3 Principles of data reductions and inference

- 3.11 Ancillary examples
- 3.12 Maximum likelihood inference
- 3.13 Maximum likelihood estimation an introduction
- 3.14 Information and likelihood

3.1 Data reduction in statistical inference

In chapter one we were dealing with different probability models. These models can be used to describe the population of interest. Finding such a good model is our final goal. On the way towards this goal, we use the data to identify the parameters $\theta \in \Theta$ that best describes the model.

The statistical inference problem arises because we do not know the exact value of the parameter. The word parameter is used very generally here. The parameter could be a scalar (e.g., the probability of success in a binomial experiment), or a vector parameter (e.g., the vector with two components (μ, σ^2) for the normal distribution), or even a function if we perform nonparametric inference.

Suppose a vector $\mathbf{X} = (X_1, X_2, .., X_n)$ of n i.i.d. random variables, each with a density $f(x;\theta)$ is to be observed, and inference on $\theta \in \Theta$ based on the observations $x_1, x_2, .., x_n$ is to be made.

Let $\mathbf{X}$ takes values in $\mathcal{X}$ - the sample space. The statistician will use the information in the observations $x_1, x_2, \ldots, x_n$ to make inference about θ.

The statistician wants to summarize the information in the sample by determining a few key features of the sample values through transforming the sample values. Calculating such transformations means to calculate a **statistic.**

Typically, $\dim(T) << n$, i.e. using the statistic, we achieve the goal of data reduction. The statistic summarises the data in that, rather than reporting the entire sample $\mathbf{x}$, it reports only that $T(\mathbf{x}) = \mathbf{t}$.

The main purpose of this chapter is to discuss how the data reduction is to be performed in some sort of optimal way.

Data reduction in terms of a particular statistic can be thought of as *partitioning the sample space* $\mathcal{X}$ into disjoint subsets

$$A_t = \{\boldsymbol{X} : T(\boldsymbol{X}) = \boldsymbol{t}\}.$$

If $\tau = \{\mathbf{t} : \mathbf{t} = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ then $\mathcal{X}$ is represented as a union of disjoint sets (i.e. is partitioned):

$$\mathcal{X} = \bigcup_{t \in \tau} A_t.$$

Goal in data reduction:

When only using the value of the statistic $T(\mathbf{x})$ we want to "not to lose information" about θ. The whole information about θ will be contained in the statistic.

In particular, we will treat as equal *any* two samples $\mathbf{x}$ and $\mathbf{y}$ that satisfy

$$T(\mathbf{x}) = T(\mathbf{y})$$

even though the actual sample values may be different. Hence we arrive at the definition of sufficiency.

The information in $\mathbf{X}$ about θ can be discussed in terms of partitions of the sample space.

Definition 3.5 (Sufficient partition)

Suppose for *any* set A_t in a particular partition $\mathcal{A} = \{A_t, t \in \tau\}$ we have

$$P\{\mathbf{X} = \mathbf{x} \mid \mathbf{X} \in A_t\}$$

does not depend on θ. Then $\mathcal{A}$ is a sufficient partition for θ.

Remark 3.5

The partition is defined through a suitable statistic. If the statistic T is such that it generates a sufficient partition of the sample space then the statistic itself is sufficient.

3.2 Sufficent partition example

Exercise 3.13 (at lecture)

Suppose $\boldsymbol{X} = (X_1, X_2, .., X_n)$ are i.i.d. Bernoulli with parameter θ, i.e.

$$P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i = 0, 1.$$

The partition $\mathcal{A} = (A_0, A_1, \ldots, A_n)$ where $x \in A_r$ if and only if (iff)

$$\sum_{i=1}^{n} x_i = r,$$

is sufficient for θ. Correspondingly, the statistic

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

is sufficient for θ.

Given observed value $\mathbf{t}$ of $\mathbf{T}$, we know that the observed value $\mathbf{x}$ of $\mathbf{X}$ is in the partition set A_t.

Sufficiency means that $P(\mathbf{X} = \mathbf{x} \mid \mathbf{T} = \mathbf{t})$ is a function of $\mathbf{x}$ and $\mathbf{t}$ **only** (i.e., is **not** a function of θ).

Thus once having observed the particular realization $\mathbf{t}$ of $\mathbf{T}$, knowing in addition the particular value $\mathbf{x}$ of $\mathbf{X}$ would not help for a better identification of θ. Hence we arrive at the sufficiency principle:

3.3 Sufficiency principle

Remark 3.6 (Sufficiency principle)

The sufficiency principle implies that if T is sufficient for θ, then if x and y are such that $T(x) = T(y)$ then inference about θ should be the same whether $X = x$ or $Y = y$ is observed.

This leads to the following useful criterion:

3.4 Neyman Fisher factorization criterion

Theorem 3.10 (Neyman Fisher Factorization Criterion)

If $X_i \sim f(x, \theta)$ then $T(X) = T(X_1, X_2, \ldots, X_n)$ is sufficient for θ iff

$$L(X, \theta) = f_\theta(X_1, X_2, \ldots, X_n) = g(T(X), \theta)h(X)$$

where X, T, θ may all be vectors and $g \geq 0, h \geq 0$.

Proof: at lecture.

In other words, we try to factorise the joint distribution into a product of two non-negative functions of which one factor $(g(T(X), \theta))$ may depend on the parameter θ but depends on the sample only through the value of the statistic T whereas the other factor $(h(\mathbf{X}))$ does not depend on the parameter θ.

Sufficient partitions can be ordered and the **coarsest partition** (i.e. the one that contains the smallest number of sets) is called the **minimal sufficient partition.**

Suppose T is sufficient and $T(X) = g_1(U(X))$ where U is a statistic and g_1 is a known function. It can be seen that U must also be sufficient for θ then. Indeed:

$$\begin{aligned} L(X, \theta) &= g(T(X), \theta)h(X) \\ &= g(g_1(U(X)), \theta)h(X) \\ &= \bar{g}(U(X), \theta)h(X) \end{aligned}$$

which means that $U(X)$ is also sufficient.

But, generally speaking, U induces a finer (or at least no coarser) partition than T since it might happen that $U_1 \neq U_2$ but yet $g_1(U_1) = g_1(U_2)$. Thus a finer partition of any sufficient partition is sufficient.

In applications we look for the **coarsest** partition that is still sufficient because this means the greatest data reduction without loss of information on θ.

From the above, we see that the statistic that introduces this coarsest partition will be a function of any other sufficient statistics. Such a statistic is called the **minimal sufficient** statistic.

Properties of sufficient statistics:

i) If T is sufficient, so is any one-to-one function of T (since it generates the same partition);

ii) If T is minimal sufficient, it is necessarily a function of all other possible sufficient statistics;

iii) If T is sufficient then $P(\mathbf{x} \mid \mathbf{t})$ does not depend on θ. The observed $\mathbf{t}$ is a summary of $\mathbf{x}$ that contains all the information about θ in the data, under the given family of models. It divides the sample space $\mathcal{X}$ into disjoint subsets A_t, each containing all possible observations $\mathbf{x}$ with the same value $\mathbf{t}$.

3.5 Sufficiency examples

Exercise 3.14 (at lecture)

For the following distributions find the sufficient statistic and provide justification.

i) Bernoulli

ii) Univariate normal distribution with unknown μ and σ^2

iii) Uniform distribution in $[0, \theta)$

iv) Multivariate normal with unknown μ and Σ.

See next two slides for some useful formula.

Fundamental equality 1-dimensional

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

Proof:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2&=\sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)^2\\&=\sum_{i=1}^{n}(X_i-\bar{X})^2+2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu)+\sum_{i=1}^{n}(\bar{X}-\mu)^2\\&=\sum_{i=1}^{n}(X_i-\bar{X})^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})+n(\bar{X}-\mu)^2\\&=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2\end{aligned}$$

Fundamental equality p-dimensional

A p-dimensional version of the fundamental equality also exists:

$$\sum_{i=1}^{n}(X_i-\mu)(X_i-\mu)^\top = \sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^\top + n(\bar{X}-\mu)(\bar{X}-\mu)^\top$$

We also need to use the following property of traces:

$$X^\top AX = \text{tr}(A(XX^\top))$$

where $\text{tr}(A)$ is the sum of the diagonal elements of the matrix.

3.6 Lehmann and Scheffe's method for constructing a minimal sufficient partition

Often the sufficient statistic which has been found by the Factorization criterion, turns out to be minimal sufficient. However, to find a general method for constructing a minimal sufficient statistic is a difficult task.

Theorem 3.11 (Lehmann- Scheffe's method)

Consider a partition $\mathcal{A}$ of $\mathcal{X}$ by defining for any

$$x \in \mathcal{X} : A(x) = \left\{ y : \frac{L(y,\theta)}{L(x,\theta)} \quad \text{does not depend on } \theta \right\}$$

That is, the ratio is a function of the type $h(y,x)$. The above defined sets $\{A(x), x \in \mathcal{X}\}$ form a partition of $\mathcal{X}$ and this partition is *minimal sufficient.*

Proof. (discrete case only for simplicity; proof could be skipped if you are not interested in the details)

Step one: We have to show that the sets $\{A(x), x \in \mathcal{X}\}$ form a partition. To this end, we have to show that they are *either disjoint or coincide*.

If we assume there exists a joint element $z \in A(x) \bigcap A(u)$ then $A(x)$ and $A(u)$ must coincide. To see this, consider the following:

i) Take any other $x_0 \in A(x)$ and any $u_0 \in A(u)$.

ii) Then

$$\frac{L(z,\theta)}{L(x_0,\theta)} = \frac{L(z,\theta)}{L(x,\theta)} \Big/ \frac{L(x_0,\theta)}{L(x,\theta)}$$

is *not* a function of θ because each of the two ratios on the RHS are not functions of θ.

iii) $\dfrac{L(z,\theta)}{L(u,\theta)}$ is also *not* a function of θ since $z \in A(u)$.

iv) But then

$$\frac{L(x_0,\theta)}{L(u,\theta)} = \frac{L(x_0,\theta)}{L(x,\theta)} \cdot \frac{L(x,\theta)}{L(z,\theta)} \cdot \frac{L(z,\theta)}{L(u,\theta)}$$

is *not* a function of θ since the RHS is *not* a function of θ.

The conclusion from this chain of statements is that an *arbitrary* $x_0 \in A(x)$ must belong to $A(u)$ as well.

But in the same way as above it can be argued that $u_o \in A(x)$ and u_0 was *arbitrary* in $A(u)$. Therefore it must hold $A(x) = A(u)$ if they had one joint element z. This shows that $\{A(x), x \in \mathcal{X}\}$ is a partition.

Step two: We want to show the above defined partition $\mathcal{A}$ is *minimal sufficient*. Remember that we consider the discrete case only. First, we show that the partition is *sufficient*. Fix x and consider the conditional probability:

$$P(Y = y \mid Y \in A(x)) = \frac{P_\theta(Y = y, Y \in A(x))}{P_\theta(Y \in A(x))}$$

$$= \begin{cases} 0 & \text{if } y \text{ is not in } A(x) \\ \frac{P_\theta(Y=y)}{P_\theta(Y \in A(x))} & \text{if } y \in A(x) \end{cases}$$

But since

$$\frac{P_\theta(Y = y)}{P_\theta(Y \in A(x))} = \frac{P_\theta(Y = y)}{\sum_{z \in A(x)} P_\theta(Y = z)} = \frac{P_\theta(y)}{\sum_{z \in A(x)} h(z, x) P_\theta(x)}$$

is *not* a function of θ, we see that always $P(Y = y \mid Y \in A(x))$ does not depend on θ, that is, $\mathcal{A}$ is a sufficient partition.

Now we want to show that $\mathcal{A}$ is a *minimal sufficient partition.* Take any $A(x)$. Fix any $y \in A(x)$. Assume that $v = v(Y)$ is also sufficient and creates a coarser partition.

If y and z are such that $v(y) = v(z)$ then by the factorization theorem (v is assumed to be sufficient) we have:

$$L(y,\theta) = g(v(y),\theta)h^*(y) = g(v(z),\theta)h^*(y) = \frac{L(z,\theta)}{h^*(z)}h^*(y)$$

i.e. $\dfrac{L(z,\theta)}{L(y,\theta)}$ is not a function of θ.

But this means

$$\frac{L(z,\theta)}{L(x,\theta)} = \frac{L(z,\theta)/L(y,\theta)}{L(x,\theta)/L(y,\theta)}$$

is *not* a function of θ, because the RHS is *not*. Hence y and z are in the same $A(x)$ class. This means that the partition $A(x)$ includes the partition generated by v and so, $\mathcal{A}$ must be the *coarsest* partition. □

3.7 Minimal sufficient examples

Exercise 3.15 (at lecture)

For the following distributions find the minimal sufficient statistic and provide justification.

i) i.i.d. Bernoulli;

ii) i.i.d. normal with unknown μ and σ^2;

iii) i.i.d. uniform in $(0, \theta)$;

iv) i.i.d. Cauchy(θ) - an example that shows that sometimes the dimension of the minimal sufficient statistics can be quite large, even equal to the sample size n itself.

3.8 One parameter exponential family densities

A density $f(x,\theta)$ is a **one parameter exponential family density** if $\theta \in \Theta \in R^1$ and

$$f(x,\theta) = a(\theta)b(x)\exp(c(\theta)d(x))$$

with $c(\theta)$ strictly monotone.

The joint density of n i.i.d observations is:

$$\begin{aligned} L(x,\theta) &= \prod_{i=1}^{n} a(\theta)b(x_i)\exp(c(\theta)d(x_i)) \\ &= a(\theta)^n \prod_{i=1}^{n} b(x_i)\exp\left(c(\theta)\sum_{i=1}^{n} d(x_i)\right) \end{aligned}$$

In this case we have:

$$\frac{L(x,\theta)}{L(y,\theta)} = \prod_{i=1}^{n} \frac{b(x_i)}{b(y_i)} \exp\left(c(\theta)\left[\sum_{i=1}^{n} d(x_i) - \sum_{i=1}^{n} d(y_i)\right]\right)$$

which is *not* a function of θ iff

$$\sum_{i=1}^{n} d(x_i) = \sum_{i=1}^{n} d(y_i).$$

So, if x is any point in $\mathcal{X}$ then

$$A(x) = \left\{y : \sum_{i=1}^{n} d(x_i) = \sum_{i=1}^{n} d(y_i)\right\}$$

and the sets in the minimal sufficient partition are contours of $\sum_{i=1}^{n} d(x_i)$. Hence, $T = \sum_{i=1}^{n} d(x_i)$ is minimal sufficient.

Remark 3.7

It turns out that a lot of the standard distributions can be seen to belong to the one parameter exponential family:

- $f(x,\theta) = \theta \exp(-\theta x),\ x > 0,\ \theta > 0$
- Poisson(θ)
- Bernoulli (θ);
- $N(\theta, 1)$;
- $N(0, \theta^2)$

and others. However, that there are many distributions outside the above class, too. For example, the uniform $(0,\theta)$ distribution or the Cauchy(θ) distribution do *not* belong to exponential family.

Example 3.19

The exponential density:

$$f(x,\theta) = \theta \exp(-\theta x), \quad x > 0, \quad \theta > 0$$

belongs to the one parameter exponential family of densities, with

$$a(\theta) = \theta, \quad b(x) = 1, \quad c(\theta) = -\theta \quad \text{and} \quad d(x) = x.$$

Therefore, $T(X) = \sum_{i=1}^{n} X_i$ is minimal sufficient.

Exercise 3.16 (at lecture)

Show that the following densities belong to the exponential family of densities and identify the minimal sufficient statistic for each of the distributions.

i) Poisson(θ) : $f(x,\theta) = \dfrac{e^{-\theta}\theta^x}{x!}$, $x \in \{0,1,2,\ldots\}$, $\theta > 0$

ii) Bernoulli(θ) : $f(x,\theta) = \theta^x(1-\theta)^{1-x}$, $x \in \{0,1\}$, $\theta \in (0,1)$

iii) Normal $N(\theta,1)$: $f(x,\theta) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$, $x \in R$, $\theta \in R$

iv) Normal $N(0,\theta^2)$: $f(x,\theta) = \dfrac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{x^2}{2\theta^2}}$, $x \in R$, $\theta^2 > 0$.

3.9 Generalization to a k- parameter exponential family

It is natural to define the k-parameter exponential family $(k \geq 1)$ via:

$$f(x;\theta_1,\ldots,\theta_k) = a(\theta_1,\ldots,\theta_k)b(x)\exp\left(\sum_{j=1}^{k} c_j(\theta_1,\ldots,\theta_k)d_j(x)\right)$$

where $c_j(.)$ are certain smooth functions of the k-dimensional parameter vector $\theta = (\theta_1,\ldots,\theta_k)'$.

In order to avoid degenerate cases, it is also required that the $k \times k$ matrix of partial derivatives

$$\left\{ \frac{\partial c_j}{\partial \theta_l} \right\}, \quad j = 1, \ldots, k; \quad l = 1, \ldots, k$$

has a non-zero determinant.

The minimal sufficient (vector) statistic is:

$$T = \left(\sum_{i=1}^{n} d_1(X_i), \ldots, \sum_{i=1}^{n} d_k(X_i) \right)^{\top}.$$

Example 3.20

Consider a Normal density which belongs to the two-parameter exponential family of densities.

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x^2-2x\mu+\mu^2)\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\exp\left(\frac{\mu}{\sigma^2}x-\frac{x^2}{2\sigma^2}\right). \end{aligned}$$

We have

$$d_1(x)=x, \quad d_2(x)=x^2 \quad \text{and} \quad \bar{T}=\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)^\top$$

is minimal sufficient for $\theta=(\mu,\sigma^2)^\top$.

Exercise 3.17 (at lecture)

Show that the Beta(θ_1, θ_2) density belongs to the two-parameter exponential family of densities and identify the sufficient statistic for θ_1 and θ_2 given that

$$f(x, \theta_1, \theta_2) = \frac{1}{B(\theta_1, \theta_2)} x^{\theta_1 - 1}(1-x)^{\theta_2 - 1}, \quad x \in (0,1) \quad \theta_1, \theta_2 > 0$$

Here $B(\theta_1, \theta_2)$ is the Beta function that has given rise to the name of the density family. It serves to norm the density to integrate to one.By definition,

$$B(\theta_1, \theta_2) = \int_0^1 x^{\theta_1 - 1}(1-x)^{\theta_2 - 1} dx.$$

3.10 Ancillary statistic and ancillarity principle

Consider again $X = (X_1, X_2, \ldots, X_n)$: i.i.d. with $f(x, \theta), \theta \in R^k$.

Definition 3.6

A statistic is called ancillary if its distribution does not depend on θ.

Intuitively, it would seem that knowledge of an ancillary should not help in inference about θ. However, an ancillary, when used *in conjunction* with another statistic, sometimes *does help* in inference about θ. Inference for θ could be improved if done conditionally on the ancillary.

The most important case occurs when a statistic $\mathbf{T}$ is minimal sufficient for θ but its dimension is *greater* than that of θ.

We can write $\mathbf{T}=(\mathbf{T}_1',\mathbf{T}_2')'$ where $\mathbf{T}_2$ has a marginal distribution not depending on θ. The distribution of $\mathbf{T}_2$ is the same for all $P_\theta \in \mathcal{P}$. Then $\mathbf{T}_2$ is ancillary and $\mathbf{T}_1$ is *conditionally sufficient given* $\mathbf{T}_2$:

$$L(\mathbf{x},\theta) \propto L_1(\mathbf{t}_1 \mid \mathbf{t}_2,\theta)L_2(\mathbf{t}_2)$$

Inference about θ then, according to the **ancillarity principle** should be based on $L_1(\mathbf{t}_1 \mid \mathbf{t}_2,\theta)$.

3.11 Ancillary examples

Example 3.21 (at lecture)

Assume: n i.i.d. $X_1, \ldots, X_n$ uniform in $(\theta, 1+\theta)$. You can show: the minimal sufficient statistic for θ is $T = (X_{(1)}, X_{(n)})$. Then

$$T^* = (X_{(n)} - X_{(1)}, X_{(n)} + X_{(1)})$$

is also minimal sufficient!

Solution:

Denoting $Z_i = X_i - \theta$ we see that Z_i are i.i.d. uniformly distributed in $[0,1]$ and their distribution does not involve θ :

$$P(X_{(n)} - X_{(1)} < r) = P[(X_{(n)} - \theta) - (X_{(1)} - \theta) < r] = P(Z_{(n)} - Z_{(1)} < r).$$

Therefore no dependence on θ whatsoever. Hence the first component $X_{(n)} - X_{(1)}$ of the minimal sufficient statistic: ancillary.

If you were to make inference about θ you would like to take note of value of $X_{(n)} - X_{(1)}$ and to condition on it. Intuitively, if this value is close to 0 then your inference about θ (based on $\frac{1}{2}(X_{(1)} + X_{(n)} - 1)$ for example) may be very unprecise but it would be very precise if $X_{(n)} - X_{(1)}$ was close to 1.

Example 3.22 (Inference in case of normal mixture.)

Assume that the density of Y is given by

$$f_Y(y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-0.5(y-\mu)^2/\sigma_1^2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-0.5(y-\mu)^2/\sigma_2^2}.$$

Solution:

If we also observe an indicator random variable C (with values 1 or 2 telling us whether the first or the second component of the mixture has been observed) then it becomes clear which is the distribution that has generated Y.

Hence the joint distribution is

$$f_{C,Y}(c,y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_c}e^{-0.5(y-\mu)^2/\sigma_c^2}$$

The statistic $S = (C, Y)$ is sufficient for μ when σ_1^2, σ_2^2 are assumed known. Moreover since $P(C = 1) = P(C = 2) = 0.5$, C is ancillary. Conditioning on C can definitely help in our inference about μ.

Example 3.23

Let $X = (X_1, X_2, \ldots, X_n)$ be i.i.d. from a location family with cdf $F_\theta(x) = F(x-\theta) = F_0(x-\theta),\ \theta \in R^k$. Consider $T_2 = X_2 - X_1$. This statistic is ancillary.

Solution:

First, note that if $X \sim F_\theta$, then $X - \theta \sim F_0$ since:

$$\begin{aligned} F_{X-\theta}(x) &= P(X-\theta < x) = P(X < x+\theta) \\ &= F_\theta(x+\theta) = F_0(x+\theta-\theta) = F_0(x) \end{aligned}$$

Hence,the distribution of $X_i - \theta,\ i = 1, 2, \ldots, n$ does not depend on θ.

However,

$$F_{T_2}(y,\theta) = P_\theta(T_2 < y) = P\Big\{[(X_2-\theta)-(X_1-\theta)] < y\Big\}$$

and the latter expression obviously does not depend on θ. Hence T_2 is ancillary.

Following the same logic, $\tilde{T}_2 = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ is also ancillary.

Definition 3.7

The statistic $\hat{\theta}(X)$ is called **equivariant**, if

$$\hat{\theta}(X_1 + C, X_2 + C, \ldots, X_n + C) = \hat{\theta}(X_1, X_2, \ldots, X_n) + C$$

for any vector C with appropriate dimension.

The importance of ancillarity can be seen by the following theorem:

Theorem 3.12

If $\tilde{\theta}$ is **any** equivariant estimator with $E_0(\tilde{\theta}) < \infty$ then the estimator

$$\hat{\theta}_P = \tilde{\theta} - E_0(\tilde{\theta} \mid \tilde{T}_2), \qquad \tilde{T}_2 = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$$

is the **best equivariant** estimator (i.e. with uniformly smallest risk with respect to quadratic loss among all equivariant estimators). It is the so called **Pitman** estimator.

It can be shown that for square error loss, and $\theta \in R^1$ the Pitman estimator is given in a closed form as

$$\hat{\theta}_P = \frac{\int_{-\infty}^{\infty} \theta \prod_{i=1}^{n} f(X_i - \theta) d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} f(X_i - \theta) d\theta}$$

where $f(x-\theta) = f_\theta(x)$ denotes the density of a single observation.

3.12.1 Likelihood principle

Let $\mathbf{X} = (X_1, X_2, .., X_n)$ be i.i.d. each with density $f(x, \theta)$. Given an observation $\mathbf{x}$ of $\mathbf{X}$, we substitute in

$$L(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

which becomes a function of θ only. This is called the Likelihood function.

Other functions of θ in the form $c(\mathbf{x})L(\mathbf{x}, \theta)$ can also be called likelihood functions.

The *MLE* $\hat{\theta}$ (that maximizes L or, equivalently, g with respect to θ) will be a function of every sufficient statistic. In particular, the MLE will be a function of the *minimal sufficient statistic* when the latter exists.

If two points $\mathbf{x}$ and $\mathbf{y}$ are in the same set in the minimal sufficient partition then

$$L(\mathbf{y}, \theta) = h(\mathbf{y}, \mathbf{x}) L(\mathbf{x}, \theta),$$

which means they give rise to proportional likelihood functions and the same value of the minimal sufficient statistic. These values $\mathbf{x}$ and $\mathbf{y}$ must lead to the same inference about θ.

An even stronger version of this requirement is the weak likelihood principle:

> "Data sets with proportional likelihood functions lead to identical conclusions".

We say the version is "stronger" since it does not necessitate the sampling processes to be identical. If *sampling processes* A and B lead to likelihood functions $L_A(\mathbf{x}, \theta)$ and $L_B(\mathbf{y}, \theta)$ such that

$$\frac{L_A(\mathbf{x}, \theta)}{L_B(\mathbf{y}, \theta)}$$

does not depend on θ, inference about θ should be the same.

3.12.2 Weak likelihood principle examples

Example 3.24

Experiment A:

Observe $\boldsymbol{x} = (x_1, x_2, .., x_n)$: n i.i.d. Bernoulli with parameter θ. Then

$$L_A(\mathbf{x}, \theta) = \theta^k (1 - \theta)^{n-k}$$

if it happened that there were k outcomes equal to one in $\mathbf{x}$.

Example 3.25

Experiment B:

Observe one realization $\boldsymbol{y}$ of a single random variable $\mathbf{Y}$ =number of successes in n i.i.d. Bernoulli trials. Then

$$L_B(\boldsymbol{y}, \theta) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k}$$

if it happened that $\mathbf{y} = k$.

Example 3.26

Experiment C:

Observe realization of a random variable $\mathbf{Z}$ - number of trials until k successes:

$$P_\theta(\mathbf{Z} = z) = \begin{pmatrix} z-1 \\ k-1 \end{pmatrix} \theta^k (1-\theta)^{z-k}, z = k, k+1, \ldots$$

Here the number of trials is random but if it happened that $z = n$ then

$$L_C(n, \theta) = \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} \theta^k (1-\theta)^{n-k}.$$

Hence, in all three cases we would get proportional likelihood functions for specific realizations of the variables $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ and the conclusions about θ in these circumstances must be identical.

3.13 Maximum likelihood estimation an introduction

We have discussed important principles such as *sufficiency*, *ancillarity*, *weak likelihood principle*. Each looks reasonable but we want a *constructive procedure* for finding reasonable estimators of the parameter of interest which leads to the MLE defined as

$$\hat{\theta} = \arg\left[\sup_{\theta \in \Theta} L(\mathbf{x}, \theta)\right].$$

In the discrete case the interpretation of the above maximization is that we look at the model that makes the observed data most likely (probable).

Because here it goes about comparing different models, it makes sense to introduce the **normed likelihood**:

$$R(\mathbf{x}, \theta) = \frac{L(\mathbf{x}, \theta)}{L(\mathbf{x}, \hat{\theta})}$$

which has a range of [0,1]. For a fixed $\mathbf{x}$, it is just a function of θ and we denote it by $R(\theta)$.

Example 3.27

If a coin is tossed 100 times and yields, say, 32 heads then the maximum likelihood estimator $\hat{\theta}$ of the probability of a head to occur is $\hat{\theta} = 0.32$ and

$$R(\theta) = \frac{\theta^{32}}{0.32^{32}} . \frac{(1-\theta)^{68}}{0.68^{68}}.$$

An even more often used measure is the deviance $D(\theta)$ which is defined as

$$D(\theta) = -2\ln R(\theta) = -2[\ln L(\mathbf{x}, \theta) - \ln L(\mathbf{x}, \hat{\theta})].$$

The deviance is a non-negative number. The larger the deviance, the further the model under consideration from the most likely model. This can be used to construct confidence intervals for the parameter.

3.14 Information and likelihood

In this section, we would like to quantify the notion of Fisher Information in a single observation and in the whole data vector with respect to the parameter of interest.

We will see that having done this in a proper way, we will be able to demonstrate quantitatively (with numbers) that, indeed, when we are using a sufficient statistic, we are preserving the information about the parameter that is contained in the whole sample.

On the contrary, if we are not using sufficient statistics, we are reducing the amount of information about the parameter that is contained in the whole sample.

3.14.1 Score function

This section will focus more closely on the precise mathematical definition of Fisher Information and will investigate its properties.

We define:

$$V(\mathbf{X}, \theta) = \frac{\partial}{\partial \theta} \ln L(\mathbf{X}, \theta)$$

to be the score function. That is, the score is the derivative of the Log-Likelihood function with respect to the parameter. If the parameter is multi-dimensional (i.e., a parameter vector) then the score is a vector function consisting of the partial derivatives of $\ln L(\mathbf{X}, \theta)$ with respect to each component of the vector θ.

Generally speaking, the score can be defined in the above way even for non-i.i.d. random variables where $L(\mathbf{X}, \theta)$ is the joint density. However, in our course we are mostly interested in the case where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and the X_i are i.i.d. with a density $f(x, \theta)$. Then

$$\begin{aligned} V(\mathbf{X}, \theta) &= \frac{\partial}{\partial \theta} \log \prod_{i=1}^{n} f(X_i, \theta) \\ &= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log f(X_i, \theta) \\ &= \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \theta} f(X_i, \theta)}{f(X_i, \theta)} \end{aligned}$$

The following properties of the score are important:

i) If $\hat{\theta}$ is the MLE then $V(\mathbf{X}, \hat{\theta}) = 0$ holds. This holds because $\hat{\theta}$ maximises $L(\mathbf{X}, \theta)$ or equivalently maximises $\log L(\mathbf{X}, \theta))$ with respect to θ.

ii) $E_\theta(V(\mathbf{X}, \theta)) = 0$ holds under suitable regularity conditions allowing exchange of order of integration and differentiation in the calculation of the expected value.

Example 3.28

Show that the property $E_\theta(V(\mathbf{X},\theta)) = 0$ holds under regularity conditions allowing exchange of order of integration and differentiation in the calculation of the expected value.

Solution:

For simplicity take θ to be one-dimensional. (For a vector θ apply the argument below for each of the components of θ).

Use the short-hand notation:

$$\mathrm{d}\mathbf{X} = \mathrm{d}X_1\mathrm{d}X_2...\mathrm{d}X_n$$

and a single integral sign to denote the integration over the region in $\mathbb{R}^n$.

Then

$$E\left[\frac{\partial}{\partial\theta}\log L(\mathbf{X},\theta)\right] = \int \frac{\frac{\partial}{\partial\theta}L(\mathbf{X},\theta)}{L(\mathbf{X},\theta)} L(\mathbf{X},\theta)\mathrm{d}\mathbf{X}$$
$$= \frac{\partial}{\partial\theta}\int L(\mathbf{X},\theta)\mathrm{d}\mathbf{X} = \frac{\partial}{\partial\theta}1 = 0$$

(Where we used the fact that the integral of any density over its support is equal to one).

The above-defined notion of Information is fundamental in Statistics. It has made RA Fisher (1890-1962), a renowned applied statistician, one of the greatest of all time. Among his pivotal contributions to the field, the introduction of the Maximum Likelihood estimation method, the analysis of variance and the notion of Expected Fisher Information (as defined above) are the most outstanding.

3.14.2 Expected Fisher information about θ contained in the vector $\mathbf{X}$

Expected Fisher Information about θ contained in the data vector $\boldsymbol{X}$ is denoted by $I_{\mathbf{X}}(\theta)$ and defined as:

$$I_{\mathbf{X}}(\theta) = Var_{\theta}(V(\mathbf{X}, \theta)) = E_{\theta}\left[\frac{\partial \ln L(\mathbf{X}, \theta)}{\partial \theta}\right]^2$$

where we utilised the fact that $E_{\theta}(V(\mathbf{X}, \theta)) = 0$.

3.14.3 Some properties of information

i) Additivity over independent samples: if X and Y are independent random variables whose densities depend on θ then for the information in the vector $\mathbf{Z} = (X, Y)$ we have

$$I_{\mathbf{Z}}(\theta) = I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

In particular, when sampling n times, the information in the sample about the parameter equals n times the information in a single observation about the parameter: If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ then

$$I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta).$$

ii) If $T(X)$ is sufficient for θ then $I_T(\theta) = I_X(\theta)$ and there is no information loss.

iii) Under some regularity conditions:

$$I_X(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln L(X,\theta)\right)$$

iv) For any statistics $T(X)$ it holds: $I_T(\theta) \leq I_X(\theta)$ with equality if and only if T is sufficient for θ. This property most clearly underlines the importance of sufficiency when we try to perform data reduction without loss of information about the parameter!

Sketch of proofs: (full discussion at lecture)

i) Starting with

$$L_{(X,Y)}(x,y;\theta) = L_X(x;\theta)L_Y(y;\theta),$$

we take logarithms of both sides first and then calculate partial derivatives with respect to θ of both sides. In the resulting equality, we square both sides and take expected values. This gives us:

$$E_\theta\left(\left[\frac{\partial}{\partial\theta}\log L_{(X,Y)}(X,Y,\theta)\right]^2\right) = I_X(\theta) + I_Y(\theta) + 2E_\theta[V(X,\theta)V(Y,\theta)].$$

Since X and Y are independent:

$$E_\theta[V(X,\theta)V(Y,\theta)] = E_\theta[V(X,\theta)]E_\theta[V(Y,\theta)] = 0$$

holds (using the property of the score) and we end up with $I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$.

ii) We will only consider the discrete case. First, let us note that because of the sufficiency,

$$\begin{aligned} f_T(t,\theta) &= \sum_{x:T(x)=t} f_X(x,\theta) \\ &= \sum_{x:T(x)=t} g(T(x),\theta)h(x) \\ &= g(t,\theta) \sum_{x:T(x)=t} h(x) \end{aligned}$$

holds and hence

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\log f_T(T;\theta)\right]^2 &= E_\theta\left[\frac{\partial}{\partial\theta}\log g_T(T;\theta) + \frac{\partial}{\partial\theta}\log \sum_{x:T(x)=t} h(x)\right]^2 \\ &= E_\theta\left[\frac{\partial}{\partial\theta}\log g_T(T;\theta)\right]^2. \end{aligned}$$

Then

$$
\begin{aligned}
I_T(\theta) &= E_\theta\left[\frac{\partial}{\partial\theta}\log f_T(T;\theta)\right]^2 \\
&= E_\theta\left[\frac{\partial}{\partial\theta}\log g(T;\theta)\right]^2 \\
&= E_\theta\left[\frac{\partial}{\partial\theta}(\log g(T;\theta)+\log h(X))\right]^2 \\
&= E_\theta\left[\frac{\partial}{\partial\theta}\log L(X;\theta)\right]^2 \\
&= I_X(\theta).
\end{aligned}
$$

iii) If $f(x,\theta)$ denotes the density of a single observation and under suitable differentiability conditions, we can write:

$$\frac{\partial^2}{\partial\theta^2}(\log f(x,\theta)) = \frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)} - \left[\frac{\frac{\partial}{\partial\theta}f(x,\theta)}{f(x,\theta)}\right]^2$$

For the case of a sample size $n = 1$, we see that if we take expected values in the above equality, property iii) would be shown if we are able to show that

$$E_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}\right] = 0$$

holds.

But under suitable regularity conditions that allow for exchange of order of integration and differentiation, we have

$$E_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(x,\theta)}{f(x,\theta)}\right] = \frac{\partial^2}{\partial\theta^2}\int f(x,\theta)dx = \frac{\partial^2}{\partial\theta^2}1 = 0.$$

Therefore statement iii) is shown for the case $n = 1$. For the case of arbitrary sample size, we use the additivity of the information over independent samples to get

$$I_X(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\ln L(X,\theta)\right).$$

iv) We need two properties of conditional expected values:

For random variables Z and Y and a function $g(z)$ we can write (under "suitable conditions"):

$$E(g(Z)Y|Z=z) = g(z)E(Y|Z=z) \tag{1}$$

$$E(Y) = E_Z(E(Y|Z=z)) \tag{2}$$

Since the expected value of the square of any random variable is non-negative:

$$\begin{aligned} 0 &\leq E\left\{\frac{\partial}{\partial\theta}\log L(X,\theta) - \frac{\partial}{\partial\theta}\log f_T(T,\theta)\right\}^2 \\ &= I_X(\theta) + I_T(\theta) - 2E\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)\right] \end{aligned} \tag{3}$$

If we were able to show in (3) that

$$E\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)\right] = I_T(\theta) \tag{4}$$

holds then from (3) we would have as a consequence that $I_X(\theta) - I_T(\theta) \geq 0$ holds which means that $I_T(\theta) \leq I_X(\theta)$, that is, the information in the statistic can not exceed the information in the sample.

Now we concentrate on showing (4). Using properties (1) and (2) we can write

$$\begin{aligned} &E\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)\right] \\ &\quad = E_T\left[\frac{\partial}{\partial\theta}\log f_T(t,\theta)E\left(\frac{\partial}{\partial\theta}\log L(X,\theta)|T=t\right)\right] \end{aligned} \tag{5}$$

To show now:

$$E\left(\frac{\partial}{\partial\theta}\log L(X,\theta)|T=t\right)=\frac{\partial}{\partial\theta}\log f_T(t,\theta).$$

This is the famous Fisher's identity.

Then substitution in (5) shows that (4) holds.

We also that the only way in which we may end up with an equality $I_T(\theta) = I_X(\theta)$ is if we had equality in (4). But this is only possible if

$$\frac{\partial}{\partial\theta}\log L(X,\theta) = \frac{\partial}{\partial\theta}\log f_T(T,\theta).$$

Hence

$$\log L(X,\theta) - \log f_T(T,\theta)$$

does **not** depend on θ. Denote this difference by $\log h(X)$, say, then

$$\log L(X,\theta) = \log f_T(T,\theta) + \log h(X).$$

This means that $L(X,\theta)$ can be factorized as in the Neyman Fisher criterion and T must be sufficient. □