7.6: Sufficient, Complete and Ancillary Statistics

Basic Theory

The Basic Statistical Model

Consider again the basic statistical model, in which we have a random experiment with an observable random variable $\boldsymbol{X}$ taking values in a set S. Once again, the experiment is typically to sample n objects from a population and record one or more measurements for each item. In this case, the outcome variable has the form

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \tag{7.6.1}$$

where X_i is the vector of measurements for the ith item. In general, we suppose that the distribution of $\boldsymbol{X}$ depends on a parameter θ taking values in a parameter space T. The parameter θ may also be vector-valued. We will sometimes use subscripts in probability density functions, expected values, etc. to denote the dependence on θ.

As usual, the most important special case is when $\boldsymbol{X}$ is a sequence of independent, identically distributed random variables. In this case $\boldsymbol{X}$ is a random sample from the common distribution.

Sufficient Statistics

Let $U = u(\boldsymbol{X})$ be a statistic taking values in a set R. Intuitively, U is sufficient for θ if U contains all of the information about θ that is available in the entire data variable $\boldsymbol{X}$. Here is the formal definition:

A statistic U is *sufficient* for θ if the conditional distribution of $\boldsymbol{X}$ given U does not depend on $\theta \in T$.

Sufficiency is related to the concept of *data reduction*. Suppose that $\boldsymbol{X}$ takes values in $\mathbb{R}^n$. If we can find a sufficient statistic $\boldsymbol{U}$ that takes values in $\mathbb{R}^j$, then we can reduce the original data vector $\boldsymbol{X}$ (whose dimension n is usually large) to the vector of statistics $\boldsymbol{U}$ (whose dimension j is usually much smaller) with no loss of information about the parameter θ.

The following result gives a condition for sufficiency that is equivalent to this definition.

Let $U = u(\boldsymbol{X})$ be a statistic taking values in R, and let f_θ and h_θ denote the probability density functions of $\boldsymbol{X}$ and U respectively. Then U is suffcient for θ if and only if the function on S given below does not depend on $\theta \in T$:

$$\boldsymbol{x} \mapsto \frac{f_\theta(\boldsymbol{x})}{h_\theta[u(\boldsymbol{x})]} \tag{7.6.2}$$

Proof

The joint distribution of $(\boldsymbol{X}, U)$ is concentrated on the set $\{(\boldsymbol{x}, y) : \boldsymbol{x} \in S, y = u(\boldsymbol{x})\} \subseteq S \times R$. The conditional PDF of $\boldsymbol{X}$ given $U = u(\boldsymbol{x})$ is $f_\theta(\boldsymbol{x}) \big/ h_\theta[u(\boldsymbol{x})]$ on this set, and is 0 otherwise.

The definition precisely captures the intuitive notion of sufficiency given above, but can be difficult to apply. We must know in advance a candidate statistic U, and then we must be able to compute the conditional distribution of $\boldsymbol{X}$ given U. The *Fisher-Neyman factorization theorem* given next often allows the identification of a sufficient statistic from the form of the probability density function of $\boldsymbol{X}$. It is named for Ronald Fisher and Jerzy Neyman.

Fisher-Neyman Factorization Theorem. Let f_θ denote the probability density function of $\boldsymbol{X}$ and suppose that $U = u(\boldsymbol{X})$ is a statistic taking values in R. Then U is sufficient for θ if and only if there exists $G : R \times T \to [0, \infty)$ and $r : S \to [0, \infty)$ such that

$$f_\theta(\boldsymbol{x}) = G[u(\boldsymbol{x}), \theta] r(\boldsymbol{x}); \quad \boldsymbol{x} \in S, \ \theta \in T \tag{7.6.3}$$

Proof

Let h_θ denote the PDF of U for $\theta \in T$. If U is sufficient for θ, then from the previous theorem, the function $r(\boldsymbol{x}) = f_\theta(\boldsymbol{x}) \big/ h_\theta[u(\boldsymbol{x})]$ for $\boldsymbol{x} \in S$ does not depend on $\theta \in T$. Hence $f_\theta(\boldsymbol{x}) = h_\theta[u(\boldsymbol{x})] r(\boldsymbol{x})$ for $(\boldsymbol{x}, \theta) \in S \times T$ and so $(\boldsymbol{x}, \theta) \mapsto f_\theta(\boldsymbol{x})$ has the form given in the theorem. Conversely, suppose that $(\boldsymbol{x}, \theta) \mapsto f_\theta(\boldsymbol{x})$ has the form given in the theorem. Then there exists a positive constant C such that $h_\theta(y) = C G(y, \theta)$ for $\theta \in T$ and $y \in R$. Hence $f_\theta(\boldsymbol{x}) \big/ h_\theta[u(x)] = r(\boldsymbol{x}) / C$ for $\boldsymbol{x} \in S$, independent of $\theta \in T$.

Note that r depends only on the data $\boldsymbol{x}$ but not on the parameter θ. Less technically, $u(\boldsymbol{X})$ is sufficient for θ if the probability density function $f_\theta(\boldsymbol{x})$ depends on the data vector $\boldsymbol{x}$ and the parameter θ only through $u(\boldsymbol{x})$.

If U and V are equivalent statistics and U is sufficient for θ then V is sufficient for θ.

Minimal Sufficient Statistics

The entire data variable $\boldsymbol{X}$ is trivially sufficient for θ. However, as noted above, there usually exists a statistic U that is sufficient for θ and has smaller dimension, so that we can achieve real data reduction. Naturally, we would like to find the statistic U that has the smallest dimension possible. In many cases, this smallest dimension j will be the same as the dimension k of the parameter vector θ. However, as we will see, this is not necessarily the case; j can be smaller or larger than k. An example based on the uniform distribution is given in (38).

Suppose that a statistic U is sufficient for θ. Then U is *minimally sufficient* if U is a function of any other statistic V that is sufficient for θ.

Once again, the definition precisely captures the notion of minimal sufficiency, but is hard to apply. The following result gives an equivalent condition.

Let f_θ denote the probability density function of $\boldsymbol{X}$ corresponding to the parameter value $\theta \in T$ and suppose that $U = u(\boldsymbol{X})$ is a statistic taking values in R. Then U is minimally sufficient for θ if the following condition holds: for $\boldsymbol{x} \in S$ and $\boldsymbol{y} \in S$

$$\frac{f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{y})} \text{ is independent of } \theta \text{ if and only if } u(\boldsymbol{x}) = u(\boldsymbol{y}) \tag{7.6.4}$$

Proof

Suppose that the condition in the theorem is satisfied. Then the PDF f_θ of $\boldsymbol{X}$ must have the form given in the factorization theorem (3) so U is sufficient for θ. Next, suppose that $V = v(\boldsymbol{X})$ is another sufficient statistic for θ, taking values in R. From the factorization theorem, there exists $G : R \times T \to [0, \infty)$ and $r : S \to [0, \infty)$ such that $f_\theta(\boldsymbol{x}) = G[v(\boldsymbol{x}), \theta] r(\boldsymbol{x})$ for $(\boldsymbol{x}, \theta) \in S \times T$. Hence if $\boldsymbol{x}, \boldsymbol{y} \in S$ and $v(\boldsymbol{x}) = v(\boldsymbol{y})$ then

$$\frac{f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{y})} = \frac{G[v(\boldsymbol{x}), \theta] r(\boldsymbol{x})}{G[v(\boldsymbol{y}), \theta] r(\boldsymbol{y})} = \frac{r(\boldsymbol{x})}{r(\boldsymbol{y})} \tag{7.6.5}$$

does not depend on $\theta \in \Theta$. Hence from the condition in the theorem, $u(\boldsymbol{x}) = u(\boldsymbol{y})$ and it follows that U is a function of V.

If U and V are equivalent statistics and U is minimally sufficient for θ then V is minimally sufficient for θ.

Properties of Sufficient Statistics

Sufficiency is related to several of the methods of constructing estimators that we have studied.

Suppose that U is sufficient for θ and that there exists a maximum likelihood estimator of θ. Then there exists a maximum likelihood estimator V that is a function of U.

Proof

From the factorization theorem (3), the log likelihood function for $\boldsymbol{x} \in S$ is

$$\theta \mapsto \ln G[u(\boldsymbol{x}), \theta] + \ln r(\boldsymbol{x}) \tag{7.6.6}$$

Hence a value of θ that maximizes this function, if it exists, must be a function of $u(\boldsymbol{x})$.

In particular, suppose that V is the unique maximum likelihood estimator of θ and that V is sufficient for θ. If U is sufficient for θ then V is a function of U by the previous theorem. Hence it follows that V is minimally sufficient for θ. Our next result applies to Bayesian analysis.

Suppose that the statistic $U = u(\boldsymbol{X})$ is sufficient for the parameter θ and that θ is modeled by a random variable Θ with values in T. Then the posterior distribution of Θ given $\boldsymbol{X} = \boldsymbol{x} \in S$ is a function of $u(\boldsymbol{x})$.

Proof

Let h denote the prior PDF of Θ and $f(\cdot \mid \theta)$ the conditional PDF of $\boldsymbol{X}$ given $\Theta = \theta \in T$. By the factorization theorem (3), this conditional PDF has the form $f(\boldsymbol{x} \mid \theta) = G[u(\boldsymbol{x}), \theta] r(\boldsymbol{x})$ for $\boldsymbol{x} \in S$ and $\theta \in T$. The posterior PDF of Θ given $\boldsymbol{X} = \boldsymbol{x} \in S$ is

$$h(\theta \mid \boldsymbol{x}) = \frac{h(\theta) f(\boldsymbol{x} \mid \theta)}{f(\boldsymbol{x})}, \quad \theta \in T \tag{7.6.7}$$

where the function in the denominator is the marginal PDF of $\boldsymbol{X}$, or simply the normalizing constant for the function of θ in the numerator. Let's suppose that Θ has a continuous distribution on T, so that $f(\boldsymbol{x}) = \int_T h(t) G[u(\boldsymbol{x}), t] r(\boldsymbol{x}) dt$ for $\boldsymbol{x} \in S$. Then the posterior PDF simplifies to

$$h(\theta \mid \boldsymbol{x}) = \frac{h(\theta) G[u(\boldsymbol{x}), \theta]}{\int_T h(t) G[u(\boldsymbol{x}), t] dt} \tag{7.6.8}$$

which depends on $\boldsymbol{x} \in S$ only through $u(\boldsymbol{x})$.

Continuing with the setting of Bayesian analysis, suppose that θ is a real-valued parameter. If we use the usual mean-square loss function, then the Bayesian estimator is $V = \mathbb{E}(\Theta \mid \boldsymbol{X})$. By the previous result, V is a function of the sufficient statistics U. That is, $\mathbb{E}(\Theta \mid \boldsymbol{X}) = \mathbb{E}(\Theta \mid U)$.

The next result is the *Rao-Blackwell theorem*, named for CR Rao and David Blackwell. The theorem shows how a sufficient statistic can be used to improve an unbiased estimator.

Rao-Blackwell Theorem. Suppose that U is sufficient for θ and that V is an unbiased estimator of a real parameter $\lambda = \lambda(\theta)$. Then $\mathbb{E}_\theta(V \mid U)$ is also an unbiased estimator of λ and is uniformly better than V.

Proof

This follows from basic properties of conditional expected value and conditional variance. First, since V is a function of $\boldsymbol{X}$ and U is sufficient for θ, $\mathbb{E}_\theta(V \mid U)$ is a valid statistic; that is, it does not depend on θ, in spite of the formal dependence on θ in the expected value. Next, $\mathbb{E}_\theta(V \mid U)$ is a

function of U and $\mathbb{E}_\theta[\mathbb{E}_\theta(V \mid U)] = \mathbb{E}_\theta(V) = \lambda$ for $\theta \in \Theta$. Thus $\mathbb{E}_\theta(V \mid U)$ is an unbiased estimator of λ. Finally $\text{var}_\theta[\mathbb{E}_\theta(V \mid U)] = \text{var}_\theta(V) - \mathbb{E}_\theta[\text{var}_\theta(V \mid U)] \le \text{var}_\theta(V)$ for any $\theta \in T$.

Complete Statistics

Suppose that $U = u(\boldsymbol{X})$ is a statistic taking values in a set R. Then U is a *complete* statistic for θ if for any function $r : R \to \mathbb{R}$

$$\mathbb{E}_\theta[r(U)] = 0 \text{ for all } \theta \in T \implies \mathbb{P}_\theta[r(U) = 0] = 1 \text{ for all } \theta \in T \tag{7.6.9}$$

To understand this rather strange looking condition, suppose that $r(U)$ is a statistic constructed from U that is being used as an estimator of 0 (thought of as a function of θ). The completeness condition means that the only such unbiased estimator is the statistic that is 0 with probability 1.

If U and V are equivalent statistics and U is complete for θ then V is complete for θ.

The next result shows the importance of statistics that are both complete and sufficient; it is known as the *Lehmann-Scheffé theorem*, named for Erich Lehmann and Henry Scheffé.

Lehmann-Scheffé Theorem. Suppose that U is sufficient and complete for θ and that $V = r(U)$ is an unbiased estimator of a real parameter $\lambda = \lambda(\theta)$. Then V is a uniformly minimum variance unbiased estimator (UMVUE) of λ.

Proof

Suppose that W is an unbiased estimator of λ. By the Rao-Blackwell theorem (10), $\mathbb{E}(W \mid U)$ is also an unbiased estimator of λ and is uniformly better than W. Since $\mathbb{E}(W \mid U)$ is a function of U, it follows from completeness that $V = \mathbb{E}(W \mid U)$ with probability 1.

Ancillary Statistics

Suppose that $V = v(\boldsymbol{X})$ is a statistic taking values in a set R. If the distribution of V does not depend on θ, then V is called an *ancillary* statistic for θ.

Thus, the notion of an ancillary statistic is complementary to the notion of a sufficient statistic. A sufficient statistic contains all available information about the parameter; an ancillary statistic contains no information about the parameter. The following result, known as *Basu's Theorem* and named for Debabrata Basu, makes this point more precisely.

Basu's Theorem. Suppose that U is complete and sufficient for a parameter θ and that V is an ancillary statistic for θ. Then U and V are independent.

Proof

Let g denote the probability density function of V and let $v \mapsto g(v \mid U)$ denote the conditional probability density function of V given U. From properties of conditional expected value, $\mathbb{E}[g(v \mid U)] = g(v)$ for $v \in R$. But then from completeness, $g(v \mid U) = g(v)$ with probability 1.

If U and V are equivalent statistics and U is ancillary for θ then V is ancillary for θ.

Applications and Special Distributions

In this subsection, we will explore sufficient, complete, and ancillary statistics for a number of special distributions. As always, be sure to try the problems yourself before looking at the solutions.

The Bernoulli Distribution

Recall that the *Bernoulli distribuiton* with parameter $p \in (0, 1)$ is a discrete distribution on $\{0, 1\}$ with probability density function g defined by

$$g(x) = p^x (1 - p)^{1 - x}, \quad x \in \{0, 1\} \tag{7.6.10}$$

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the Bernoulli distribution with parameter p. Equivalently, $\boldsymbol{X}$ is a sequence of *Bernoulli trials*, so that in the usual langauage of reliability, $X_i = 1$ if trial i is a success, and $X_i = 0$ if trial i is a failure. The Bernoulli distribution is named for Jacob Bernoulli and is studied in more detail in the chapter on Bernoulli Trials

Let $Y = \sum_{i=1}^n X_i$ denote the number of successes. Recall that Y has the binomial distribution with parameters n and p, and has probability density function h defined by

$$h(y) = \binom{n}{y} p^y (1 - p)^{n - y}, \quad y \in \{0, 1, \ldots, n\} \tag{7.6.11}$$

Y is sufficient for p. Specifically, for $y \in \{0, 1, \ldots, n\}$, the conditional distribution of $\boldsymbol{X}$ given $Y = y$ is uniform on the set of points

$$D_y = \{(x_1, x_2, \ldots, x_n) \in \{0, 1\}^n : x_1 + x_2 + \cdots + x_n = y\} \tag{7.6.12}$$

Proof

The joint PDF f of $\boldsymbol{X}$ is defined by

$$f(\boldsymbol{x}) = g(x_1)g(x_2)\cdots g(x_n) = p^y(1-p)^{n-y}, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n \tag{7.6.13}$$

where $y = \sum_{i=1}^n x_i$. Now let $y \in \{0, 1, \ldots, n\}$ Given $Y = y$, $\boldsymbol{X}$ is concentrated on D_y and

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y) = \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(Y = y)} = \frac{p^y(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}}, \quad \boldsymbol{x} \in D_y \tag{7.6.14}$$

Of course, $\binom{n}{y}$ is the cardinality of D_y.

This result is intuitively appealing: in a sequence of Bernoulli trials, all of the information about the probability of success p is contained in the number of successes Y. The particular *order* of the successes and failures provides no additional information. Of course, the sufficiency of Y follows more easily from the factorization theorem (3), but the conditional distribution provides additional insight.

Y is complete for p on the parameter space $(0, 1)$.

Proof

If $r : \{0, 1, \ldots, n\} \to \mathbb{R}$, then

$$\mathbb{E}[r(Y)] = \sum_{y=0}^n r(y) \binom{n}{k} p^y (1-p)^{n-y} = (1-p)^n \sum_{y=0}^n r(y) \binom{n}{y} \left(\frac{p}{1-p}\right)^y \tag{7.6.15}$$

The last sum is a polynomial in the variable $t = \frac{p}{1-p} \in (0, \infty)$. If this polynomial is 0 for all $t \in (0, \infty)$, then all of the coefficients must be 0. Hence we must have $r(y) = 0$ for $y \in \{0, 1, \ldots, n\}$

The proof of the last result actually shows that if the parameter space is any subset of $(0, 1)$ containing an interval of positive length, then Y is complete for p. But the notion of completeness depends very much on the parameter space. The following result considers the case where p has a finite set of values.

Suppose that the parameter space $T \subset (0, 1)$ is a finite set with $k \in \mathbb{N}_+$ elements. If the sample size n is at least k, then Y is not complete for p.

Proof

Suppose that $r : \{0, 1, \ldots, n\} \to \mathbb{R}$ and that $\mathbb{E}[r(Y)] = 0$ for $p \in T$. Then we have

$$\sum_{y=0}^n \binom{n}{y} p^y (1-p)^{n-y} r(y) = 0, \quad p \in T \tag{7.6.16}$$

This is a set of k linear, homogenous equations in the variables $(r(0), r(1), \ldots, r(n))$. Since $n \geq k$, we have at least $k+1$ variables, so there are infinitely many nontrivial solutions.

The sample mean $M = Y/n$ (the sample proportion of successes) is clearly equivalent to Y (the number of successes), and hence is also sufficient for p and is complete for $p \in (0, 1)$. Recall that the sample mean M is the method of moments estimator of p, and is the maximum likelihood estimator of p on the parameter space $(0, 1)$.

In Bayesian analysis, the usual approach is to model p with a random variable P that has a prior beta distribution with left parameter $a \in (0, \infty)$ and right parameter $b \in (0, \infty)$. Then the posterior distribution of P given $\boldsymbol{X}$ is beta with left parameter $a + Y$ and right parameter $b + (n - Y)$. The posterior distribution depends on the data only through the sufficient statistic Y, as guaranteed by theorem (9).

The sample variance S^2 is an UMVUE of the distribution variance $p(1-p)$ for $p \in (0, 1)$, and can be written as

$$S^2 = \frac{Y}{n-1}\left(1 - \frac{Y}{n}\right) \tag{7.6.17}$$

Proof

Recall that the sample variance can be written as

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1} M^2 \tag{7.6.18}$$

But $X_i^2 = X_i$ since X_i is an indicator variable, and $M = Y/n$. Substituting gives the representation above. In general, S^2 is an unbiased estimator of the distribution variance σ^2. But in this case, S^2 is a function of the complete, sufficient statistic Y, and hence by the Lehmann Scheffé theorem (13), S^2 is an UMVUE of $\sigma^2 = p(1-p)$.

The Poisson Distribution

Recall that the *Poisson distribution* with parameter $\theta \in (0, \infty)$ is a discrete distribution on $\mathbb{N}$ with probability density function g defined by

$$g(x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x \in \mathbb{N} \tag{7.6.19}$$

The Poisson distribution is named for Simeon Poisson and is used to model the number of "random points" in region of time or space, under certain ideal conditions. The parameter θ is proportional to the size of the region, and is both the mean and the variance of the distribution. The Poisson distribution is studied in more detail in the chapter on Poisson process.

Suppose now that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the Poisson distribution with parameter θ. Recall that the sum of the scores $Y = \sum_{i=1}^n X_i$ also has the Poisson distribution, but with parameter $n\theta$.

The statistic Y is sufficient for θ. Specifically, for $y \in \mathbb{N}$, the conditional distribution of $\boldsymbol{X}$ given $Y = y$ is the multinomial distribution with y trials, n trial values, and uniform trial probabilities.

Proof

The joint PDF f of $\boldsymbol{X}$ is defined by

$$f(\boldsymbol{x}) = g(x_1)g(x_2) \cdot g(x_n) = \frac{e^{-n\theta}\theta^y}{x_1!x_2! \cdots x_n!}, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n \tag{7.6.20}$$

where $y = \sum_{i=1}^n x_i$. Given $Y = y \in \mathbb{N}$, random vector $\boldsymbol{X}$ takes values in the set $D_y = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{N}^n : \sum_{i=1}^n x_i = y\}$. Moreover,

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y) = \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(Y = y)} = \frac{e^{-n\theta}\theta^y / (x_1!x_2! \cdots x_n!)}{e^{-n\theta}(n\theta)^y / y!} = \frac{y!}{x_1!x_2! \cdots x_n!}\frac{1}{n^y}, \quad \boldsymbol{x} \in D_y \tag{7.6.21}$$

The last expression is the PDF of the multinomial distribution stated in the theorem. Of course, the important point is that the conditional distribution does not depend on θ.

As before, it's easier to use the factorization theorem to prove the sufficiency of Y, but the conditional distribution gives some additional insight.

Y is complete for $\theta \in (0, \infty)$.

Proof

If $r : \mathbb{N} \to \mathbb{R}$ then

$$\mathbb{E}[r(Y)] = \sum_{y=0}^\infty e^{-n\theta}\frac{(n\theta)^y}{y!}r(y) = e^{-n\theta}\sum_{y=0}^\infty \frac{n^y}{y!}r(y)\theta^y \tag{7.6.22}$$

The last sum is a power series in θ with coefficients $n^y r(y)/y!$ for $y \in \mathbb{N}$. If this series is 0 for all θ in an open interval, then the coefficients must be 0 and hence $r(y) = 0$ for $y \in \mathbb{N}$.

As with our discussion of Bernoulli trials, the sample mean $M = Y/n$ is clearly equivalent to Y and hence is also sufficient for θ and complete for $\theta \in (0, \infty)$. Recall that M is the method of moments estimator of θ and is the maximum likelihood estimator on the parameter space $(0, \infty)$.

An UMVUE of the parameter $\mathbb{P}(X = 0) = e^{-\theta}$ for $\theta \in (0, \infty)$ is

$$U = \left(\frac{n-1}{n}\right)^Y \tag{7.6.23}$$

Proof

The probability generating function of Y is

$$P(t) = \mathbb{E}(t^Y) = e^{n\theta(t-1)}, \quad t \in \mathbb{R} \tag{7.6.24}$$

Hence

$$\mathbb{E}\left[\left(\frac{n-1}{n}\right)^Y\right] = \exp\left[n\theta\left(\frac{n-1}{n} - 1\right)\right] = e^{-\theta}, \quad \theta \in (0, \infty) \tag{7.6.25}$$

So $U = [(n-1)/n]^Y$ is an unbiased estimator of $e^{-\theta}$. Since U is a function of the complete, sufficient statistic Y, it follows from the Lehmann Scheffé theorem (13) that U is an UMVUE of $e^{-\theta}$.

The Normal Distribution

Recall that the *normal distribution* with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$ is a continuous distribution on $\mathbb{R}$ with probability density function g defined by

$$g(x) = \frac{1}{\sqrt{2\,\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad x \in \mathbb{R} \tag{7.6.26}$$

The normal distribution is often used to model physical quantities subject to small, random errors, and is studied in more detail in the chapter on Special Distributions. Because of the central limit theorem, the normal distribution is perhaps the most important distribution in statistics.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the normal distribution with mean μ and variance σ^2. Then each of the following pairs of statistics is minimally sufficient for (μ, σ^2)

1. (Y, V) where $Y = \sum_{i=1}^n X_i$ and $V = \sum_{i=1}^n X_i^2$.
2. (M, S^2) where $M = \frac{1}{n} \sum_{i=1}^n X_i$ is the sample mean and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - M)^2$ is the sample variance.
3. (M, T^2) where $T^2 = \frac{1}{n} \sum_{i=1}^n (X_i - M)^2$ is the biased sample variance.

Proof

1. The joint PDF f of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = g(x_1) g(x_2) \cdots g(x_n) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right], \quad \boldsymbol{x} = (x_1, x_2 \ldots, x_n) \in \mathbb{R}^n \tag{7.6.27}$$

   After some algebra, this can be written as

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-n\mu^2/\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{2\mu}{\sigma^2} \sum_{i=1}^n x_i\right), \quad \boldsymbol{x} = (x_1, x_2 \ldots, x_n) \in \mathbb{R}^n \tag{7.6.28}$$

   It follows from the factorization theorem (3) that (Y, V) is sufficient for (μ, σ^2). Minimal sufficiency follows from the condition in theorem (6).
2. Note that $M = \frac{1}{n} Y$, $S^2 = \frac{1}{n-1} V - \frac{n}{n-1} M^2$. Hence (M, S^2) is equivalent to (Y, V) and so (M, S^2) is also minimally sufficient for (μ, σ^2).
3. Similarly, $M = \frac{1}{n} Y$ and $T^2 = \frac{1}{n} V - M^2$. Hence (M, T^2) is equivalent to (Y, V) and so (M, T^2) is also minimally sufficient for (μ, σ^2).

Recall that M and T^2 are the method of moments estimators of μ and σ^2, respectively, and are also the maximum likelihood estimators on the parameter space $\mathbb{R} \times (0, \infty)$.

Run the normal estimation experiment 1000 times with various values of the parameters. Compare the estimates of the parameters in terms of bias and mean square error.

Sometimes the variance σ^2 of the normal distribution is known, but not the mean μ. It's rarely the case that μ is known but not σ^2. Nonetheless we can give sufficient statistics in both cases.

Suppose again that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the normal distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$. If

1. If σ^2 is known then $Y = \sum_{i=1}^n X_i$ is minimally sufficient for μ.
2. If μ is known then $U = \sum_{i=1}^n (X_i - \mu)^2$ is sufficient for σ^2.

Proof

1. This results follow from the second displayed equation for the PDF $f(\boldsymbol{x})$ of $\boldsymbol{X}$ in the proof of the previous theorem.
2. This result follows from the first displayed equation for the PDF $f(\boldsymbol{x})$ of bsX in the proof of the previous theorem.

Of course by equivalence, in part (a) the sample mean $M = Y/n$ is minimally sufficient for μ, and in part (b) the special sample variance $W = U/n$ is minimally sufficient for σ^2. Moreover, in part (a), M is complete for μ on the parameter space $\mathbb{R}$ and the sample variance S^2 is ancillary for μ (Recall that $(n-1)S^2/\sigma^2$ has the chi-square distribution with $n-1$ degrees of freedom.) It follows from Basu's theorem (15) that the sample mean M and the sample variance S^2 are independent. We proved this by more direct means in the section on special properties of normal samples, but the formulation in terms of sufficient and ancillary statistics gives additional insight.

The Gamma Distribution

Recall that the *gamma distribution* with shape parameter $k \in (0, \infty)$ and scale parameter $b \in (0, \infty)$ is a continuous distribution on $(0, \infty)$ with probability density function g given by

$$g(x) = \frac{1}{\Gamma(k) b^k} x^{k-1} e^{-x/b}, \quad x \in (0, \infty) \tag{7.6.29}$$

The gamma distribution is often used to model random times and certain other types of positive random variables, and is studied in more detail in the chapter on Special Distributions.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the gamma distribution with shape parameter k and scale parameter b. Each of the following pairs of statistics is minimally sufficient for (k, b)

1. (Y, V) where $Y = \sum_{i=1}^n X_i$ is the sum of the scores and $V = \prod_{i=1}^n X_i$ is the product of the scores.
2. (M, U) where $M = Y/n$ is the sample (arithmetic) mean of $\boldsymbol{X}$ and $U = V^{1/n}$ is the sample geometric mean of $\boldsymbol{X}$.

Proof

1. The joint PDF f of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = g(x_1) g(x_2) \cdots g(x_n) = \frac{1}{\Gamma^n(k) b^{nk}} (x_1 x_2 \ldots x_n)^{k-1} e^{-(x_1 + x_2 + \cdots + x_n)/b}, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in (0, \infty)^n \tag{7.6.30}$$

   From the factorization theorem (3), (Y, V) is sufficient for (k, b). Minimal sufficiency follows from condition (6).

2. Clearly $M = Y/n$ is equivalent to Y and $U = V^{1/n}$ is equivalent to V. Hence (M, U) is also minimally sufficient for (k, b).

Recall that the method of moments estimators of k and b are M^2/T^2 and T^2/M, respectively, where $M = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean and $T^2 = \frac{1}{n}\sum_{i=1}^n (X_i - M)^2$ is the biased sample variance. If the shape parameter k is known, $\frac{1}{k}M$ is both the method of moments estimator of b and the maximum likelihood estimator on the parameter space $(0, \infty)$. Note that T^2 is not a function of the sufficient statistics (Y, V), and hence estimators based on T^2 suffer from a loss of information.

Run the gamma estimation experiment 1000 times with various values of the parameters and the sample size n. Compare the estimates of the parameters in terms of bias and mean square error.

The proof of the last theorem actually shows that Y is sufficient for b if k is known, and that V is sufficient for k if b is known.

Suppose again that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample of size n from the gamma distribution with shape parameter $k \in (0, \infty)$ and scale parameter $b \in (0, \infty)$. Then $Y = \sum_{i=1}^n X_i$ is complete for b.

Proof

Y has the gamma distribution with shape parameter nk and scale parameter b. Hence, if $r : [0, \infty) \to \mathbb{R}$, then

$$\mathbb{E}[r(Y)] = \int_0^\infty \frac{1}{\Gamma(nk) b^{nk}} y^{nk-1} e^{-y/b} r(y)\, dy = \frac{1}{\Gamma(nk) b^{nk}} \int_0^\infty y^{nk-1} r(y) e^{-y/b}\, dy \tag{7.6.31}$$

The last integral can be interpreted as the Laplace transform of the function $y \mapsto y^{nk-1} r(y)$ evaluated at $1/b$. If this transform is 0 for all b in an open interval, then $r(y) = 0$ almost everywhere in $(0, \infty)$.

Suppose again that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the gamma distribution on $(0, \infty)$ with shape parameter $k \in (0, \infty)$ and scale parameter $b \in (0, \infty)$. Let $M = \frac{1}{n}\sum_{i=1}^n X_i$ denote the sample mean and $U = (X_1 X_2 \ldots X_n)^{1/n}$ the sample geometric mean, as before. Then

1. M/U is ancillary for b.
2. M and M/U are independent.

Proof

1. We can take $X_i = b Z_i$ for $i \in \{1, 2, \ldots, n\}$where $\boldsymbol{Z} = (Z_1, X_2, \ldots, Z_n)$ is a random sample of size n from the gamma distribution with shape parameter k and scale parameter 1 (the *standard* gamma distribution with shape parameter k). Then

$$\frac{M}{U} = \frac{1}{n}\sum_{i=1}^n \frac{X_i}{(X_1 X_2 \cdots X_n)^{1/n}} = \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i^n}{X_1 X_2 \cdots X_n}\right)^{1/n} = \frac{1}{n}\sum_{i=1}^n \left(\prod_{j \neq i} \frac{X_i}{X_j}\right)^{1/n} \tag{7.6.32}$$

   But $X_i / X_j = Z_i / Z_j$ for $i \neq j$, and the distribution of $\{Z_i / Z_j : i, j \in \{1, 2, \ldots, n\},\ i \neq j\}$ does not depend on b. Hence the distribution of M/U does not depend on b.
2. This follows from Basu's theorem (15), since M is complete and sufficient for b and M/U is ancillary for b.

The Beta Distribution

Recall that the *beta distribution* with left parameter $a \in (0, \infty)$ and right parameter $b \in (0, \infty)$ is a continuous distribution on $(0, 1)$ with probability density function g given by

$$g(x) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1}, \quad x \in (0, 1) \tag{7.6.33}$$

where B is the beta function. The beta distribution is often used to model random proportions and other random variables that take values in bounded intervals. It is studied in more detail in the chapter on Special Distribution

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the beta distribution with left parameter a and right parameter b. Then (P, Q) is minimally sufficient for (a, b) where $P = \prod_{i=1}^n X_i$ and $Q = \prod_{i=1}^n (1 - X_i)$.

Proof

The joint PDF f of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = g(x_1) g(x_2) \cdots g(x_n) = \frac{1}{B^n(a, b)} (x_1 x_2 \cdots x_n)^{a-1} [(1 - x_1)(1 - x_2) \cdots (1 - x_n)]^{b-1}, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in (0, 1)^n \tag{7.6.34}$$

From the factorization theorem (3), it follows that (U, V) is sufficient for (a, b). Minimal sufficiency follows from condition (6).

The proof also shows that P is sufficient for a if b is known, and that Q is sufficient for b if a is known. Recall that the method of moments estimators of a and b are

$$U = \frac{M\left(M - M^{(2)}\right)}{M^{(2)} - M^2}, \quad V = \frac{(1 - M)\left(M - M^{(2)}\right)}{M^{(2)} - M^2} \tag{7.6.35}$$

respectively, where $M = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean and $M^{(2)} = \frac{1}{n}\sum_{i=1}^n X_i^2$ is the second order sample mean. If b is known, the method of moments estimator of a is $U_b = bM/(1-M)$, while if a is known, the method of moments estimator of b is $V_a = a(1-M)/M$. None of these estimators is a function of the sufficient statistics (P, Q) and so all suffer from a loss of information. On the other hand, if $b = 1$, the maximum likelihood estimator of a on the interval $(0, \infty)$ is $W = -n/\sum_{i=1}^n \ln X_i$, which is a function of P (as it must be).

Run the beta estimation experiment 1000 times with various values of the parameters. Compare the estimates of the parameters.

The Pareto Distribution

Recall that the *Pareto distribution* with shape parameter $a \in (0, \infty)$ and scale parameter $b \in (0, \infty)$ is a continuous distribution on $[b, \infty)$ with probability density function g given by

$$g(x) = \frac{ab^a}{x^{a+1}}, \quad b \le x < \infty \tag{7.6.36}$$

The Pareto distribution, named for Vilfredo Pareto, is a heavy-tailed distribution often used to model income and certain other types of random variables. It is studied in more detail in the chapter on Special Distribution.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the Pareto distribution with shape parameter a and scale parameter b. Then $\left(P, X_{(1)}\right)$ is minimally sufficient for (a, b) where $P = \prod_{i=1}^n X_i$ is the product of the sample variables and where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ is the first order statistic.

Proof

The joint PDF f of $\boldsymbol{X}$ at $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is given by

$$f(\boldsymbol{x}) = g(x_1)g(x_2)\cdots g(x_n) = \frac{a^n b^{na}}{(x_1 x_2 \cdots x_n)^{a+1}}, \quad x_1 \ge b, x_2 \ge b, \ldots, x_n \ge b \tag{7.6.37}$$

which can be rewritten as

$$f(\boldsymbol{x}) = g(x_1)g(x_2)\cdots g(x_n) = \frac{a^n b^{na}}{(x_1 x_2 \cdots x_n)^{a+1}} \mathbf{1}\left(x_{(n)} \ge b\right), \quad (x_1, x_2, \ldots, x_n) \in (0, \infty)^n \tag{7.6.38}$$

So the result follows from the factorization theorem (3). Minimal sufficiency follows from condition (6).

The proof also shows that P is sufficient for a if b is known (which is often the case), and that $X_{(1)}$ is sufficient for b if a is known (much less likely). Recall that the method of moments estimators of a and b are

$$U = 1 + \sqrt{\frac{M^{(2)}}{M^{(2)} - M^2}}, \quad V = \frac{M^{(2)}}{M}\left(1 - \sqrt{\frac{M^{(2)} - M^2}{M^{(2)}}}\right) \tag{7.6.39}$$

respectively, where as before $M = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean and $M^{(2)} = \sum_{i=1}^n X_i^2$ the second order sample mean. These estimators are not functions of the sufficient statistics and hence suffers from loss of information. On the other hand, the maximum likelihood estimators of a and b on the interval $(0, \infty)$ are

$$W = \frac{n}{\sum_{i=1}^n \ln X_i - n \ln X_{(1)}}, \quad X_{(1)} \tag{7.6.40}$$

respectively. These are functions of the sufficient statistics, as they must be.

Run the Pareto estimation experiment 1000 times with various values of the parameters a and b and the sample size n. Compare the method of moments estimates of the parameters with the maximum likelihood estimates in terms of the empirical bias and mean square error.

The Uniform Distribution

Recall that the *continuous uniform distribution* on the interval $[a, a+h]$, where $a \in \mathbb{R}$ is the location parameter and $h \in (0, \infty)$ is the scale parameter, has probability density function g given by

$$g(x) = \frac{1}{h}, \quad x \in [a, a+h] \tag{7.6.41}$$

Continuous uniform distributions are widely used in applications to model a number chosen "at random" from an interval. Continuous uniform distributions are studied in more detail in the chapter on Special Distributions. Let's first consider the case where both parameters are unknown.

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the uniform distribution on the interval $[a, a+h]$. Then $\left(X_{(1)}, X_{(n)}\right)$ is minimally sufficient for (a, h), where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ is the first order statistic and $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$ is the last order statistic.

Proof

The PDF f of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = g(x_1)g(x_2)\cdots g(x_n) = \frac{1}{h^n}, \quad \boldsymbol{x} = (x_1, x_2, \ldots x_n) \in [a, a+h]^n \tag{7.6.42}$$

We can rewrite the PDF as

$$f(\boldsymbol{x}) = \frac{1}{h^n}\mathbf{1}[x_{(1)} \geq a]\mathbf{1}[x_{(n)} \leq a+h], \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \tag{7.6.43}$$

It then follows from the factorization theorem (3) that $\left(X_{(1)}, X_{(n)}\right)$ is sufficient for (a, h). Next, suppose that $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and that $x_{(1)} \neq y_{(1)}$ or $x_{(n)} \neq y_{(n)}$. For a given $h \in (0, \infty)$, we can easily find values of $a \in \mathbb{R}$ such that $f(\boldsymbol{x}) = 0$ and $f(\boldsymbol{y}) = 1/h^n$, and other values of $a \in \mathbb{R}$ such that $f(\boldsymbol{x}) = f(\boldsymbol{y}) = 1/h^n$. By condition (6), $\left(X_{(1)}, X_{(n)}\right)$ is minimally sufficient.

If the location parameter a is known, then the largest order statistic is sufficient for the scale parameter h. But if the scale parameter h is known, we still need both order statistics for the location parameter a. So in this case, we have a single real-valued parameter, but the minimally sufficient statistic is a pair of real-valued random variables.

Suppose again that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is a random sample from the uniform distribution on the interval $[a, a+h]$.

1. If $a \in \mathbb{R}$ is known, then $X_{(n)}$ is sufficient for h.
2. If $h \in (0, \infty)$ is known, then $\left(X_{(1)}, X_{(n)}\right)$ is minimally sufficient for a.

Proof

Both parts follow easily from the analysis given in the proof of the last theorem.

Run the uniform estimation experiment 1000 times with various values of the parameter. Compare the estimates of the parameter.

Recall that if both parameters are unknown, the method of moments estimators of a and h are $U = 2M - \sqrt{3}T$ and $V = 2\sqrt{3}T$, respectively, where $M = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean and $T^2 = \frac{1}{n}\sum_{i=1}^n (X_i - M)^2$ is the biased sample variance. If a is known, the method of moments estimator of h is $V_a = 2(M - a)$, while if h is known, the method of moments estimator of h is $U_h = M - \frac{1}{2}h$. None of these estimators are functions of the minimally sufficient statistics, and hence result in loss of information.

The Hypergeometric Model

So far, in all of our examples, the basic variables have formed a random sample from a distribution. In this subsection, our basic variables will be dependent.

Recall that in the *hypergeometric model*, we have a population of N objects, and that r of the objects are *type 1* and the remaining $N - r$ are *type 0*. The population size N is a positive integer and the type 1 size r is a nonnegative integer with $r \leq N$. Typically one or both parameters are unknown. We select a random sample of n objects, without replacement from the population, and let X_i be the type of the ith object chosen. So our basic sequence of random variables is $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. The variables are identically distributed indicator variables with $\mathbb{P}(X_i = 1) = r/N$ for $i \in \{1, 2, \ldots, n\}$, but are dependent. Of course, the sample size n is a positive integer with $n \leq N$.

The variable $Y = \sum_{i=1}^n X_i$ is the number of type 1 objects in the sample. This variable has the *hypergeometric distribution* with parameters N, r, and n, and has probability density function h given by

$$h(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}} = \binom{n}{y}\frac{r^{(y)}(N-r)^{(n-y)}}{N^{(n)}}, \quad y \in \{\max\{0, N-n+r\}, \ldots, \min\{n, r\}\} \tag{7.6.44}$$

(Recall the *falling power* notation $x^{(k)} = x(x-1)\cdots(x-k+1)$). The hypergeometric distribution is studied in more detail in the chapter on Finite Sampling Models.

Y is sufficient for (N, r). Specifically, for $y \in \{\max\{0, N-n+r\}, \ldots, \min\{n, r\}\}$, the conditional distribution of $\boldsymbol{X}$ given $Y = y$ is uniform on the set of points

$$D_y = \{(x_1, x_2, \ldots, x_n) \in \{0, 1\}^n : x_1 + x_2 + \cdots + x_n = y\} \tag{7.6.45}$$

Proof

By a simple application of the multiplication rule of combinatorics, the PDF f of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = \frac{r^{(y)}(N-r)^{(n-y)}}{N^{(n)}}, \quad \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0, 1\}^n \tag{7.6.46}$$

where $y = \sum_{i=1}^n x_i$. If $y \in \{\max\{0, N-n+r\}, \ldots, \min\{n, r\}\}$, the conditional distribution of $\boldsymbol{X}$ given $Y = y$ is concentrated on D_y and

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} \mid Y = y) = \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(Y = y)} = \frac{r^{(y)}(N-r)^{(n-y)}/N^{(n)}}{\binom{n}{y}r^{(y)}(N-r)^{(n-y)}/N^{(n)}} = \frac{1}{\binom{n}{y}}, \quad \boldsymbol{x} \in D_y \tag{7.6.47}$$

Of course, $\binom{n}{y}$ is the cardinality of D_y.

There are clearly strong similarities between the hypergeometric model and the Bernoulli trials model above. Indeed if the sampling were *with* replacement, the Bernoulli trials model with $p = r/N$ would apply rather than the hypergeometric model. It's also interesting to note that we have a single real-valued statistic that is sufficient for two real-valued parameters.

Once again, the sample mean $M = Y/n$ is equivalent to Y and hence is also sufficient for (N, r). Recall that the method of moments estimator of r with N known is NM and the method of moment estimator of N with r known is r/M. The estimator of r is the one that is used in the capture-recapture experiment.

Exponential Families

Suppose now that our data vector $\boldsymbol{X}$ takes values in a set S, and that the distribution of $\boldsymbol{X}$ depends on a parameter vector $\boldsymbol{\theta}$ taking values in a parameter space Θ. The distribution of $\boldsymbol{X}$ is a k-parameter *exponential family* if S does not depend on $\boldsymbol{\theta}$ and if the probability density function of $\boldsymbol{X}$ can be written as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \alpha(\boldsymbol{\theta}) r(\boldsymbol{x}) \exp\left(\sum_{i=1}^{k} \beta_i(\boldsymbol{\theta}) u_i(\boldsymbol{x})\right); \quad \boldsymbol{x} \in S, \ \boldsymbol{\theta} \in \Theta \tag{7.6.48}$$

where α and $(\beta_1, \beta_2, \ldots, \beta_k)$ are real-valued functions on Θ, and where r and $(u_1, u_2, \ldots, u_k)$ are real-valued functions on S. Moreover, k is assumed to be the smallest such integer. The parameter vector $\boldsymbol{\beta} = (\beta_1(\boldsymbol{\theta}), \beta_2(\boldsymbol{\theta}), \ldots, \beta_k(\boldsymbol{\theta}))$ is sometimes called the *natural parameter* of the distribution, and the random vector $\boldsymbol{U} = (u_1(\boldsymbol{X}), u_2(\boldsymbol{X}), \ldots, u_k(\boldsymbol{X}))$ is sometimes called the *natural statistic* of the distribution. Although the definition may look intimidating, exponential families are useful because they have many nice mathematical properties, and because many special parametric families are exponential families. In particular, the sampling distributions from the Bernoulli, Poisson, gamma, normal, beta, and Pareto considered above are exponential families. Exponential families of distributions are studied in more detail in the chapter on special distributions.

$\boldsymbol{U}$ is minimally sufficient for $\boldsymbol{\theta}$.

Proof

That U is sufficient for θ follows immediately from the factorization theorem. That U is minimally sufficient follows since k is the smallest integer in the exponential formulation.

It turns out that $\boldsymbol{U}$ is complete for $\boldsymbol{\theta}$ as well, although the proof is more difficult.