

1 Sufficient Statistics

When the population is modeled by a probability distribution or probability density function that depends on a parameter θ , we collect a random sample to make inferences about θ . For any particular family of distributions $f(x|\theta)$, is there is a function $T(X_1, \dots, X_n)$ of the random sample that carries all of the information concerning θ ?

To make this notion precise, we say that $T = T(X_1, \dots, X_n)$ is a **sufficient statistic** for θ if the conditional distribution, or density, of the sample, given the statistic is free of the parameter θ .

$$f(x_1, x_2, \dots, x_n | T = t; \theta) = \frac{f(x_1, x_2, \dots, x_n, t; \theta)}{g(t; \theta)}$$

Example [*A Sufficient Statistic For Bernoulli Trials*] Consider a random sample X_1, X_2, \dots, X_n , of size n , from the Bernoulli distribution with

$$P[X_1 = 0] = 1 - p = 1 - P[X_1 = 1]$$

The parameter is the population proportion p .

Show that $T = \sum_{i=1}^n X_i$ is a sufficient statistic.

Solution

By independence, the joint distribution of the random sample is

$$\prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}$$

and the joint probability of $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ and $T = \sum_{i=1}^n X_i = t$ is

$$f(x_1, x_2, \dots, x_n, t; p) = p^t (1-p)^{n-t} \quad \text{where } \sum_{i=1}^n x_i = t$$

and is zero elsewhere. Further, we know that $T = \sum_{i=1}^n X_i$ has the binomial distribution

$$g(t; p) = \binom{n}{t} p^t (1-p)^{n-t} \quad \text{for } t = 0, 1, \dots, n$$

Consequently, the conditional distribution of the sample, given $T = t$ is

$$\begin{aligned}
f(x_1, x_2, \dots, x_n | T = t; p) &= \frac{f(x_1, x_2, \dots, x_n, t; p)}{g(t; p)} \\
&= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}}
\end{aligned}$$

which does not depend on the parameter p . By definition, $T = \sum_{i=1}^n X_i$ is a sufficient statistic.

EXAMPLE Let X_1, X_2 be a random sample of size 2 from the Poisson distribution $f(x_i; \lambda) = \lambda^{x_i} e^{-\lambda} / x_i!$. Show that $T = X_1 + X_2$ is a sufficient statistic.

Solution The joint distribution of the random sample is

$$\prod_{i=1}^2 f(x_i; \lambda) = \frac{\lambda^{x_1+x_2}}{x_1! x_2!} e^{-2\lambda}$$

and the joint probability of $X_1 = x_1, X_2 = x_2$ and $T = X_1 + X_2 = t$ is

$$f(x_1, x_2, t; \lambda) = \frac{\lambda^t}{x_1! x_2!} e^{-2\lambda} \quad \text{where } x_1 + x_2 = t$$

and is zero elsewhere. Further, we know that the sum $T = X_1 + X_2$ has the Poisson distribution with parameter 2λ .

$$g(t; \lambda) = \frac{(2\lambda)^t}{t!} e^{-2\lambda}$$

Consequently, the conditional distribution of the sample, given $T = t$ is

$$\begin{aligned}
f(x_1, x_2 | T = t; \lambda) &= \frac{f(x_1, x_2, t; \lambda)}{g(t; \lambda)} \\
&= \frac{\lambda^t}{x_1! x_2!} e^{-2\lambda} / \frac{(2\lambda)^t}{t!} e^{-2\lambda} \\
&= \binom{t}{x_1} \frac{1}{2^t}
\end{aligned}$$

which does not depend on the parameter λ . By definition, $T = X_1 + X_2$ is a sufficient statistic.

An Interpretation of Sufficiency In what sense does a sufficient statistic $T(X_1, \dots, X_n)$ carry all of the sample information concerning the parameter? Consider the following two stage procedure for obtaining a random sample from $f(x; \theta)$

Step 1. Observe the sufficient statistic T which is distributed as $g(t; \theta)$. Note that the distribution of T depends on the value of θ that prevails.

Step 2. Given the value t observed in step 1, generate (X_1, \dots, X_n) from the conditional distribution $f(x_1, x_2, \dots, x_n | T = t)$, which by the definition is free of theta. That is, at this second step, it is not necessary to know the value of θ .

In the Poisson example, suppose we observe $T = X_1 + X_2 = 5$. The five events are assigned to the two components of the random sample (X_1, X_2) according to the binomial 'fair coin' model for five flips. The sample $(x_1, 5 - x_1)$ is generated when x_1 is the value of the binomial variable where

$$f(x_1, x_2 | T = t) = \binom{5}{x_1} \frac{1}{2^5}$$

This last experiment could be conducted by flipping a fair coin five times and letting x_1 be the number of heads. Clearly $T = X_1 + X_2$ has the information on λ because the second step is conducted with random numbers unrelated to the parameter.

Fortunately, there is an easier way to obtain sufficient statistics. J. Neyman and R. A. Fisher produced equivalent conditions for a sufficient statistic to exist. It connects sufficiency to a factorization of the joint distribution or density, $f(x_1, x_2, \dots, x_n; \theta)$, of X_1, X_2, \dots, X_n

Neyman-Fisher Factorization Criteria A statistic $T(X_1, \dots, X_n)$ is sufficient for θ if and only if the joint distribution or density can be factored as

$$f(x_1, x_2, \dots, x_n; \theta) = g(t, \theta)h(x_1, \dots, x_n)$$

where the function $g(t, \theta)$ only depends on $t = t(x_1, x_2, \dots, x_n)$ and θ while $h(x_1, x_2, \dots, x_n)$ does not depend on θ .

We illustrate the use of the factorization criterion with the following examples.

Example Let X have the binomial distribution.

Show that X is a sufficient statistic

Solution The distribution of X

$$\binom{n}{x} p^x (1 - p)^{n-x}$$

factors with $h(x) = 1$.

Example Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with known variance σ_0^2 .

Obtain a sufficient statistic.

Solution The joint density function is

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \sigma_0} e^{-(x_i - \mu)^2 / 2\sigma_0^2} \\ &= \left(\frac{1}{\sqrt{2\pi} \sigma_0} \right)^n e^{(-n\mu^2 + 2\mu \sum_{i=1}^n x_i) / 2\sigma_0^2} \times e^{-(\sum_{i=1}^n x_i^2) / 2\sigma_0^2} \end{aligned}$$

Taking the first term to be the function of t and μ , $g(t; \mu)$, and $h(x_1, x_2, \dots, x_n) = e^{-(\sum_{i=1}^n x_i^2) / 2\sigma_0^2}$ we obtain a factorization that shows that $T = \sum_{i=1}^n X_i$ is sufficient.

Example Let X_1, X_2, \dots, X_n be a random sample from an exponential family with probability distribution or density

$$f(x_1; \theta) = e^{-A(\theta) + B(\theta)t(x_1)} k(x_1)$$

Show that $T = \sum_{i=1}^n t(X_i)$ is a sufficient statistic

Solution

The joint distribution

$$f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n e^{-A(\theta) + B(\theta)t(x_i)} k(x_i) = e^{-nA(\theta) + B(\theta) \sum_{i=1}^n t(x_i)} \prod_{i=1}^n k(x_i)$$

factors with $h(x_1, x_2, \dots, x_n) = \prod_{i=1}^n k(x_i)$ and the other term is $g(t; \theta)$.

This last example applies to the normal distributions with known variance, the normal distributions with known mean and the Poisson distributions, among others. Somewhat surprisingly, there is a one dimensional sufficient statistic whatever the sample size.

The factorization criteria even applies when there is more than one parameter.

Example Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with mean and variance both unknown.

Obtain the sufficient statistics.

Solution

We first write

$$(x_i - \mu)^2 = (x_i - \bar{x} + \bar{x} - \mu)^2 = (x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu)$$

Summing, we see that the last term vanishes since $\sum_{i=1}^n (x_i - \bar{x}) = 0$, so

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

Then, the joint probability density

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi} \sigma} e^{-(x_i - \mu)^2 / 2\sigma^2} \\ &= \left(\frac{1}{\sqrt{2\pi} \sigma} \right)^n e^{-(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2) / 2\sigma^2} \end{aligned}$$

which is a function only of \bar{x} and $\sum_{i=1}^n (x_i - \bar{x})^2$. Together, \bar{x} and $\sum_{i=1}^n (x_i - \bar{x})^2$ are sufficient for μ and σ^2 .

Note that, in the last example, we could also factor the joint density as a function of \bar{x} and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$. In fact any one-to-one transformation of a sufficient statistic is still sufficient.