48              pv2 <- p
49              pv3 <- p
50              pv4 <- p
51              pv1[i] <- pv1[i] + h/2
52              pv1[j] <- pv1[j] + h/2
53              pv1[54] <- 1-sum(pv1[1:53])
54              pv2[i] <- pv2[i] + h/2
55              pv2[j] <- pv2[j] - h/2
56              pv2[54] <- 1-sum(pv2[1:53])
57              pv3[i] <- pv3[i] - h/2
58              pv3[j] <- pv3[j] + h/2
59              pv3[54] <- 1-sum(pv3[1:53])
60              pv4[i] <- pv4[i] - h/2
61              pv4[j] <- pv4[j] - h/2
62              pv4[54] <- 1-sum(pv4[1:53])
63              H[i,j] <- ((Like(pv1)-Like(pv2))-(Like(pv3)-Like(pv4)))/(h^2)
64          }
65      }
66      pn <- p[1:53] - solve(H) %*% S     # Newton step
67      pn <- c(pn,1-sum(pn))
68      print(pn)
69      print(Like(pn))
70      diff <- abs(Like(p)-Like(pn))
71      p <- pn
72  }
73
74  sd <- c(sqrt(diag(solve(-H))),0)
75  print(cbind(p,sd))
```

References

Berndt, E., B. Hall, R. Hall, and J. Hausman, (1974), "Estimation and Inference in Nonlinear Structural Models", *Annals of Social Measurement*, vol. 3, 653-665.

Buse, A. (1982), "Likelihood Ratio, Wald, Lagrange Multiplier Test: An Expository Note", *American Statistician*, vol. 36, 153-157.

Johnson, R., and Klotz, J. (1993), "Estimating Hot Numbers and Testing Uniformity for the Lottery", *Journal of the American Statistical Association*, vol. 88, 662-668.

Manski, C.F., and MacFadden D., (1981) " Alternative Estimators and Sample Designs for Discrete Choice Analysis", in *Structural Analysis of Discrete Data with Econometric Applications*, eds. Manski, C.F., and MacFadden D., MIT Press.

Pagan, A.R. (1981), "Reflection on Australian Macro-Modeling", *Working Papers in Economics and Econometrics*, no. 48, Australian National University.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., (1997), "Numerical Recipes in Fortran 77", 2nd edition, Cambridge Press.

> By shape of likelihood, the news were told.
>
> – *William Shakespeare,*
> *King Henry IV, Act 1, Scene 1.*

"You haven't told me yet," said Lady Nuttal, "What it is your fiance does for a living."

"He's a statistician," replied Lamia, with an annoying sense of being on the defensive.

Lady Nuttal was obviously taken aback. It had not occurred to her that statisticians entered into normal social relationships. The species, she would have surmised, was perpetuated in some collateral manner, like mules.

"But Aunt Sara, it's a very interesting profession," said Lamia warmly.

"I don't doubt it," said her aunt, who obviously doubted it very much. "To express anything important in mere figures is so plainly impossible

that there must be endless scope for well-paid advice on how to do it. But don't you think that life with a statistician would be rather, shall we say, humdrum?"

Lamia was silent. She felt reluctant to discuss the surprising depth of emotional possibility which she had discovered below Edward's numerical veneer.

"It's not the figures themselves," she said finally, "It's what you do with them that matters."

K.A.C. Mandeville, *The Undoing of Lamia Gurdleneck*