3 Sufficient Statistics

1.Sufficiency Principle An experimenter uses the information in a sample $X_1, X_2, \ldots, X_n$ to make inferences about an unknown parameter θ. If n is large, then the observed sample $x_1, x_2, \ldots, x_n$ is a long list of numbers that may be hard to interpret. An experimenter might wish to summarize the information in a sample by determining a few key features of the sample values. This is usually done by computing statistics, functions of the sample. Any statistic, $T(\mathbf{X})$, defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\mathbf{x})$, rather the entire observed sample, $\mathbf{x}$, will treat as equal two samples, $\mathbf{x}$ and $\mathbf{y}$, that satisfy $T(\mathbf{x}) = T(\mathbf{y})$ even though actual sample values may be different in some ways.

A *sufficient statistic* of a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in a sample. Any additional information in a sample besides the value of the sufficient statistic, does not contain any more information about θ. These considerations lead to the data reduction technique known as the Sufficiency principle.

Sufficiency Principle: If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points, such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

2. Example. In the earlier courses in statistics we relied on our intuition to define the statistics we used for estimation and hypothesis testing. For example, in a sequence of Bernoulli trials in each of which there is a probability θ of success we used $\sum_{i=1}^{n} x_i$ as the base for our estimate of θ and in our hypothesis tests about θ. [All x_i are 0 or 1.] Other information, such as the order in which the 0s and 1s appear was not considered relevant and the statistic $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ (number of X_is that equal 1), was *sufficient* for our purposes. We now give this notion a general mathematical formulation.

Suppose that we are given a value of $T(\mathbf{X}) = \sum_{i=1}^{n} X_i = t$. If we know the value t, can we gain any further information about θ by looking at other functions of $X_1, X_2, \ldots, X_n$?

One way to answer this question is by looking at the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T(\mathbf{X}) = t$:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T(\mathbf{X}) = t) = \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}.$$

Thus, the conditional probability is independent of θ, that is , once $\sum_{i=1}^{n} X_i$ is known, no other function of $X_1, X_2, \ldots, X_n$ will shed additional light on the possible value of θ.

3. Definition. A statistic $T(\mathbf{X})$ is a **sufficient statistic** for θ if the conditional distribution of the sample $X_1, X_2, \ldots, X_n$ given the value of $T(\mathbf{X})$ does not depend on θ.

4. Factorization criterion. Let $f(\mathbf{x}, \theta)$ denote the joint p.d.f/p.f. of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for the parameter θ if and only if we can write $f(\mathbf{x},\theta) = h(T(\mathbf{x}), \theta)c(\mathbf{x})$ for appropriate functions h and c of the indicated variables - that is, c does not depend on θ. [Recall that this includes the statement that the set of values $\mathbf{x}$ where $c(\mathbf{x}) \neq 0$ does not depend on θ.]

5. Examples. Let $X_1, X_2..., X_n$ be i.i.d. (a) N(μ, σ^2) r.v.s, where σ is known; (b) M(θ); (c) $U(0, \theta)$.

6. Note. In all the previous examples, the sufficient statistic is a real valued function of the sample. All the information about θ in a sample $\mathbf{x}$ is summarised in the single number $T(\mathbf{x})$. Sometimes the information cannot be summarised in a single number and several numbers are required instead. In such cases a sufficient statistic is a vector, say $T(\mathbf{x}) = (T_1(\mathbf{x}), \dots, T_r(\mathbf{x}))$. This situation often occurs when the parameter is also a vector, say, $\theta = (\theta_1, ..., \theta_k)$, and it is usually the case that the sufficient statistic and the vector of parameters are of equal length, that is $r = k$.

7. Example. Let $X_1, X_2..., X_n$ be i.i.d. N(μ, σ^2) r.v.s, where both μ and σ^2 are unknown.

8. Proposition Let $X_1, X_2..., X_n$ be i.i.d. observations from a p.d.f./p.f. $f(x, \theta)$ that belongs to the exponential family, i.e.,

$$f(x, \theta) = \exp\{\textstyle\sum_{i=1}^m p_i(\theta) K_i(x) + S(x) + q(\theta)\}$$

for all $\{x : f(x, \theta) \neq 0\}$, where $\theta = (\theta_1, ..., \theta_k)$ and the set $\{x : f(x, \theta) \neq 0\}$ does not depend on θ. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^n K_1(X_j), \sum_{j=1}^n K_2(X_j), \dots, \sum_{j=1}^n K_m(X_j) \right)$$

is a sufficient statistic for θ.

9. Example. Let $X_1, X_2..., X_n$ be i.i.d. N(μ, σ^2) r.v.s, where both μ and σ^2 are unknown.

10. Remarks. (i) In the examples above we found one sufficient statistic for each model considered. In any problem, there are many sufficient statistics. It is always true that the complete sample, $\mathbf{X}$, is a sufficient statistic.

(ii) It follows that any one-to-one function of a sufficient statistic is a sufficient statistic.

(iii) Because of the numerous sufficient statistics in a model, we might ask whether one sufficient statistic is any better than the another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still retaining all the information about θ might be considered as preferable. The definition of such statistic is formalized below.

11. Definition. A sufficient statistic $T(\mathbf{X})$ is called a **minimal sufficient statistic**, if for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.

12. Example. (*Two normal sufficient statistics*) In Example 5(a) with N(μ,σ^2) r.v.s and σ^2 known, we concluded that $T(\mathbf{X}) = \sum X_i$ is a sufficient statistic for μ. Instead we could write down factorisation from Example 7 for this problem (σ^2 is a known value now) and correctly conclude that $T'(\mathbf{X}) = (\sum X_i, \sum X_i^2)$ is a sufficient statistic for μ in this problem. Clearly, $T(\mathbf{X})$ achieves a greater data reduction than $T'(\mathbf{X})$; we can write $T(\mathbf{X})$ as a function of $T'(\mathbf{X})$. (Indeed, define $v(a, b) = a$, then $T(\mathbf{X}) = \sum X_i = v(\sum X_i, \sum X_i^2) = v(T'(\mathbf{X}))$.) Since $T(\mathbf{X})$ and $T'(\mathbf{X})$ are both sufficient statistics, they both contain the same information about μ. Thus the additional information about the value of $\sum X_i^2$ does not add to our knowledge of μ since σ^2 is known. Of course, if σ^2 is unknown, as in Example 7, $T(\mathbf{X}) = \sum X_i$ is not a sufficient statistic and $T'(\mathbf{X})$ contains more information about (μ,σ^2) than does $T(\mathbf{X})$.

13. Theorem. (*Lehmann-Scheffé criterion*) Let $f(\mathbf{x}, \theta)$ be the joint p.d.f/p.f. of a sample $\mathbf{X}$. Suppose that there exists a function $T(\mathbf{x})$ such that, for any two sample points $\mathbf{x}$ and

$\mathbf{y}$, the ratio $f(\mathbf{x},\theta)/f(\mathbf{y},\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ.

14. Example. Let $X_1, X_2..., X_n$ be i.i.d. (a) $B(1,\theta)$, (b)$\Gamma(\alpha,\beta)$.

15. Theorem. If T is a sufficient statistic for the parameter θ and $\widehat{\theta}$ is a maximum likelihood estimator of θ, then we may write $\widehat{\theta} = \widehat{\theta}(T)$.

Proof. By the factorization, we have$f(\mathbf{x},\theta) = h(T(\mathbf{x}),\theta)c(\mathbf{x})$ where $c(\mathbf{x})$ is independent of θ. Thus, maximizing $f(\mathbf{x},\theta)$ over θ for fixed $\mathbf{x}$ (and, therefore fixed $T(\mathbf{x})$) is equivalent to maximizing $h(T(\mathbf{x}),\theta)$ and the solution $\widehat{\theta}$ will be a function of T. The corresponding estimator $\widehat{\theta}$ is thus a function of T.