General

Applications of Basu's Theorem

Dennis D. BOOS and Jacqueline M. HUGHES-OLIVER

This article reviews some basic ideas of classical inference and the roles they play in Basu's theorem. The usefulness of this theorem is demonstrated by applications to various statistical problems.

KEY WORDS: Ancillary statistic; Behrens–Fisher problem; Complete family; Monte Carlo swindle; Pivotal quantity; Sufficient statistic.

1. INTRODUCTION

Basu's theorem is one of the most elegant results of classical statistics. Succinctly put, the theorem says: if $\boldsymbol{T}$ is a complete sufficient statistic for a family of probability measures $\mathcal{P}$, and $\boldsymbol{V}$ is an ancillary statistic, then $\boldsymbol{T}$ and $\boldsymbol{V}$ are independent. (See, for example, Lehmann 1983, p. 46; Casella and Berger 1990, p. 262; or the original source, Basu 1955, theorem 2.) We often think of $\mathcal{P}$ as a parametric family $\mathcal{P} = \{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Omega\}$, where $\boldsymbol{\theta}$ has dimension q, but it is important to note that the theorem holds for more general families $\mathcal{P}$ such as classes of continuous distributions.

Basu's theorem is actually one of several theorems given in Basu (1955) and Basu (1958). Theorem 1 of Basu (1955) (updated in Basu 1958) states that ancillarity follows from sufficiency and independence. The main result above is taken from theorem 2 of Basu (1955) which actually uses the concept of bounded completeness (see Sec. 2.1) rather than completeness, but that is a minor broadening of the basic result. Basu (1982) said that the historical interest in these theorems was because they showed the connection between sufficiency, ancillarity, and independence, concepts that previously had seemed unrelated. In a different thread, Lehmann (1981) emphasized the connection between Basu's theorem and completeness.

The main emphasis of this article is to illustrate the wide variety of technical results that may be obtained from Basu's theorem. These applications are given in Section 3. In Section 2 we review some of the basic definitions and ideas from classical inference that are related to Basu's theorem.

Dennis D. Boos is Professor, and Jacqueline M. Hughes-Oliver is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695–8203 (E-mail: boos@stat.ncsu.edu and hughesol@stat.ncsu.edu). The authors thank Roger L. Berger for his careful reading of the manuscript and helpful suggestions.

2. STATISTICAL INFERENCE

The goal of statistical inference is to make statements about the underlying distribution that generated a given observed sample. Suppose the n-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ represents the sample. We assume that $\boldsymbol{X}$ has a distribution from some family $\mathcal{P}$, but we do not know the specific member, call it P_0, that generated $\boldsymbol{X}$. The basic goal is to determine P_0 based on $\boldsymbol{X}$. In the case of a parametric family, we need to determine the value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$, that corresponds to the distribution of $\boldsymbol{X}$.

To understand the role that Basu's theorem plays in inference, we first need to define some basic concepts.

2.1 Definitions

A *sufficient statistic* $\boldsymbol{T}$ is defined by the property that the conditional distribution of the data $\boldsymbol{X}$ *given* $\boldsymbol{T}$ is the same for each member of $\mathcal{P}$. For discrete $\boldsymbol{X}$, the density of $\boldsymbol{X}$ factors into two parts: the conditional density of $\boldsymbol{X}$ given $\boldsymbol{T}$, which is the same for every member of $\mathcal{P}$ and thus provides no information about P_0; and the density of $\boldsymbol{T}$, which contains all the information in the data that points to P_0 as the distribution of $\boldsymbol{X}$. It is thus reasonable that statistical inference be based solely on the sufficient statistic $\boldsymbol{T}$, which is usually of much smaller dimension that the data vector $\boldsymbol{X}$.

Reduction to a sufficient statistic, however, is neither unique nor guaranteed to yield maximum reduction. When it exists, a *minimal sufficient statistic* achieves the greatest reduction in data without losing information; that is, $\boldsymbol{T}$ is minimal sufficient if it is sufficient and can be computed from (is a function of) *every* sufficient statistic.

An *ancillary statistic* $\boldsymbol{V}$ is one whose distribution is the same for all members of $\mathcal{P}$. Therefore, $\boldsymbol{V}$ contains no information about the distribution of $\boldsymbol{X}$.

A statistic $\boldsymbol{T}$ is said to be *complete* with respect to the family $\mathcal{P}$ if there are no functions ϕ such that $E\phi(\boldsymbol{T}) = 0$ for all members of $\mathcal{P}$, except $\phi(t) \equiv 0$ a.e. $\mathcal{P}$. (*Bounded completeness* is a slightly more general concept. It restricts the class of functions ϕ under consideration to be bounded, allowing more families to have this property. Thus, completeness implies bounded completeness, but the distinction is not important for the applications in this article.) The relationship between completeness and sufficiency may now be discussed.

2.2 Basu's Theorem in Inference

The essential difference between a sufficient (or even minimal sufficient) statistic and a complete sufficient statistic is that the sufficient statistic $\boldsymbol{T}$ may contain extraneous information that is not relevant for determining P_0. That is, there may exist a function of $\boldsymbol{T}$ that is ancillary. On the other hand, Basu's theorem tells us that if $\boldsymbol{T}$ is complete in addition to being sufficient, then no ancillary statistics can be computed from $\boldsymbol{T}$ (except, of course, constant functions). Therefore, $\boldsymbol{T}$ has been successful in "squeezing out" any extraneous information.

Note that minimality of $\boldsymbol{T}$ does not guarantee this "squeezing out" but completeness does. Complete sufficient statistics are, however, always minimal sufficient (Lehmann 1983, p. 46). A nice discussion of the ideas in this subsection is found in Lehmann (1981).

Consider, for example, a random sample from the Laplace location family where

$$\mathcal{P} = \left\{ \prod_{i=1}^{n} f_\theta(x_i) : f_\theta(x) = f_0(x-\theta), \right.$$
$$\left. f_0(x) = \frac{1}{2}\exp(-|x|), -\infty < x < \infty, -\infty < \theta < \infty \right\}.$$

The order statistic $\boldsymbol{T} = (X_{(1)}, \ldots, X_{(n)})$ is sufficient for $\mathcal{P}$ but $\boldsymbol{V} = (X_{(n)} - X_{(1)}, \ldots, X_{(n)} - X_{(n-1)})$ is ancillary even though it may be obtained from $\boldsymbol{T}$.

Is $\boldsymbol{T}$ complete? Arguing from the contrapositive to Basu's theorem, we can say that since $\boldsymbol{T}$ and $\boldsymbol{V}$ are not independent (since $\boldsymbol{V}$ is a function of $\boldsymbol{T}$), then $\boldsymbol{T}$ is not a complete sufficient statistic for $\mathcal{P}$. A more typical method of showing lack of completeness comes from noting that $E[X_{(n)} - X_{(1)}] = c_1$, where c_1 just depends on $f_0(x) = (1/2)\exp(-|x|)$ and can be computed. Thus, $\boldsymbol{T}$ is not complete because $E\phi(\boldsymbol{T}) = 0$ for all members of $\mathcal{P}$, where $\phi(\boldsymbol{T}) = X_{(n)} - X_{(1)} - c_1$.

The Fisherian tradition suggests that when a sufficient statistic is not complete, then we should condition on ancillary statistics for the purposes of inference. There are at least two reasons for this. The first is so that probability calculations are relevant to the inference to be made (Cox 1988, p. 316). The second reason is to "recover" information lost in reducing to a sufficient statistic (Basu 1964).

This approach runs into problems because there are many situations where several ancillary statistics exist but there is no "maximal ancillary statistic" (Basu 1964). So which ancillary statistic should one condition on?

Fortunately, when a complete sufficient statistic exists, Basu's theorem assures us that we need not worry about conditioning on ancillary statistics because they are all independent of the complete sufficient statistic.

2.3 Examples of Complete Sufficient and Ancillary Statistics

Before moving to applications, it may be helpful to remember that complete sufficient statistics exist for regular full rank exponential families (Lehmann 1983, p. 46). These include, for example, the Poisson, gamma, beta, binomial, many normal distribution models (univariate, multivariate, regression, ANOVA, and so on), truncated versions (Lehmann 1983, prob. 5.31, p. 68), and censored versions (Bhattacharyya, Johnson, and Mehrotra 1977).

The order statistic $(X_{(1)}, \ldots, X_{(n)})$ from a random sample is sufficient for regular families and is complete sufficient if $\mathcal{P}$ is formed from products of members of the set of all continuous distributions or the set of all absolutely continuous distributions (Lehmann 1986, pp. 143–144). This is very useful for applications to nonparametric problems.

In location models with densities of the form $f(x-\theta)$, random variables such as $\overline{X} - \theta$ are called pivotal quantities. In scale models with densities of the form $f(x/\theta)/\theta$, $\overline{X}/\theta$ is a pivotal quantity. Pivotal quantities are similar to ancillary statistics in that their distributions are the same for each member of the family, but pivotals are not statistics because their computation requires unknown quantities. In fact, a common method of verifying ancillarity of $\boldsymbol{V}$ is to re-express it as a function of a pivotal quantity $\boldsymbol{W}$.

It is easy to show that all location-invariant statistics are ancillary if $\mathcal{P}$ is formed from a location family; all scale-invariant statistics are ancillary if $\mathcal{P}$ is formed from a scale family; and all location-and-scale-invariant statistics are ancillary if $\mathcal{P}$ is formed from a location-and-scale family.

3. TECHNICAL APPLICATIONS

Basu's theorem may be applied to a variety of problems. A selective presentation follows.

3.1 Independence of $\overline{X}$ and S^2

Suppose $\boldsymbol{X}$ represents a random sample from a $N(\mu, \sigma^2)$, distribution, where σ^2 is known. Standard exponential family results yield that the sample mean $\overline{X}$ is complete sufficient for this normal location family. Moreover, the residual vector $\boldsymbol{V} = (X_1 - \overline{X}, \ldots, X_n - \overline{X})$ is seen to be ancillary because it is location-invariant and so may be written as $((X_1 - \mu) - (\overline{X} - \mu), \ldots, (X_n - \mu) - (\overline{X} - \mu))$, which is a function of the pivotal vector $\boldsymbol{W} = (X_1 - \mu, \ldots, X_n - \mu)$ whose distribution clearly does not depend on μ. Thus, $\boldsymbol{V}$ has the same distribution for each μ.

Basu's theorem now tells us that since $\overline{X}$ is complete sufficient and S^2 is a function of the ancillary statistic $\boldsymbol{V}$, then $\overline{X}$ and S^2 are independent. Although we assumed that σ^2 is known, this result holds true for *any* normal distribution since our knowledge of σ^2 has nothing to do with the joint distribution of $\overline{X}$ and S^2. In fact, assuming σ^2 unknown only unnecessarily complicates the task of finding the joint distribution of $\overline{X}$ and S^2.

In general a key to applying Basu's theorem is deciding which parameters to assume known and which to leave unknown.

3.2 Monte Carlo Swindles

A *Monte Carlo swindle* (Johnstone and Velleman 1985) is a simulation technique that allows a small number of replications to produce statistical accuracy at the level one would expect from a much larger number of replications. In some swindles, Basu's theorem can be used to derive the basic independence needed to define the method.

Suppose, for example, that we want to estimate the variance of the sample median M of a random sample $\boldsymbol{X}$ of size n from a $N(\mu, \sigma^2)$ distribution. The straightforward approach is to generate N samples of size n from a $N(\mu, \sigma^2)$ distribution, compute the median for each sample resulting in $M_1, \dots, M_N$, and then compute the sample variance of those medians as the estimate of var(M).

The *Monte Carlo swindle* used here is to estimate var$(M - \overline{X})$ (instead of var(M)) by the N Monte Carlo samples and add var$(\overline{X}) = \sigma^2/n$ to get a more precise estimate of var(M).

Why does the swindle work? First note that as in the previous subsection, $\overline{X}$ and $\boldsymbol{V} = (X_1 - \overline{X}, \dots, X_n - \overline{X})$ are independent by Basu's theorem because $\overline{X}$ is complete sufficient and $\boldsymbol{V}$ is ancillary. Consequently, $\overline{X}$ is also independent of the median of $\boldsymbol{V}$, which is median$(X_1 - \overline{X}, \dots, X_n - \overline{X}) = M - \overline{X}$. Hence

$$\text{var}(M) = \text{var}(M - \overline{X} + \overline{X}) = \text{var}(M - \overline{X}) + \text{var}(\overline{X}).$$

Thus, from the simulation we would find the sample variance of $M_1 - \overline{X}_1, \dots, M_N - \overline{X}_N$ and add it to σ^2/n to get our estimate of var(M).

To see that the swindle estimate is more precise than the straightforward estimate, note that the variances of these estimates are approximately $2[\text{var}(M - \overline{X})]^2/(N-1)$ and $2[\text{var}(M)]^2/(N-1)$, respectively, due to the asymptotic normality of $M - \overline{X}$ and M. Thus, the swindle estimate is seen to have smaller variance than the first approach since var$(M - \overline{X}) <$ var(M) due to the high correlation between M and $\overline{X}$. In fact the asymptotic correlation is $\sqrt{2/\pi} = .798$, with var$(\overline{X})$/var$(M) \approx 2/\pi$ for large n. The result is var$(M - \overline{X})$/var$(M) \approx .36$ for large n.

More intuitively, we note that the Monte Carlo swindle approach is more efficient because the var$(\overline{X}) = \sigma^2/n$ contribution to var(M) is much larger than the contribution from var$(M - \overline{X})$ that is being estimated by simulation. The error in estimation by simulation is therefore limited to a very small part of var(M).

3.3 Moments of Ratios

It is easy to derive the following simple result: if X/Y and Y are independent and appropriate moments exist, then

$$E\left(\frac{X}{Y}\right)^k = \frac{E(X^k)}{E(Y^k)}. \qquad (1)$$

We now use Basu's theorem and (1) in several applications.

3.3.1 Score Tests

In a number of testing problems, the score statistic has the form $(\sum X_i)^2/\sum X_i^2$ and is typically compared to a chi-squared distribution with one degree of freedom. In a particular example from Liang (1985), Zhang and Boos (1997) discovered by simulation that this asymptotic testing procedure is conservative in small samples; that is, the χ_1^2 critical values are too large on average. Initially, this result seems counter-intuitive since the form of the statistic appears closer to that of a t^2 statistic than to a Z^2 statistic.

To investigate further, we may simplify the problem by assuming that $\boldsymbol{X}$ is a random sample from a $N(\mu, \sigma^2)$ distribution with σ^2 unknown and that the goal is to test the null hypothesis $\mu = 0$ using the statistic

$$t_s^2 = \frac{(\sum X_i)^2}{\sum X_i^2} = \frac{n\overline{X}^2}{\frac{1}{n}\sum X_i^2}.$$

Basu's theorem can help us calculate the null moments of t_s^2.

If $\boldsymbol{X}$ is a random sample from the scale family $N(0, \sigma^2)$, then $T = \sum X_i^2$ is complete sufficient and $V = (\sum X_i)^2/\sum X_i^2 = t_s^2$ is ancillary because it is scale-invariant and so may be written as the ratio of the two pivotal quantities $(\sum X_i/\sigma)^2$ and $\sum(X_i/\sigma)^2$. Basu's theorem therefore yields independence of T and V. In addition, result (1) tells us that

$$E(V^k) = \frac{E[(\sum X_i)^{2k}]}{E[(\sum X_i^2)^k]},$$

so that the mean and variance of t_s^2 are 1 and $2 - 6/(n+2)$. Considering that the mean and variance of the χ_1^2 distribution are 1 and 2, respectively, we can see that comparing t_s^2 to the percentiles of a χ_1^2 distribution will give conservative test levels. In a class setting, it might be worth mentioning than an exact solution to this simplified testing problem is to transform monotonically to the square of the usual one sample t statistic and use $F(1, n-1)$ distribution critical values.

3.3.2 Minimum Variance Unbiased Estimation

Suppose that $\boldsymbol{X}$ is a random sample from a Gamma(α, β) distribution when α is known. The search for minimum variance unbiased estimators is limited to functions of $T = \sum X_i$, since T is a complete sufficient statistic. One way to get unbiased functions of T is to use conditional expectations.

Consider, for example, $X_{(n)}$ the largest order statistic. If $aX_{(n)} + b$ is unbiased for the quantity to be estimated, then a minimum variance unbiased estimator is based on the the conditional expectation of $X_{(n)}$ given T. Basu's theorem helps us find this conditional expectation. Because T is complete sufficient and $V = X_{(n)}/T$ is ancillary (it is scale-invariant), T and V are also independent. Thus,

$$E(X_{(n)}|T) = E\left[X_{(n)}\frac{T}{T} \mid T\right] = E(VT|T) = TE(V).$$

Moreover, using result (1) we have that $E(V) = E(X_{(n)})/E(T)$.

3.4 Behrens–Fisher Problem

Suppose that $X_1, \dots, X_m$ is a random sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_n$ is a random sample from $N(\mu_2, \sigma_2^2)$. When no assumption is made that $\sigma_1^2 = \sigma_2^2$, the testing situation $H_0 : \mu_1 = \mu_2$ is called the Behrens–Fisher problem. Numerous approaches have been proposed including Welch's solution (see Wang 1971): compare

$$t_w = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$$

to a t distribution with estimated degrees of freedom f. Wang (1971) showed how to calculate the exact probability of a Type I error for Welch's solution:

$$P(t_w \geq t_{\alpha, f}) = P(t(\rho) \geq h(\alpha, m, n, S_1^2/S_2^2)).$$

where h is a somewhat messy function,

$$t(\rho) = \frac{\overline{X} - \overline{Y}}{\sqrt{\left(\frac{1}{m} + \frac{\rho}{n}\right) \frac{(m-1)S_1^2 + (n-1)S_2^2/\rho}{(m+n-2)}}},$$

and $\rho = \sigma_2^2/\sigma_1^2$. The key to evaluating this probability is to note that $t(\rho)$ has a t distribution with $m + n - 2$ degrees and that $t(\rho)$ is independent of S_1^2/S_2^2.

But why are $t(\rho)$ and S_1^2/S_2^2 independent? This fact was merely stated in Wang (1971) without argument. If ρ is assumed known, then S_1^2/S_2^2 is ancillary, and $t(\rho)$ is a function of the complete sufficient statistic $\boldsymbol{T} = (\overline{X}, \overline{Y}, (m-1)S_1^2 + (n-1)S_2^2/\rho)$. Basu's theorem then says that $t(\rho)$ and S_1^2/S_2^2 are independent. As with the example in Section 3.1, the assumption that ρ is known has no consequence on the joint distribution of $t(\rho)$ and S_1^2/S_2^2; so the result holds true even in the case that ρ is unknown.

As previously mentioned the trick in many applications is deciding what to assume is known in order to get both a useful complete sufficient statistic and an ancillary statistic. Consider, for example, assuming that ρ is unknown. Then the complete sufficient statistic is $\boldsymbol{T} = (\overline{X}, \overline{Y}, S_1^2, S_2^2)$, and while S_1^2/S_2^2 is a function of $\boldsymbol{T}$, $t(\rho)$ is not ancillary (even though it is pivotal under H_0).

[Received January 1997. Revised December 1997.]

REFERENCES

Basu, D. (1955), "On Statistics Independent of a Complete Sufficient Statistic," *Sankhya*, 15, 377–380.

——— (1958), "On Statistics Independent of Sufficient Statistics," *Sankhya*, 20, 223–226.

——— (1964), "Recovery of Ancillary Information," *Sankhya*, Ser. A, 26, 3-16.

——— (1982), "Basu Theorems," in *The Encyclopedia of Statistical Sciences*, eds. N. L. Johnson and S. Kotz, 1, pp. 193–196.

Bhattacharyya, G. K., Johnson, R. A., and Mehrotra, K. G. (1977), "On the Completeness of Minimal Sufficient Statistics with Censored Observations," *The Annals of Statistics*, 5, 547–553.

Casella, G., and Berger, L. R. (1990), *Statistical Inference*, Pacific Grove, CA: Wadsworth & Brooks/Cole.

Cox, D. R. (1988), "Some Aspects of Conditional and Asymptotic Inference," *Sankhya*, Series A, 30, 314–337.

Johnstone, I. M., and Velleman, P. F. (1985), "Efficient Scores, Variance Decompositions, and Monte Carlo Swindles," *Journal of The American Statistical Association*, 80, 851–862.

Lehmann, E. L. (1981), "An Interpretation of Completeness and Basu's Theorem," *Journal of The American Statistical Association*, 76, 335–340.

——— (1983), *Theory of Point Estimation*, New York: Wiley.

——— (1986), *Testing Statistical Hypotheses*, New York: Wiley.

Liang, K. Y. (1985), "Odds Ratio Inference with Dependent Data," *Biometrika*, 72, 678–682.

Wang, Y. Y. (1971), "Probabilities of the Type I Errors of the Welch Tests for the Behrens-Fisher Problem," *Journal of the American Statistical Association*, 66, 605–608.

Zhang, J., and Boos, D. D. (1997), "Mantel-Haenszel Test Statistics for Correlated Binary Data," *Biometrics*, 53, 1185–1198.