Does the MLE Maximize the Likelihood?

David Olive*

Southern Illinois University

August 28, 2004

Abstract

One of the most useful properties of the maximum likelihood estimator (MLE), often called the invariance property, is that if $\hat{\theta}$ is the MLE of θ, then $h(\hat{\theta})$ is the MLE of $h(\theta)$. Many texts either define the MLE of $h(\theta)$ to be $h(\hat{\theta})$, say that the property is immediate from the definition of the MLE, or quote Zehna (1966). A little known paper, Berk (1967), gives an elegant proof of the invariance property that can be used in introductory statistical courses.

KEY WORDS: Point Estimation.

*David J. Olive is Associate Professor, Department of Mathematics, Mailcode 4408, Southern Illinois University, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. This research was supported by NSF grant DMS 0202922.

1 INTRODUCTION

The method of maximum likelihood is a popular technique for deriving estimators. Recall that if $X_1, ..., X_n$ is an iid sample from a population with pdf or pmf $f(x|\theta)$ then the likelihood function

$$L(\theta) = L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1.1}$$

The X_i can be random variables or vectors, and the parameter θ can take many forms. If the population is Poisson(λ), then $\theta = \lambda$ is a scalar. If the population is normal N(μ, σ^2), then θ is the vector $(\mu, \sigma^2)^T$, and if the population is multivariate normal $N_p(\mu, \Sigma)$, then θ is the pair (μ, Σ) where μ is a $p \times 1$ vector and Σ is a $p \times p$ matrix.

Following Casella and Berger (2002), let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{x} = (x_1, ..., x_n)$. For each sample point $\boldsymbol{x}$, let $\hat{\theta}(\boldsymbol{x})$ be the parameter value at which $L(\theta|\boldsymbol{x})$ attains its maximum as a function of θ with $\boldsymbol{x}$ held fixed. Then the maximum likelihood estimator (MLE) of the parameter θ based on the sample $\boldsymbol{X}$ is $\hat{\theta}(\boldsymbol{X})$.

The MLE may not exist, and finding the global maximizer of a likelihood can be very difficult. The fact that the range of the MLE coincides with the range of θ is useful, and most introductory texts suggest the following four techniques for finding the MLE.

- Potential candidates can be found by differentiating the likelihood.
- Potential candidates can be found by differentiating log $L(\theta|\boldsymbol{x})$, the log likelihood.
- The MLE can sometimes be found by direct maximization of the likelihood.
- If $\hat{\theta}$ is the MLE of θ, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

The last technique, often called the invariance property of the MLE, is usually stated without proof. Bickel and Doksum (1977, p. 99), Devore (1991, p. 250), and Lehmann (1983, p. 112) state that any function h can be used (implicitly assuming that $h(\hat{\theta})$ is a well defined estimator). Cox and Hinkley (1974, p. 287) and Serfling (1980, p. 149) state that the invariance property is immediate from the definition of the MLE. Several authors note that the invariance principle can be proved when h is one to one. See Anderson (1984, p. 64), DeGroot (1975, p. 291), Casella and Berger (2002, p. 320), Hogg and Craig (1995, p. 265), and Lindgren (1962, pp. 224-225). For general h, Anderson (1984, p. 64), Johnson and Wichern (1988, p. 141), and Casella and Berger (2002, p. 320) refer to Zehna (1966).

The next section will show that Berk(1967) answers some questions about the MLE which can not be answered using Zehna (1966).

2 TWO PROOFS OF THE INVARIANCE PRINCIPLE

The argument of Zehna (1966) also appears in Zehna (1970, p. 367-369) and Casella and Berger (2002, p. 320). Let $\theta \in \Theta$ and let $h : \Theta \to \Lambda$ be a function. Since the MLE $\hat{\theta} \in \Theta$, $h(\hat{\theta}) = \hat{\lambda} \in \Lambda$. If h is not one to one, then many values of θ may be mapped to λ. Let

$$\Theta_\lambda = \{\theta : h(\theta) = \lambda\}$$

and define the induced likelihood function $M(\lambda)$ by

$$M(\lambda) = \sup_{\theta \in \Theta_\lambda} L(\theta). \tag{2.1}$$

Then for any $\lambda \in \Lambda$,

$$M(\lambda) = \sup_{\theta \in \Theta_\lambda} L(\theta) \leq \sup_{\theta \in \Theta} L(\theta) = L(\hat{\theta}) = M(\hat{\lambda}). \tag{2.2}$$

Hence $h(\hat{\theta}) = \hat{\lambda}$ maximizes the induced likelihood $M(\lambda)$. Zehna (1966) says that since $h(\hat{\theta})$ maximizes the induced likelihood, we should call $h(\hat{\theta})$ the MLE of $h(\theta)$, but the definition of MLE says that we should be maximizing a genuine likelihood.

This argument raises a two important questions.

- If we call $h(\hat{\theta})$ the MLE of $h(\theta)$ and h is not one to one, does $h(\hat{\theta})$ maximize a likelihood or should $h(\hat{\theta})$ be called a maximum induced likelihood estimator?

- If $h(\hat{\theta})$ is an MLE, what is the likelihood function $K(h(\theta))$?

Some examples might clarify these questions.

- If the population come from a $N(\mu, \sigma^2)$ distribution, we say that the MLE of μ/σ is $\bar{X}_n/S_n$ where

  $$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  and

  $$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

  are the MLE's of μ and σ^2. Since the function $h(x, y) = x/\sqrt{y}$ is not one to one, what is the likelihood $K(h(\mu, \sigma^2)) = K(\mu/\sigma)$ that is being maximized?

- If X_i comes from a Bernoulli(p) population, why is $\bar{X}_n(1-\bar{X}_n)$ the MLE of $p(1-p)$?

Examining the invariance principle for one to one functions h is also useful. When h is one to one, let $\eta = h(\theta)$. Then the inverse function h^{-1} exists and $\theta = h^{-1}(\eta)$. Hence

$$f(\boldsymbol{x}|\theta) = f(\boldsymbol{x}|h^{-1}(\eta)) \tag{2.3}$$

is the joint pdf or pmf of $\boldsymbol{x}$. So the likelihood function of $h(\theta) = \eta$ is

$$K(\eta) = L(h^{-1}(\eta)). \tag{2.4}$$

Also note that

$$\sup_{\eta} K(\eta|\boldsymbol{x}) = \sup_{\eta} L(h^{-1}(\eta)|\boldsymbol{x}) = L(\hat{\theta}|\boldsymbol{x}). \tag{2.5}$$

Thus

$$\hat{\eta} = h(\hat{\theta}) \tag{2.6}$$

is the MLE of $\eta = h(\theta)$ when h is one to one.

If h is not one to one, then the new parameters $\eta = h(\theta)$ do not give enough information to define $f(\boldsymbol{x}|\eta)$. Hence we cannot define the likelihood. That is, a $N(\mu, \sigma^2)$ density cannot be defined by the parameter μ/σ alone. Before concluding that the MLE does not exist if h is not one to one, note that if $\boldsymbol{X} = (X_1, ..., X_n)$ is an iid Gaussian random sample, then $\boldsymbol{X}$ is *still* an iid Gaussian random sample even if the statistician did not rename the parameters wisely. Berk (1967) said that if h is not one to one, define

$$w(\theta) = (h(\theta), u(\theta)) = (\eta, \gamma) = \xi \tag{2.7}$$

such that $w(\theta)$ is one to one. Note that the choice

$$w(\theta) = (h(\theta), \theta)$$

works. In other words, we can always take u to be the identity function.

The choice of w is not unique, but the inverse function

$$w^{-1}(\xi) = \theta$$

is unique. Hence the likelihood is well defined, and $w(\hat{\theta})$ is the MLE of ξ. Thus calling $h(\hat{\theta})$ the MLE of $h(\theta)$ is analogous to calling $\bar{X}_n$ the MLE of μ when the data are from a $N(\mu, \sigma^2)$ population. It is often possible to choose the function u so that if θ is a $p \times 1$ vector, then so is ξ. For the $N(\mu, \sigma^2)$ example with $h(\mu, \sigma^2) = \mu/\sigma$ we can take $u(\theta) = \mu$ or $u(\theta) = \sigma^2$. For the Ber(p) example, $w(p) = (p(1-p), p)$ is a reasonable choice.

To summarize, the statistician who changes the names of the parameters of a sample does not change the distribution of the sample. Hence if one starts with a Gaussian likelihood but changes the names of the parameters, a Gaussian likelihood is still being maximized. Lastly, the invariance principle holds if $h(\hat{\theta})$ is a random variable, not only for one to one functions h, and Berk's proof should be widely used.

3 References

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed. John Wiley and Sons, Inc., NY.

Berk, R. (1967). "Review 1922 of 'Invariance of Maximum Likelihood Estimators' by Peter W. Zehna," *Mathematical Reviews,* 33, 344-343.

Bickel, P.J., and Doksum, K.A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics,* Holden-Day, San Francisco, CA.

Casella, G., and Berger, R.L. (2002), *Statistical Inference,* 2nd ed., Duxbury, Belmont, CA.

Cox, D.R., and Hinckley, D.V. (1974), *Theoretical Statistics,* Chapman and Hall, London.

DeGroot, M.H. (1975), *Probability and Statistics,* Addison-Wesley Publishing Company, Reading MA.

Devore, J.L. (1991), *Probability and Statistics for Engineering and the Sciences,* Wadsworth, Inc, Belmont, Ca.

Hogg, R.V., and Craig, A.T (1995), *Introduction to Mathematical Statistics,* 5th ed. Prentice Hall, Englewood Cliffs, NJ.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Lehmann, E.L. (1983), *Theory of Point Estimation,* John Wiley and Sons, Inc., NY.

Lindgren, B.W., (1962) *Statistical Theory,* Macmillan, New York.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* John Wiley and Sons, Inc., NY.

Zehna, P.W. (1966) "Invariance of Maximum Likelihood Estimators," *Annals of Mathematical Statistics,* 37, 744.

Zehna, P.W. (1970), *Probability Distributions and Statistics,* Allyn and Bacon, Inc. 1970.