Statistical Sufficiency

Nancy Reid, University of Toronto, Toronto, ON, Canada
© 2015 Elsevier Ltd. All rights reserved.

Abstract

In the language of statistical theory, a statistic is a function of a set of data, and a sufficient statistic contains as much information about the statistical model as the original set of data. Statistical sufficiency has served as a powerful concept in the theory of inference in helping to clarify the role of models and data and in providing a framework for data reduction.

Introduction

Statistical sufficiency is a concept in the theory of statistical inference that is meant to capture an intuitive notion of summarizing a large and possibly complex set of data by relatively few summary numbers that carry the relevant information in the larger data set. For example, a public opinion poll might sample one or two thousand people, but will typically report simply the number of respondents and the percentage of respondents in each of a small number of categories. A study to investigate the effectiveness of a new treatment for weight loss may well report simply the average weight loss of a group of patients receiving the new treatment. This study would usually report as well some measure of the range of weight loss observed; variability in this setting would be a crucial piece of information for evaluating the effectiveness of the new treatment.

In order to make precise the vague notion of an informative data summary, it is necessary to have a means of defining what we mean by relevant information. This in turn is highly dependent on the use we wish to make of the data. The formal mechanism for doing this is to use a statistical model to represent an idealized version of the problem we are studying. Statistical sufficiency and information are defined relative to this model. A statistical model is a family of probability distributions, the central problem of statistical inference being to identify which member of the family generated the data currently of interest. The basic concepts of statistical models and sufficiency were set out in Fisher (1922, 1925) and continue to play a major role in the theory of inference.

For example, we might in an idealized version of the weight loss problem assume that the potential weight loss can be described by a model in the family of normal distributions. This specifies a mathematical formula for the probability that the weight loss of an individual subject falls in any possible interval, and the mathematical formula depends on two unknown parameters, the mean and standard deviation. If we are willing to further assume that the same statistical model applies to all the potential recipients of the weight loss treatment, then the data collected in a careful study will enable us to make an inference on the two parameters that depends only on two data summaries: the average weight loss in the data and the standard deviation of the weight loss in the data.

The normal distribution is also referred to as the 'bell curve,' but there are in fact a family of bell curves, each differing from the other by a simple change of mean, which fixes the location of the bell curve, and standard deviation, which fixes the scale of the bell curve. Standardized tests for measuring IQ are designed so that the scores follow a bell curve model with parameter values for the mean and standard deviation of 100 and 15, respectively. The parameters for the potential weight loss under our hypothetical new treatment are unknown, and the role of the data collected on the treatment is to shed some light on their likely values. Should the weight loss bell curve turn out to be centered at or near zero, one would be wise not to invest too heavily in the new company's stock just yet. As long as we are willing to accept the bell curve model for the weight loss data, the information about that model is contained in two numbers: the average and the standard deviation of the weight loss recorded in our data set.

An idealized model that lies behind most media reporting of polls is a Bernoulli or binomial model, in which we assume that potential respondents will choose one of two categories with a fixed but unknown probability, that respondents choose categories independently of each other, and with the same, constant probability. In such an idealized version of the problem, we have complete information about this unknown probability from the number of responses and the proportion of responses in each of the two categories.

These two examples are by construction highly idealized, and it is easy to think quickly of a number of complexities that have been overlooked. An educated reader of a report on the weight loss study would want to know how the subjects were selected to receive the new treatment, whether the subjects were in any way representative of a population of interest to that reader, whether any comparison has been made with alternative, possibly conventional, treatments, the length of time over which the study was conducted, and so on. Public opinion polls typically offer more than two categories of response, even to a yes/no question, including something equivalent to 'don't know,' or 'won't say,' and also often break down the results by gender, geographic area, or some other classification. Randomized selection of respondents by computerized telephone polls is much different than a reader response survey often conducted by magazines and media outlets. These and a wide variety of other issues are studied in the fields of experimental design and sample survey design, by statisticians and researchers in a variety of subject matter areas, from economics to geophysics. See Observational Studies: Overview; Sample Surveys: The Field and Methods; Experimental Design: Overview.

Statistical sufficiency is concerned with a much narrower issue, that of formalizing the notion that for a given family of models, a complex set of data can be summarized by a smaller
collection of numbers, and lead to the same statistical inference. The following sections describe statistical sufficiency with reference to families of models and related concepts in the theory of statistics.

Statistical Models and Sufficiency

Basic Notation

We describe a statistical model by a random variable $Y$ and a family of probability distributions for $Y, \mathcal{F}$. A random variable is a variable that takes values on a sample space, and a probability distribution describes the probability of observing a particular value or range of values in the sample space. The sample space is in a great many applications either discrete, taking values on a countable set such as (some subset of) the nonnegative integers, or continuous taking values on (some subset of) the real line $\mathbb{R}$ or $p$-dimensional Euclidean space $\mathbb{R}^p$. A typical data set will be a number of realizations of $Y$, and the goal is to use these observations to deduce which member of $\mathcal{F}$ generated these realizations. The members of $\mathcal{F}$ are often conveniently indexed by the probability mass function or probability density function $f(y)$, where

$$f(y) = \text{Prob}(Y = y) \quad [1]$$

if the sample space is discrete, and

$$f(y) = \lim_{h \to 0} h^{-1} \{\text{Prob}(Y \leq y + h) - \text{Prob}(Y \leq y)\} \quad [2]$$

if the sample space is continuous. If $Y$ is a vector then $\text{Prob}(Y \leq y)$ is defined component-wise as $\text{Prob}(Y_1 \leq y_1, \ldots, Y_p \leq y_p)$.

Very often the mathematical form of the probability distributions in $\mathcal{F}$ is determined up to a finite number of unknown constants called parameters, so we write $\mathcal{F}_\theta = \{f(\cdot; \theta); \theta \in \Theta\}$, where $\Theta$ is a set of possible values for $\theta$, such as the real numbers $\mathbb{R}$, or real-valued vectors of fixed length $p$, $\mathbb{R}^p$, or $[0, 1]$, and so on. The problem of using data to make an inference about the probability distribution is now the problem of making inference about the parameter $\theta$, and is referred to as parametric inference. Sufficiency also plays a role in nonparametric inference as well, and we illustrate this at the end of this section.

We assume in what follows that the random variable $Y = (Y_1, \ldots, Y_n)$ is a vector of length $n$ with observed value $y = (y_1, \ldots, y_n)$. A statistic is a function $T = t(Y)$ whose distribution can be computed from the distribution of $Y$. Examples of statistics include the average, $\bar{T}_1 = n^{-1} \Sigma Y_i$, the sample variance $T_2 = (n - 1)^{-1} \Sigma (Y_i - \bar{T}_1)^2$, the standard deviation $T_3 = T_2^{1/2}$, the range $T_4 = \max Y_i - \min Y_i$, and so on. The probability distribution of any statistic is obtained from the probability distribution for $Y$ by a mathematical calculation.

Sufficiency and Minimal Sufficiency

A statistic $T = t(Y)$ is sufficient for $\theta$ in the family of models $\mathcal{F} = \{f(y; \theta); \theta \in \Theta\}$ if and only if its conditional distribution

$$f_{Y|T}(y|t) \quad [3]$$

does not depend on $\theta$.

Example 1. Suppose $Y_1, \ldots, Y_n$ are independent and identically distributed with probability mass function

$$f(y; \theta) = \theta^y (1 - \theta)^{1-n}, \quad y = 0, 1; \quad \theta \in [0, 1] \quad [4]$$

which can be used to model the number of successes in $n$ independent Bernoulli trials, when the probability of success is $\theta$. The joint density for $Y = (Y_1, \ldots, Y_n)$ is

$$f(y; \theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-n-y_i} = \theta^{\Sigma y_i} (1 - \theta)^{n-\Sigma y_i} \quad [5]$$

and the marginal density for $T = \Sigma Y_i$ is

$$f(t; \theta) = \sum_{y: \Sigma y_i = t} \theta^{\Sigma y_i} (1 - \theta)^{n-\Sigma y_i} = \theta^t (1 - \theta)^{n-t} c(t, n) \quad [6]$$

where $c(t, n) = n! / t! (n-t)!$ is the number of vectors $y$ of length $n$ with $t$ ones and $n-t$ zeroes. The conditional density of $Y$ given $T$ is the ratio of [5] and [6] and is thus free of $\theta$.

The intuitive notion that the definition is meant to encapsulate is that once the sufficient statistic $T$ is observed no further information about the $\theta$ is available from the original data $Y$.

The definition of sufficiency in terms of the conditional distribution is not very convenient to work with, especially in the case of continuous sample spaces where some care is needed in the definition of a conditional distribution. The following result is more helpful:

Factorization theorem: A statistic $T = t(Y)$ is sufficient for $\theta$ in the family of models $\mathcal{F} = \{f(y; \theta); \theta \in \Theta\}$ if and only if there exist functions $g(t; \theta)$ and $h(y)$ such that for all $\theta \in \Theta$

$$f(y; \theta) = g(t; \theta) h(y) \quad [7]$$

Example 2. Let $Y_1, \ldots, Y_n$ be independent, identically distributed from the normal distribution with mean $\mu$ and standard deviation $\sigma$:

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{1}{2\sigma^2} (y - \mu)^2 \right\} \quad [8]$$

The joint density of $Y = (Y_1, \ldots, Y_n)$ is

$$f(y; \mu, \sigma) = \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \mu)^2 \right\} = \frac{1}{\sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \bar{Y})^2 + \frac{n}{\sigma^2} \bar{Y}^2 - \bar{Y}^2 \right\} \left( \frac{1}{\sqrt{2\pi}} \right)^n \quad [9]$$

where $\bar{Y} = n^{-1} \Sigma Y_i$ is the sample average, $\theta = (\mu, \sigma)$, and $\Theta = \mathbb{R} \times \mathbb{R}^+$. From [9] we see that $T = \{Y, \Sigma(Y_i - \bar{Y})^2\}$ is sufficient for $\theta$, with $g(t; \theta)$ identified with all but the final factor in [9].

Strictly speaking, there are a number of other sufficient statistics in both the examples above. In Example 2, $(\Sigma Y_i, \Sigma Y_i^2, \Sigma(Y_i - \bar{Y})^2, \Sigma(Y_i - \bar{Y})^4)$ is also sufficient, as is, trivially, the original vector $Y_i$ itself. Clearly $(\bar{Y}, \Sigma(Y_i - \bar{Y})^2)$ is in some sense ‘smaller’ than these other sufficient statistics, and to be preferred for that reason. A minimal sufficient statistic is defined to be a function of every other sufficient statistic. One-to-one functions of the minimal sufficient statistic are also minimal sufficient and are not distinguished, so in Example 2, $(\Sigma Y_i, \Sigma Y_i^2)$ is also minimal sufficient. Any statistic defines a partition of the sample space in which all sample points leading to the same value of the statistic fall in the
same partition, and the minimal sufficient statistic defines the coarsest possible partition.

Since the central problem of parametric statistical inference is to reason from the data back to the parameter \( \theta \), extensive use is made of the \textit{likelihood function}, which is (proportional to) the joint density of the data, regarded as a function of \( \theta \):

\[
L(\theta; Y) = c(Y)f(y; \theta) \tag{10}
\]

See Likelihood in Statistics. The factorization theorem suggests, and it can be proved, that the minimal sufficient statistic in the family of models \( \mathcal{F} \) is the likelihood statistic \( L(\cdot; Y) \). This result is easily derived from the factorization theorem if \( \Theta \) is a finite set, but extending it to cases of more practical interest involves some mathematical intricacies (Barndorff-Nielsen, 1978: Chapter 4; Fraser and Naderi, 2007). See Likelihood Function and Maximum Likelihood Estimation.

\textit{Example 3.} To illustrate the concept of sufficiency in a nonparametric setting, suppose that \( Y_1, \ldots, Y_n \) are independent, identically distributed on \( \mathbb{R} \) with probability density function \( f; i.e., \mathcal{F} \) consists of all probability distributions which have densities with respect to Lebesgue measure. Denote by \( Y(\cdot) \) the vector of ordered values \( Y(\cdot) = (Y(1) \leq Y(2) \leq \cdots \leq Y(n)) \). Then \( Y(\cdot) \) is sufficient for \( \mathcal{F} \), as

\[
f_{Y|Y(\cdot)}(y|y(\cdot)) = \begin{cases} \frac{1}{n!} & \text{if } y \text{ is a permutation of } y(\cdot) \\ 0 & \text{otherwise} \end{cases} \tag{11}
\]

is the same for all members of \( \mathcal{F} \).

\textit{Example 4.} Let \( Y_1, \ldots, Y_n \) be independent, identically distributed according to the uniform distribution on \( (\theta - 1, \theta + 1) \), where \( \theta \in \mathbb{R} \). The order statistic \( Y(n) \) is sufficient, as this is a special case of Example 3 above, but is not minimal sufficient. The minimal sufficient statistic is determined from the likelihood function:

\[
L(\theta; Y) = \left( \prod_{i=1}^{n} \left( \frac{1}{2} \right) \right)^{1} \frac{1}{Y(n) - \theta} \tag{12}
\]

from which we see that the minimal sufficient statistic is \( T = (Y(n), Y(1)) \), as this is a one to one function of the likelihood statistic.

\textbf{Exponential Families}

A random vector \( Y \in \mathbb{R}^p \) is said to be distributed according to an exponential family of dimension \( k \) if the probability mass or density function can be expressed in the form

\[
f(y; \theta) = \exp \left\{ \phi(\theta)^\top t(y) - c(\theta) - d(y) \right\} \tag{13}
\]

for some functions \( \phi(\theta) = (\phi_1(\theta), \ldots, \phi_k(\theta)) \), \( t(y) = (t_1(y), \ldots, t_k(y)) \), \( c(\theta) \), and \( d(y) \). Many common distributions can be expressed in exponential family form, including Example 1 with \( \phi(\theta) = \log(\theta/(1-\theta)) \), \( t(y) = \sum y_i \), and Example 2 with

\[
\phi(\theta) = \frac{\mu}{\sigma} - \frac{1}{2\sigma^2}, \quad t(y) = \left( \sum y_i, \sum y_i^2 \right) \tag{14}
\]

Other examples of exponential families include the Poisson, geometric, negative binomial, multinomial, exponential, gamma, and inverse Gaussian distributions. Generalized linear models are regression models built on exponential families that have found wide practical application. See Analysis of Variance; Generalized Linear Models.

From [13] we can see that exponential families are closed under independent, identically distributed sampling, and that the minimal sufficient statistic for \( \theta \) is the \( k \)-dimensional vector \( T = \sum t(Y_i) \). Exponential families are said to permit a \textit{sufficiency reduction} under sampling; we can replace the original \( n \) observations by a vector of length \( k \). Possibly for this reason exponential families are widely used as models in applied work.

A partial converse of this is also true, that if a finite dimensional minimal sufficient statistic exists in independent sampling from the same density, and the range of values over which the density is positive does not depend on \( \theta \), then that density must be of exponential family form.


\textbf{Some Related Concepts}

\textbf{Ancillary Statistics}

A companion concept for the reduction of data to a number of smaller quantities is that of \textit{ancillarity}. A statistic \( A = A(Y) \) is said to be ancillary for \( \theta \) in the family of models \( \mathcal{F} \) if the marginal distribution of \( A \)

\[
f_A(a) = \int_{\{y: \theta = a\}} f(y; \theta) dy \tag{15}
\]

is free of \( \theta \). If every other ancillary statistic is a function of a given ancillary statistic, the given statistic is a maximal ancillary statistic.

\textit{Example 4 cont'd.} In independent sampling from the \( U(\theta, \theta + 1) \) distribution, \( T = (Y(n), Y(1)) \) is a minimal sufficient statistic, and \( A = Y(n) - Y(1) \) is a maximal ancillary statistic.

If an ancillary statistic exists all the information about \( \theta \) is in the conditional distribution of \( Y \), given the ancillary statistic. If a minimal sufficient statistic exists, all the information about \( \theta \) is in the marginal distribution of \( T \). These two data reductions, to \( A \) or to \( T \), are complementary, and can sometimes be combined.

\textit{Example 5.} Let \( Y_1, \ldots, Y_n \) be independent, identically distributed from the one parameter location family \( f(y; \theta) = f_0(y - \theta) \), \( \theta \in \mathbb{R} \), where \( f_0(\cdot) \) is known. Then the vector of residuals \( A = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}) \) is ancillary. An extension of this to the location scale family is possible: let \( f(y; \theta) = \sigma^{-1} f_0(\sigma^{-1}(y - \mu)) \), then

\[
A = \left\{ (Y_1 - \bar{Y})/S, \ldots, (Y_n - \bar{Y})/S \right\} \tag{16}
\]

where \( S^2 = \sum (Y_i - \bar{Y})^2 \) is a maximal ancillary. The normal distribution of Example 4 is a location-scale model, with minimal sufficient statistic \( T = (\bar{Y}, S^2) \), which in this case is independent of \( A \).
A result due to Basu (1964) is that a sufficient statistic that is boundedly complete is independent of any ancillary statistic. Bounded completeness of a sufficient statistic is a technical condition that is occasionally useful in obtaining optimality results, particularly with respect to minimum variance unbiased estimation, a topic that is somewhat out of fashion.

The importance for the theory of statistics in concepts of dimension reduction such as ancillarity and sufficiency is that inference for a $k$-dimensional parameter is encapsulated by a probability distribution on a $k$-dimensional space. This immediately provides tests of significance, confidence intervals, and point estimates for the unknown parameter. From this point of view, it now appears that ancillarity is possibly a more central notion to the theory of statistics, although sufficiency is widely regarded as more natural and intuitive.

It should be noted that inference from a Bayesian point of view automatically provides a $k$-dimensional distribution for inference for a $k$-dimensional parameter, which is the posterior distribution given the observed data $y$. This is achieved at the expense of constructing a probability model for the unknown parameter $\theta$. See Bayesian Statistics.

### Asymptotic Sufficiency

The normal distribution is widely used for modeling data from a variety of sources, but is also very important in the theory of inference in providing approximate inferences, when exact inferences may be unavailable. While the likelihood function is equivalent to the minimal sufficient statistic, a special role in approximate inference is played by some functions computed from the likelihood function.

The maximum likelihood estimator of $\theta$ in the model $f(y; \theta)$ is defined by:

$$
\sup_{\theta} L(\theta; y) = L(\hat{\theta}; y)
$$

and the observed and expected information functions for $\theta$ are

$$
j(\theta) = -\delta^2 [\log(L(\theta; y))] / \theta \theta', \quad i(\theta) = E \{ j(\theta) \}
$$

Under regularity conditions on the family of models for $Y = (Y_1, \ldots, Y_n)$, we have the following asymptotic result for a scalar parameter $\theta$ as $n \to \infty$:

$$
(\hat{\theta} - \theta_0)_{n}^{1/2} \sqrt{n} \rightarrow N(0, 1)
$$

which means that for sufficiently large $n$, the distribution of the maximum likelihood estimator is arbitrarily close to a normal distribution with mean zero and variance $i^{-1}(\theta_0)$, under sampling from $f(y; \theta_0)$. A similar result is available for vector parameters. Usually in practice, the asymptotic variance of $\hat{\theta}$ will be estimated by $j^{-1}(\hat{\theta})$.

The maximum likelihood estimator has by construction the same dimension as the unknown parameter $\theta$, so [19] provides a means of constructing an inference for $\theta$ using the normal approximation. This suggests that a notion of asymptotic sufficiency could be formalized, the maximum likelihood estimator having this property. This formalization relies on showing that the likelihood function is asymptotically of normal form, with minimal sufficient statistic $\hat{\theta}$; see Fraser (1976: Chapter 8). The argument is very similar to that showing that the posterior distribution for $\theta$ is asymptotically normal. More generally, in models with smooth likelihood functions, a Taylor series expansion of the likelihood function around $\hat{\theta}$ suggests a refinement of asymptotic sufficiency: the set of statistics $\theta$ and second and higher order derivatives of the likelihood function evaluated at $\hat{\theta}$ form a set of approximately sufficient statistics in a certain sense (Barndorff-Nielsen and Cox, 1994: Chapter 7).

In general models, the maximum likelihood estimator is not the minimal sufficient statistic, although it is always a (many-to-one) function of the minimal sufficient statistic. Much of the recent development of the parametric theory of inference has been concerned with combining sufficiency or approximate sufficiency with ancillarity or an appropriate notion of approximate ancillarity, to effect a dimension reduction with minimal loss of information. In general models if the maximum likelihood estimator is used in place of the minimal sufficient statistic, information is lost. The idea of recovering this information, via conditioning on an ancillary statistic, was put forth in Fisher (1934). With some exceptions this idea was not fully exploited in the statistical literature until about 1980, when it was discovered that an asymptotic version of Fisher’s argument holds in very general situations. The past 20 years have seen continued development and refinement of Fisher’s original idea. A brief overview is given in Reid (2000), and the books Barndorff-Nielsen and Cox (1994) and Pace and Salvan (1997) give detailed accounts.

### Inference in the Presence of Nuisance Parameters

Many statistical models have vector valued parameters $\theta$, but often some components or scalar valued functions of $\theta$ are of particular interest. For example, in evaluating the hypothetical new weight loss treatment under a normal model, the mean parameter $\mu$ is in the first instance of more interest than the standard deviation $\sigma$. In studying results from a poll, we may wish to introduce parameters describing various aspects of the population, such as gender, age group, income bracket, and so on, but the main purpose of the poll is to construct an inference about the proportion likely to vote for our candidate, for example.

The idealized modeling of these types of situations is to describe $\theta$ as consisting of two components $\theta = (\psi, \lambda)$, of parameters of interest, $\psi$, and nuisance parameters, $\lambda$. An extension of this would be to define $\psi = g(\theta)$ as determined by a set of constraints on $\theta$, but only the component version will be considered here.

There are in the literature several different definitions of sufficiency for a parameter of interest $\psi$ in the presence of a nuisance parameter $\lambda$. To illustrate some of the issues we consider again the normal distribution with $\theta = (\mu, \sigma)$, $T = (\bar{Y}, S^2)$. We have

$$
f_T(t; \mu, \sigma) = f_T(\bar{Y}; \mu, \sigma)f_{S^2}(s^2; \sigma^2) = f_T(\bar{Y}; \mu, \sigma)f_{S^2}(s^2; \sigma^2)
$$

where $f_T(\cdot)$ is the density of a normal distribution with mean $\mu$ and standard deviation $\sigma/n$, and $f_{S^2}(\cdot)$ is proportional to that of a chi-squared distribution on $n-1$ degrees of freedom. As the conditional distribution of $S^2$, given $\bar{Y}$ is free of $\mu$, we could define $T$ as sufficient for $\mu$ by analogy to [3]. However, it is not sufficient for $\mu$ in the sense of providing everything we need for...
inference about $\mu$, since its distribution depends also on $\sigma$. In this latter sense $S^2$ is sufficient for $\sigma$.

There have been a number of attempts to formalize the notions of partial sufficiency and partial ancillarity, but none has been particularly successful. However, the notion of dimension reduction through exact or approximate sufficiency and ancillarity has played a very important role in the theory of parametric inference.


Complex Models

In some applications, the model is too complex to enable a likelihood function to be constructed; it may entail a large number of integrals over latent random variables, for example. Several approaches to inference in such settings involve computing some version of approximate likelihoods, either by simulation, or by ignoring some higher-order dependencies. Two examples of these approaches are the method of indirect inference (Gouriéroux and Montfort, 1996) and approximate Bayesian computation (ABC) (Marin et al., 2011); the latter is focused on Bayesian inference, but may be useful for approximate likelihood inference as well. Since the likelihood map is sufficient, this might suggest that there is no role for the concept of sufficient or approximate sufficiency in such settings, but Robert et al. (2011) show that the ABC method will not be consistent for model selection unless it is built on sufficient statistics for the model of interest. Indirect inference also uses sufficient statistics, although for an auxiliary model, as a step in simulating the likelihood of the model of interest. Connections between indirect inference and ABC methods are mentioned in Fearnhead and Prangle (2011) and a subject of ongoing research.

Conclusion

A sufficient statistic is defined relative to a statistical model, usually a parametric model, and provides all the information in the data about that model or the parameters of that model. In many cases, the sufficient statistic provides a substantial reduction of complexity or dimension from the original data, and as a result the concept of sufficiency makes the task of statistical inference simpler. The related concept of ancillarity has a similar function. Recent developments in likelihood inference have emphasized the construction of approximately sufficient and approximately ancillary statistics in cases where further dimension reduction is needed after invoking sufficiency and ancillarity. Excellent textbook references for the definitions and examples include Casella and Berger (1990) and Azzalini (1996). More advanced accounts of recent work are given in Barndorff-Nielsen and Cox (1994) and Pace and Salvan (1997).

See also: Analysis of Variance and Generalized Linear Models; Bayesian Statistics; Distributions, Statistical: Approximations; Experimental Design: Overview; Likelihood in Statistics; Nonparametric Regression; Observational Studies: Overview; Order Statistics.

Bibliography