Sufficiency

Sufficiency is an important concept in theoretical statistics but it also has practical and computational implications. It formalizes the intuition that certain summaries of the data contain all the information about the parameters in the statistical model that we are considering. Once we have sufficient statistics we can use them to improve the estimates, find optimal tests, find better algorithms for maximum likelihood estimation, or simply have a better intuition about the models that we are using.

Two examples

To start we will look at two examples to motivate the definitions.

Example 1: When testing roulette wheels we usually assume that the subsequent spins are independent but not necessarily that the probabilities of outcomes are equal. The simplest statistics to use is the χ^2 defined by

$$\chi^2 = \sum_{i=0}^{36} \frac{(O_i - E_i)^2}{E_i}$$

where O_i stands for the observed occurencies of the outcome i and E_i for the expected occurencies for $i = 1, 2, \ldots, 36$. If we have a long series of outcomes we can "monitor the progress" of the χ^2 statistics through time, or look at segments that look "suspicious". If we simulate $n = 370.000$ outcomes on an ideal wheel and look for the segment of length $m = 37.000$ which produces the highest χ^2 statistics we get the following results after 200.000 repetitions:

Critical value	% exceeding for all data	% exeeding, worst segment
58,62	1,03	99,57
67,99	0,11	73,75
76,36	0,011	25,42

Table 1 Simulated percentages of rejection of H_0.

If we have the above assumptions of independence and constant probabilities through time we will reject the correct $H_0 \colon p_0 = p_1 = \cdots = p_{36}$ with probability almost 1 at the 1% significance level. This is an instance of *data snooping* i.e. looking for patterns in the data and basing judgement of these patterns. The simple example above shows that this is clearly wrong.

Reference:

https://en.wikipedia.org/wiki/Data_dredging

How can we justify that looking at selected segments of the data is not the right thing to look at. Denote by $N_0(n), N_1(n), \ldots, N_{36}(n)$ the random number of occurencies of individual outcomes of the roulette wheel after n spins. Denote by $X_1, X_2, \ldots, X_n$ the outcomes themselves which means that $X_1, X_2, \ldots, X_n$ are independent random variables uniformly distributed on the set $\{0, 1, \ldots, 36\}$. Probability gives us that

$$P\left(X_1 = x_1, \ldots, X_n = x_n | N_0(n) = n_0, \ldots, N_{36}(n) = n_{36}\right) = \left(\frac{n!}{n_0! \cdots n_{36}!}\right)^{-1}$$

whatever the probabilities of individual outcomes. This means that if we know the counters of outcomes there is no "residual information" about the parameters in knowing the individual outcomes. This means that the counters capture all the information there is about the parameters in the data. The mathematical way to say that is that the conditional distribution of the data given a set of statistics does not depend on the parameters. This is an instance of a set of sufficent statistics.

Example 2: One of the psychometric models is the Rasch model. The data are vectors of 0 and 1 indicating the correct response by subject i to question j. If we denote by X_{ij} the indicatorof the response of subject i to question j the Rasch model specifies that

$$P\left(X_{ij} = x_{ij}, i \leq m, j \leq n\right) = \prod_{i,j} \frac{e^{(\alpha_i - \delta_j)x_{ij}}}{1 + e^{(\alpha_i - \delta_j)}}$$

for parameters $\alpha = (\alpha_1, \ldots, \alpha_m)$ and $\delta = (\delta_1, \ldots, \delta_n)$. The parameters are interpreted as abilities of subjects and the difficulty of the problems. Given the data we need to estimate the parameters. But what quantities capture the information about the parameters? Denote by

$$X_{i\cdot} = \sum_j X_{ij} \quad \text{and} \quad X_{\cdot j} = \sum_i X_{ij}$$

the row or column sums in the data matrix. Some elementary mathematics gives that

$$P\left(X_{ij} = x_{ij}, i \leq m, j \leq n | X_{i\cdot} = x_{i\cdot}, X_{\cdot j} = x_{\cdot j} i \leq m, j \leq n\right) = \frac{1}{M}$$

where M is the total number of possible data matrices with given row and column sums. Again we see that the parameters do not appear in the conditional distribution. The row and column sums have captured all the information there is about the parameters in the data. All estimation procedures and hypothesis tests should be functions of these sufficient statistics only. This example has and additional feature that we should point out. We are mostly interested in the estimation of abilities α_i not so much the levels of difficulty δ_j. We can compute the conditional probabilities

$$P\left(X_{ij} = x_{ij}, i \leq m, j \leq n | X_{.j} = x_{.j}, j \leq n\right) = \frac{\prod_i e^{\alpha_i x_{i.}}}{\sum \prod_i e^{\alpha_i u_{i.}}} .$$

The sum runs over all possible data matrices with prescribed column sums. The conditional distributions does not contain the parameters δ_j and can be used as conditional likelihood function to estimate teh α_i. We can say that the column sums are sufficient for part of the (nuisance) parameters. Estimation based on conditional likelihood in this case has advantages like aymptotic normality and consistency.

Reference:

`https://en.wikipedia.org/wiki/Rasch_model`

Definitions and factorisation theorem

In the two examples we have seen the importance of finding sufficient statistics in practical situations of judgement and estimation. We need a precise mathematical definition of the concept of sufrficiency and an easy way to judge whether a set of statistics is sufficient for the parameters. The setup will be that we will assume that the data is a sample from a distribution of a vector or matrix $\mathbf{X}$ and the distribution of $\mathbf{X}$ is from a parametric family indexed by the parameter $\boldsymbol{\theta} \in \Theta$. Let $\mathbf{T}(\mathbf{X})$ be a vector of statistics i.e. functions of $\mathbf{X}$. For each $\boldsymbol{\theta}$ and each bounded function f we can compute the conditional expectation

$$E_{\boldsymbol{\theta}}\left(f(\mathbf{X})|\mathbf{T}(\mathbf{X})\right) = \psi_{\boldsymbol{\theta}}\left(\mathbf{X}\right)$$

In general this conditional expectation will depend on $\boldsymbol{\theta}$. If that is not the case, however, we can claim that the conditional distribution does not depend on the parameter $\boldsymbol{\theta}$.

Definition: If for every bounded function f the conditional expectation

$$E_{\boldsymbol{\theta}}\left(f(\mathbf{X})|\mathbf{T}(\mathbf{X})\right)$$

is a function of $\mathbf{T}$ only then $\mathbf{T}$ is a sufficient statistic for the parameter $\boldsymbol{\theta}$.

REMARK: Sufficiency means that $\mathbf{T}$ captures all the information about the parameter $\boldsymbol{\theta}$ contained in the data. The conditional distribution may depend on part of the parameters. Then $\mathbf{T}$ is sufficient for those parameters that do not appear in the conditional distribution.

In the examples we found sufficient statistics explicitely. But how does one find sufficient statistics easily? The answer is given by the factorisation theorem. The theorem is valid in great generality but here we will only treat the case when the distributions involved have a density or a probability function. The problem is treated in utmost generality in P. Billingsley, Probability and Measure, John Wiley and Sons, 1979, p. 400.

THEOREM: Suppose we have a family of probability functions or densities of the form $\{p(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Suppose further that T is a function (possibly vector valued) defined for all $\mathbf{x}$. The statistic $\mathbf{T}(\mathbf{X})$ is sufficient if and only if the probability function or the density can be factorised as

$$p(\mathbf{x}, \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x})$$

where g and h are functions and h does not depend on $\boldsymbol{\theta}$.

Proof: First assume that $\mathbf{X}$ is discrete. There are countably many points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ such that

$$P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}_i) > 0$$

and $\sum_i P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}_i) = 1$. Suppose the probability function can be factorized as above. Suppose $P_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{t}) > 0$. By definition of conditional probabilities

$$\begin{aligned} & P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} | \mathbf{T}(\mathbf{X}) = \mathbf{t}) \\ = & \frac{P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}, \mathbf{T}(\mathbf{X}) = \mathbf{t})}{P_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{t})} \\ = & \frac{P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x})}{P_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))} \end{aligned}$$

Observe that if $\mathbf{T}(\mathbf{x}) = \mathbf{t}$ then

$$P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}, \mathbf{T}(\mathbf{X}) = \mathbf{t}) = g(\mathbf{t}, \boldsymbol{\theta}) h(\mathbf{x})$$

and

$$P_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{t}) = \sum_{\{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}\}} g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) \,.$$

So

$$P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} | \mathbf{T}(\mathbf{X}) = \mathbf{t}) = \frac{h(\mathbf{x})}{\sum_{\{\mathbf{x}:\mathbf{T}(\mathbf{x})=\mathbf{t}\}} h(\mathbf{x})} \,.$$

This proves that $\mathbf{T}(\mathbf{X})$ is sufficient because the right side does not depend on $\boldsymbol{\theta}$.

Assume now that $\mathbf{T}(\mathbf{X})$ is sufficient. By the law of total probabilities

$$P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{t}} P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x}, \mathbf{T}(\mathbf{X}) = \mathbf{t}) \,.$$

Rewrite to get

$$P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x} | \mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) P_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x})) \,.$$

By definition the conditional probability in the last line above only depends on $\mathbf{x}$ but not on $\boldsymbol{\theta}$ so we can take it to be the function h. The probability $P_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) = \mathbf{T}(\mathbf{x}))$ depends on $\mathbf{x}$ only through $\mathbf{T}(\mathbf{x}))$ and is therefore of the form $g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})$ for some g.

The proof in the continuous case is harder and depends on calculations with conditional expectations.

The most obvious example where sufficient statistics can be found are exponential families of distributions.

DEFINITION: An exponential family of distributions is given by either probability functions or densities of the form

$$p(\mathbf{x}, \boldsymbol{\theta}) = \exp\left(\sum_{k=1}^{r} c_k(\boldsymbol{\theta}) T_k(\mathbf{x}) \right) h(\mathbf{x})$$

for some functions $c_1, \ldots, c_r$.

From the factorization theorem it follows immediately that $\mathbf{T} = (T_1, \ldots, T_r)$ is a sufficient statistic for the parameter ***theta***. All the usual families (normal, gamma, Poisson) are exponential distributions.

EXAMPLE 3: If $X_1, \ldots, X_n$ are independent normal then we have that for $\mathbf{X} = (X_1, \ldots, X_n)$ that

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{k=1}^{n} \left[(x_k - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right) .$$

It follows that the pair

$$\left(\sum_{k=1}^{(} X_k - \bar{X})^2, \bar{X} \right)$$

is a set of sufficient statistics for the parameters μ and σ. This is not so easy to see directly.