STA3000 Sufficiency and Ancillarity

Sufficiency

Definition:
A statistic $S = s(Y)$ is sufficient for θ, in the family of models $f(y;\theta); \theta \in \Theta$, if and only if $f(y|s)$ is free of θ.

Factorization Theorem: A statistic S is sufficient for θ (in the family of distributions $\{f(y;\theta), \theta \in \Theta\}$), if and only if there exist functions g and h such that

$$f(y;\theta) = g\{s(y);\theta\}h(y) \qquad \forall \quad \theta.$$

Interpretation: The usual interpretation of sufficiency (see, e.g. CH Ch. 2) is that a statistician observing Y should come to the same conclusion about θ as a statistician observing only S, because the second statistician could recover the joint density for Y from the marginal density for S by multiplying by a fixed density (that of Y given S). That is, an observation from the conditional density of Y given S can be generated without any knowledge of θ. Thus it could be generated by a random number generator. It is argued (and generally accepted) that this additional randomness cannot provide any information about the unknown parameter θ. This discussion assumes that the model is correct, of course. In fact, we shall see that the conditional density of Y given S can be used for checking the model.

Minimal sufficiency: A sufficient statistic partitions the sample space $\mathcal{Y}$ into subspaces. Two elements of $\mathcal{Y}$ are in the same partition if they have the same value of $s(y)$. Typically the 'size' of the partition is substantially smaller that the 'size' of the sample space. For example, if Y is a random variable on $\mathbb{R}^n$, $S(Y) = \sum Y_i$ takes values on $\mathbb{R}$. The biggest reduction in dimension is obtained with the coarsest partition. The statistic inducing the coarsest possible partition (if it exists) is called the minimal sufficient statistic.
Definition. A statistic $S = s(Y)$ is a minimal sufficient statistic if it is a function of every other sufficient statistic.

This definition is obviously difficult to work with, but can be avoided by using the following theorem:
Theorem. The likelihood statistic is minimal sufficient.
Proof. The likelihood statistic is the partition of the sample space that puts two values of y into the same partition if they have the same likelihood function (i.e. their likelihood functions are in the same equivalence class). Thus y and y' are in the same partition if

$$\begin{aligned} L(\theta;y) &= L(\theta;y') \\ c(y)f(y;\theta) &= c(y')f(y';\theta) \\ f(y;\theta) &= k(y',y)f(y';\theta). \end{aligned}$$

If this holds, we write $s(y) = s(y')$ to indicate that $s(\cdot)$ is the function that indexes this partition.

Suppose $T = t(Y)$ is some other sufficient statistic. By the factorization theorem

$$\begin{aligned} f(y;\theta) &= g\{t(y);\theta\}h(y) \\ f(y';\theta) &= g\{t(y');\theta\}h(y') \end{aligned}$$

If $t(y) = t(y')$ then

$$\begin{aligned} f(y;\theta) &= g\{t(y');\theta\}h(y) \\ &= f(y';\theta)\{h(y)/h(y')\} \\ &= f(y';\theta)k(y,y') \end{aligned}$$

so $s(y) = s(y')$, which shows that $s(\cdot)$ is a function of $t(\cdot)$.

References

The sketch of the proof given here is taken from CH, Ch.2. There is a good discussion of sufficiency as well in SM, §4.2.

The "likelihood statistic" is a confusing term, as normally functions of y and θ are not statistics. The term "likelihood map" is used in Fraser & Naderi (2006); this paper (available as `238.pdf` at `http://www.utstat.toronto.edu/dfraser/`) is the latest is a series of studies of abstract notions of likelihood and minimal sufficiency.

The result is also proved in Lehmann and Romano (TSH), §1.9 and 2.6 or in Lehmann and Casella (TPE), Theorem 1.6.12 and Corollary 1.6.13, p. 37.

Ancillarity

Definition:

A statistic $A = a(Y)$ is ancillary for θ, in the family of models $f(y;\theta); \theta \in \Theta$, if and only if $f(a)$ is free of θ.

A key feature of transformation models is that the model for a sample of size n permits a reduction in dimension of the sufficient statistic. This reduction is obtained by conditioning. Thus, there exist functions of the data, say $s(Y)$ and $a(Y)$, for which we can write

$$f(y;\theta) \propto f_1\{s(y)|a(y);\theta\}f_2\{a(y)\}$$

where the marginal density of $a(Y)$ does not depend on the parameter θ. More importantly, the conditional density f_1 is itself a transformation family density, with parameter θ and sample space variable $s(Y)$.

Location family Suppose $Y_1, \ldots, Y_n$ are i.i.d. observations from the location family $f_0(y-\theta)$. Letting $s(y) = y_n$ and $a_i = y_i - y_n, i = 1, \ldots, n-1$, we can write

$$f(y;\theta) = \prod f_0(y_i - \theta) = f_1(y_n|a;\theta)f_2(a) \tag{1}$$

where

$$f_2(a) = \int f_0(a_1 + t)\cdots f_0(a_{n-1} + t)f_0(t)\mathrm{d}t \tag{2}$$

and

$$f_1(y_n|a;\theta) = \frac{f_0(y_n + a_1 - \theta)\cdots f_0(y_n + a_{n-1} - \theta) f_0(y_n - \theta)}{f_2(a)}. \tag{3}$$

The numerator is just a rewriting of $\prod f(y_i;\theta)$, and an expression equivalent to (3) is

$$f_1(y_n|a;\theta) = \frac{f_0(y_1 - \theta)\cdots f_0(y_n - \theta)\mathrm{d}\theta}{\int f_0(y_1 - \theta)\cdots f_0(y_n - \theta)\mathrm{d}\theta}. \tag{4}$$

The density of A does not depend on θ, and the conditional density of Y_n given A is a location family density on $\mathbb{R}$.

The functions $s(Y)$ and $a(Y)$ are not uniquely determined, but they are uniquely determined up to a location transformation. We could for example let $S = \bar{Y}$ and $A_i = Y_i - \bar{Y}$. The vector A has n components but lies in $\mathbb{R}^{n-1}$ (as all its components must sum to 0, or in other words it is orthogonal to the 1-vector). The marginal density for a is again given by (2), and

$$f(y;\theta) \propto f(\bar{y}, a;\theta) = f_1(\bar{y}|a;\theta) f_2(a).$$

We might choose instead to let S be $\hat{\theta}$, the maximum likelihood estimate of θ, and define $A_i = Y_i - \hat{\theta}$.

Location-scale model A version of S and A that can be used for the location-scale model is $s(Y) = (\bar{Y}, s_Y)$, where $s_Y^2 = (n-1)^{-1}\sum(Y_i - \bar{Y})^2$, and $a_i(Y) = (Y_i - \bar{Y})/s_Y, \quad i = 1,\ldots,n$. The vector A has n components, but is restricted to lie in $\mathbb{R}^{n-2}$. To prove that the distribution of A is indeed free of θ, we write

$$\begin{aligned} f(y;\theta)\mathrm{d}y &= \theta_2^{-n}\prod\{f_0\left(\frac{y_i - \theta_1}{\theta_2}\right)\}\mathrm{d}y_1\ldots\mathrm{d}y_n \\ &= \theta_2^{-n}\prod f_0\left(\frac{a_i s + \bar{y} - \theta_1}{\theta_2}\right)|J|\mathrm{d}a_1\ldots\mathrm{d}a_n\mathrm{d}\bar{y}\mathrm{d}s \end{aligned} \tag{5}$$

where $|J|$ is the Jacobian of the transformation from y to $(a, \bar{y}, s)$. To compute $f(a)$ we need to integrate out $\bar{y}$ and s from this expression, so we need to figure out the dependence of $|J|$ on $\bar{y}$ and s. The computation is a little bit tricky, but by writing $y_i = a_i s + \bar{y}$ we can see that $\mathrm{d}y_i = s\mathrm{d}a_i$, and since a has $n-2$ free dimensions, the factor s^{n-2} will be part of the Jacobian. It turns out that this is the only part that depends on $\bar{y}$ and s. The details are presented in the next section. The result is

$$\begin{aligned} f(a)\mathrm{d}a &= \int\int \frac{s^{n-2}}{\theta_2^n} f_0\left(\frac{a_1 s + \bar{y} - \theta_1}{\theta_2}\right)\ldots f_0\left(\frac{a_n s + \bar{y} - \theta_1}{\theta_2}\right)\mathrm{d}\bar{y}\mathrm{d}s \\ &= \int\int \frac{(\theta_2 v)^{n-2}}{\theta_2^n} f_0\left(\frac{a_1\theta_2 v + \bar{y} - \theta_1}{\theta_2}\right)\ldots f_0\left(\frac{a_n\theta_2 v + \bar{y} - \theta_1}{\theta_2}\right)\mathrm{d}\bar{y}\theta_2\mathrm{d}v \\ &= \int\int \frac{v^{n-2}}{\theta_2} f_0\left(a_1 v + \frac{\bar{y} - \theta_1}{\theta_2}\right)\ldots f_0\left(a_n v + \frac{\bar{y} - \theta_1}{\theta_2}\right)\mathrm{d}\bar{y}\mathrm{d}v \\ &= \int\int v^{n-2} f_0(a_1 v + t)\ldots f_0(a_n v + t)\mathrm{d}t\mathrm{d}v \end{aligned} \tag{6}$$

which shows that the marginal distribution of A does not depend on θ. The conditional distribution of $s(Y)$, given A, is simply the ratio of the joint density to this marginal density. Again, the ancillary is not uniquely determined, but it is unique up to choice of location and scale variable. We could use $(a', y_{(1)}, y_{(n)} - y_{(1)})$, where $a_i' = (y_{(i)} - y_{(1)})/(y_{(n)} - y_{(1)})$, instead of $(a, \bar{y}, s)$, or many other equivalent formulations.

Details on the Jacobian:

As mentioned above, it is necessary to compute the Jacobian in the transformation from $\bar{y}$ to $(\bar{y}, s, a)$, where $a = (a_1, \ldots a_n)$ and $a_i = (y_i - \bar{y})/s$. This will be done below both algebraically and geometrically.

Since $s^2 = \sum (y_i - \bar{y})^2$, we can see that the vector a, although it has n components, in fact lies in $\mathbb{R}^{n-2}$, because $\sum a_i = a \cdot 1 = 0$ and $\sum a_i^2 = ||a||^2 = 1$. To compute the Jacobian we make the transformation one-to-one by letting

$$t_1 = \bar{y}, \quad t_2 = s, \quad t_i = a_i \qquad i = 3, \ldots, n;$$

note that we are explicitly using only $n - 2$ components of a. To find the inverse transformation we have

$$y_i = t_1 + t_2 t_i \qquad i = 3, \ldots, n$$

and using the restrictions on a we have

$$\begin{aligned} a_1 + a_2 &= -\sum_3^n t_i \\ 1 - a_1 - a_2 &= \sum_3^n t_i^2 \end{aligned}$$

from which we can write $a_1 = f_1(t_3, \ldots, t_n) = f_1(t_{(2)})$; $\quad a_2 = f_2(t_3, \ldots, t_n) = f_2(t_{(2)})$, say. Then

$$\begin{aligned} y_1 &= \bar{y} + s \cdot f_1(t_3, \ldots, t_n) \\ y_2 &= \bar{y} + s \cdot f_2(t_3, \ldots, t_n) \\ y_3 &= \bar{y} + s \cdot t_3 \\ &\vdots \\ y_n &= \bar{y} + s \cdot t_n \end{aligned}$$

is now a one-to-one transformation, with Jacobian determinant

$$\left| \frac{\partial y}{\partial t} \right| = \begin{vmatrix} 1 & f_1(t_{(2)}) & s f_{13}(t_{(2)}) & \ldots & s f_{1n}(t_{(2)}) \\ 1 & f_2(t_{(2)}) & s f_{13}(t_{(2)}) & \ldots & s f_{1n}(t_{(2)}) \\ 1 & t_3 & s & \ldots & 0 \\ 1 & t_4 & 0 & s & \ldots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_n & 0 & \ldots & s \end{vmatrix}$$

where $f_{ij}(t_{(2)}) = \partial f_i(t_{(2)})/\partial t_j, i = 1, 2; j = 3, \ldots, n$. A well-known formula for the determinant of a partitioned matrix shows that the Jacobian is of the form

$$s^{n-2}h(t_{(2)}) = s^{n-2}h(a)$$

which is the result we were looking for.

In the above derivation the location and scale estimates were $\bar{y}$ and s, with the result that a is orthogonal to the 1-vector, and has length 1. However, the same derivation applies for a variety of other location and scale estimates. For example, suppose we wanted to use the maximum likelihood estimates of μ and σ, which are defined as the solutions to the equations $\partial \log f(y; \hat{\mu}, \hat{\sigma})/\partial\mu = 0, \quad \partial \log f(y; \hat{\mu}, \hat{\sigma})/\partial\sigma = 0$ i.e.

$$\begin{aligned} \frac{-1}{\hat{\sigma}} \sum g'(\frac{y_i - \hat{\mu}}{\hat{\sigma}}) &= 0 \\ \frac{-n}{\hat{\sigma}} + \frac{y_i - \hat{\mu}}{\hat{\sigma}^2} \sum g'(\frac{y_i - \hat{\mu}}{\hat{\sigma}}) &= 0 \end{aligned}$$

where $g(y_i) = \log f(y_i)$. The ancillary statistic a is defined by $a_i = (y_i - \hat{\mu})/\hat{\sigma}$, so that $y_i = \hat{\sigma}a_i + \hat{\mu}$, and these two equations can be reexpressed as

$$\begin{aligned} \sum g'(a_i) &= 0 \\ \sum a_i g'(a_i) &= n \end{aligned}$$

which gives two restrictions on the a_i. Thus we can proceed as above and express $(\hat{\mu}, \hat{\sigma}, a_3, \ldots a_n)$ as a one-to-one function of $\bar{y}$, and find the inverse transformation. It is of exactly the same form (with s replaced by $\hat{\sigma}$), but the functions called f_1 and f_2 in the above derivation are different.

The geometric derivation of the result is actually very similar, but a little more elegant. It again uses $\bar{y}$ and s as coordinates to get the result, and then argues that this choice of coordinates is arbitrary. Although it's not necessary, it's a little bit easier to first define $z_i = (y_i - \mu)/\sigma$; we want to construct the conditional distribution of $\bar{z}, s(z)$, where $\bar{z} = n^{-1}\sum z_i$, and $s^2(\mathbf{z}) = \sum(z_i - \bar{z})^2$, given $\mathbf{d}(\mathbf{z}) = (\mathbf{z} - \bar{z}\,\mathbf{1})/||\mathbf{z} - z\,\mathbf{1}||$. Note that in terms of the original variables $\bar{z} = (\bar{y} - \mu)/\sigma$, $s(\mathbf{z}) = \sigma s$, and $\mathbf{d}(\mathbf{z}) = \mathbf{a}$. (Bold font is used for vectors here to try to clarify the geometric argument.)

Now we compute the Jacobian of the transformation from $\mathbf{z}$ to $\{\bar{z}, s(\mathbf{z}), \mathbf{d}(\mathbf{z})\}$ by figuring out what the differential element $d\mathbf{z}$ is in the new coordinates. That is, in the joint density of $\mathbf{z}$,

$$f(\mathbf{z})d\mathbf{z} = \prod f(z_i)dz_i$$

we consider the differential element $\prod dz_i$ as giving the volume of a small box at the point $\mathbf{z}$. We want to express this volume in the new coordinates. The coordinates $\{\bar{z}, s(\mathbf{z}), \mathbf{d}(\mathbf{z})\}$ provide locally orthogonal coordinates at the point $\mathbf{z} = \bar{z}\,1 + s(\mathbf{z})\mathbf{d}(\mathbf{z})$, and we want to know how they change as we change to point $\mathbf{z}$ to $\mathbf{z} + d\mathbf{z}$. The

coordinate specified by $\bar{z}$ lies on the $\mathbf{1}$-vector, so a small change in the coordinates $\mathbf{z}$ cause a change in $\bar{z}$ of $\sqrt{n}d\bar{z}$. Since $s(\mathbf{z})$ measures the length of $\mathbf{z} - \bar{z}\,\mathbf{1}$, its rate of change is simply ds. (Think of the picture in $\mathbb{R}^2$.) Now $\mathbf{d}(\mathbf{z})$ is orthogonal to the $\mathbf{1}$-vector, and lies on a unit sphere in the $n-1$-dimensional subspace of $\mathbb{R}^n$ that is orthogonal to the $\mathbf{1}$-vector. Thus the volume element is the surface volume on the sphere defined by $s(\mathbf{z})\mathbf{d}(\mathbf{z})$, i.e. the sphere of radius $s(\mathbf{z})$. This volume is $s(\mathbf{z})^{n-1}du$, where du is surface volume on the unit sphere in $\mathbb{R}^{n-1}$ (which is $n-2$-dimensional). (In fact an explicit expression for the surface area of the unit sphere in $\mathbb{R}^d$ is $(2\pi)^{d/2}/\Gamma(d/2)$.)

The coordinates in this development are orthogonal, so the volume element is the product of the three pieces. For this reason these coordinates are a convenient choice for computing the differential. However, if we choose to coordinatize the point using other location and scale functions, the result is unchanged. The only thing we need to make sure of is that the location coordinate, say $\tilde{\mu}(\mathbf{z})$, satisfies the property $\tilde{\mu}(a\mathbf{z}+b\mathbf{1}) = a\tilde{\mu}(\mathbf{z})+b$, the scale coordinate, say $\tilde{\sigma}(\mathbf{z})$, satisfies $\tilde{\sigma}(a\mathbf{z}+b\mathbf{1}) = a\tilde{\sigma}(\mathbf{z})$, and $\mathbf{d}$ is appropriately defined in terms of these two coordinates. We can show that such location and scale coordinates must themselves be location scale transformations of $\bar{z}$ and $s(\mathbf{z})$, so that we can convert the above result to a more general one. (Although $\hat{\mu}\mathbf{1}$ and $\hat{\sigma}$ will not give orthogonal coordinates, so that in these two dimensions the 'box' is a parallelogram, we can still figure out the volume by multiplying the base by the height!) Using the more general coordinates will not provide an explicit expression for the normalizing constant in terms of the surface area on the sphere, because the new vector $\mathbf{d}$ isn't forced to lie in the orthogonal complement of the $\mathbf{1}$-vector.

The geometric argument is from Chapter 2 of *Inference and Linear Models* (Fraser, 1979).