1 Sufficient statistics

A *statistic* is a function $T = r(X_1, X_2, \cdots, X_n)$ of the random sample $X_1, X_2, \cdots, X_n$. Examples are

$$\begin{aligned} \bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{(the sample mean)} \\ s^2 = &= \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \quad \text{(the sample variance)} \\ T_1 &= \max\{X_1, X_2, \cdots, X_n\} \\ T_2 &= 5 \end{aligned} \tag{1}$$

The last statistic is a bit strange (it completely igonores the random sample), but it is still a statistic. We say a statistic T is an estimator of a population parameter if T is usually close to θ. The sample mean is an estimator for the population mean; the sample variance is an estimator for the population variation.

Obviously, there are lots of functions of $X_1, X_2, \cdots, X_n$ and so lots of statistics. When we look for a good estimator, do we really need to consider all of them, or is there a much smaller set of statistics we could consider? Another way to ask the question is if there are a few key functions of the random sample which will by themselves contain all the information the sample does. For example, suppose we know the sample mean and the sample variance. Does the random sample contain any more information about the population than this? We should emphasize that we are always assuming our population is described by a given family of distributions (normal, binomial, gamma or ...) with one or several unknown parameters. The answer to the above question will depend on what family of distributions we assume describes the population.

We start with a heuristic definition of a *sufficient statistic*. We say T is a sufficient statistic if the statistician who knows the value of T can do just as good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample.

The mathematical definition is as follows. A statistic $T = r(X_1, X_2, \cdots, X_n)$ is a sufficient statistic if for each t, the conditional distribution of $X_1, X_2, \cdots, X_n$ given $T = t$ and θ does not depend on θ.

To motivate the mathematical definition, we consider the following "experiment." Let $T = r(X_1, \cdots, X_n)$ be a sufficient statistic. There are two statisticians; we will call them A and B. Statistician A knows the entire random sample $X_1, \cdots, X_n$, but statistician B only knows the value of T, call it t. Since the conditional distribution of $X_1, \cdots, X_n$ given θ and T does not depend on θ, statistician B knows this conditional distribution. So he can use his computer to generate a random sample $X_1', \cdots, X_n'$ which has this conditional distribution. But then his random sample has the same distribution as a random sample drawn from the population (with its unknown value of θ). So statistician B can use his random sample $X_1', \cdots, X_n'$ to compute whatever statistician A computes using his random sample $X_1, \cdots, X_n$, and he will (on average) do as well as statistician A. Thus the mathematical definition of sufficient statistic implies the heuristic definition.

It is difficult to use the definition to check if a statistic is sufficient or to find a sufficient statistic. Luckily, there is a theorem that makes it easy to find sufficient statistics.

Theorem 1. *(Factorization theorem) Let $X_1, X_2, \cdots, X_n$ be a random sample with joint density $f(x_1, x_2, \cdots, x_n|\theta)$. A statistic $T = r(X_1, X_2, \cdots, X_n)$ is sufficient if and only if the joint density can be factored as follows:*

$$f(x_1, x_2, \cdots, x_n|\theta) = u(x_1, x_2, \cdots, x_n)\, v(r(x_1, x_2, \cdots, x_n), \theta) \tag{2}$$

where u and v are non-negative functions. The function u can depend on the full random sample $x_1, \cdots, x_n$, but not on the unknown parameter θ. The function v can depend on θ, but can depend on the random sample only through the value of $r(x_1, \cdots, x_n)$.

It is easy to see that if $f(t)$ is a one to one function and T is a sufficient statistic, then $f(T)$ is a sufficient statistic. In particular we can multiply a sufficient statistic by a nonzero constant and get another sufficient statistic.

We now apply the theorem to some examples.

Example *(normal population, unknown mean, known variance)* We consider a normal population for which the mean μ is unknown, but the the variance σ^2 is known. The joint density is

$$f(x_1, \cdots, x_n|\mu) = (2\pi)^{-n/2}\, \sigma^{-n} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$= (2\pi)^{-n/2}\,\sigma^{-n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n\mu^2}{2\sigma^2}\right)$$

Since σ^2 is known, we can let

$$u(x_1,\cdots,x_n) = (2\pi)^{-n/2}\,\sigma^{-n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)$$

and

$$v(r(x_1,x_2,\cdots,x_n),\mu) = \exp\left(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}r(x_1,x_2,\cdots,x_n)\right)$$

where

$$r(x_1,x_2,\cdots,x_n) = \sum_{i=1}^{n} x_i$$

By the factorization theorem this shows that $\sum_{i=1}^{n} X_i$ is a sufficient statistic. It follows that the sample mean $\bar{X}_n$ is also a sufficient statistic.

Example *(Uniform population)* Now suppose the X_i are uniformly distributed on $[0,\theta]$ where θ is unknown. Then the joint density is

$$f(x_1,\cdots,x_n|\theta) = \theta^{-n}\,1(x_i \le \theta,\, i = 1,2,\cdots,n)$$

Here $1(E)$ is an indicator function. It is 1 if the event E holds, 0 if it does not. Now $x_i \le \theta$ for $i = 1,2,\cdots,n$ if and only if $\max\{x_1,x_2,\cdots,x_n\} \le \theta$. So we have

$$f(x_1,\cdots,x_n|\theta) = \theta^{-n}\,1(\max\{x_1,x_2,\cdots,x_n\} \le \theta)$$

By the factorization theorem this shows that

$$T = \max\{X_1,X_2,\cdots,X_n\}$$

is a sufficient statistic.

What about the sample mean? Is it a sufficient statistic in this example? By the factorization theorem it is a sufficient statistic only if we can write $1(\max\{x_1,x_2,\cdots,x_n\} \le \theta)$ as a function of just the sample mean and θ. This is impossible, so the sample mean is not a sufficient statistic in this setting.

Example *(Gamma population, α unknown, β known)* Now suppose the population has a gamma distribution and we know β but α is unknown. Then the joint density is

$$f(x_1, \cdots, x_n|\alpha) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i^{\alpha-1} \right) \exp(-\beta \sum_{i=1}^{n} x_i)$$

We can write

$$\prod_{i=1}^{n} x_i^{\alpha-1} = \exp\left((\alpha - 1) \sum_{i=1}^{n} \ln(x_i) \right)$$

By the factorization theorem this shows that

$$T = \sum_{i=1}^{n} \ln(X_i)$$

is a sufficient statistic. Note that $\exp(T) = \prod_{i=1}^{n} X_i$ is also a sufficient statistic. But the sample mean is not a sufficient statistic.

Now we reconsider the example of a normal population, but suppose that both μ and σ^2 are both unknown. Then the sample mean is not a sufficient statistic. In this case we need to use more than one statistic to get sufficiency. The definition (both heuristic and mathematical) of sufficiency extends to several statistics in a natural way.

We consider k statistics

$$T_i = r_i(X_1, X_2, \cdots, X_n), \quad i = 1, 2, \cdots, k \tag{3}$$

We say $T_1, T_2, \cdots, T_k$ are jointly sufficient statistics if the statistician who knows the values of $T_1, T_2, \cdots, T_k$ can do just as good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample. In this setting θ typically represents several parameters and the number of statistics, k, is equal to the number of unknown parameters.

The mathematical definition is as follows. The statistics $T_1, T_2, \cdots, T_k$ are jointly sufficient if for each $t_1, t_2, \cdots, t_k$, the conditional distribution of $X_1, X_2, \cdots, X_n$ given $T_i = t_i$ for $i = 1, 2, \cdots, n$ and θ does not depend on θ.

Again, we don't have to work with this definition because we have the following theorem:

Theorem 2. *(Factorization theorem) Let $X_1, X_2, \cdots, X_n$ be a random sample with joint density $f(x_1, x_2, \cdots, x_n|\theta)$. The statistics*

$$T_i = r_i(X_1, X_2, \cdots, X_n), \quad i = 1, 2, \cdots, k \tag{4}$$

are jointly sufficient if and only if the joint density can be factored as follows:

$$\begin{aligned} f(x_1, x_2, \cdots, x_n|\theta) &= u(x_1, x_2, \cdots, x_n) \\ & \quad v(r_1(x_1, \cdots, x_n), r_2(x_1, \cdots, x_n), \cdots, r_k(x_1, \cdots, x_n), \theta) \end{aligned}$$

where u and v are non-negative functions. The function u can depend on the full random sample $x_1, \cdots, x_n$, but not on the unknown parameters θ. The function v can depend on θ, but can depend on the random sample only through the values of $r_i(x_1, \cdots, x_n)$, $i = 1, 2, \cdots, k$.

Recall that for a single statistic we can apply a one to one function to the statistic and get another sufficient statistic. This generalizes as follows. Let $g(t_1, t_2, \cdots, t_k)$ be a function whose values are in $\mathbb{R}^k$ and which is one to one. Let $g_i(t_1, \cdots, t_k)$, $i = 1, 2, \cdots, k$ be the component functions of g. Then if $T_1, T_2, \cdots, T_k$ are jointly sufficient, then $g_i(T_1, T_2, \cdots, T_k)$ are jointly sufficient.

We now revisit two examples. First consider a normal population with unknown mean and variance. Our previous equations show that

$$T_1 = \sum_{i=1}^{n} X_i, \quad T_2 = \sum_{i=1}^{n} X_i^2$$

are jointly sufficient statistics. Another set of jointly sufficent statistics is the sample mean and sample variance. (What is $g(t_1, t_2)$?)

Now consider a population with the gamma distribution with both α and β unknown. Then

$$T_1 = \sum_{i=1}^{n} X_i, \quad T_2 = \sum_{i=1}^{n} \ln(X_i)$$

are jointly sufficient statistics.

2 Exercises

1. Gamma, both parameters unknown, show sum and product form a sufficient

2. Uniform on $[\theta_1, \theta_2]$. Find two statistics that are sufficient.

3. Poisson or geometric - find sufficient statistic.

4. Beta ?