3 Lecture 3: PRINCIPLES OF DATA REDUCTION AND INFERENCE

3.1 Data Reduction in Statistical Inference

Given vector $\mathbf{X}=(X_1, X_2, .., X_n)$ of n i.i.d. random variables, each with a density $f(x;\theta)$, we are meant to conduct inference on $\theta \in \Theta$ based on the observations $x_1, x_2, .., x_n$. Let $\mathbf{X}$ takes values in $\mathcal{X}$ - the sample space. The statistician uses the information in the observations $x_1, x_2, .., x_n$ to conduct the inference. His/her wish is 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 (i.e. functions of the sample) means to calculate a **statistic.** Typically , $\dim(\mathbf{T}) << n$, i.e. using the statistic, we achieve the goal of data reduction: rather than reporting the entire sample $\mathbf{x}$, the statistic reports only that $\mathbf{T}(\mathbf{x})=\mathbf{t}$. Data reduction in terms of a particular statistic can be thought of as a *partition of the sample space*. We partition $\mathcal{X}$ into disjoint subsets $A_t = \{\mathbf{X}:\mathbf{T}(\mathbf{X})=\mathbf{t}\}$. If $\tau = \{\mathbf{t}:\mathbf{t}=\mathbf{T}(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ then the sample space $\mathcal{X}$ is represented as a union of the following disjoint sets (i.e. is partitioned) : $\mathcal{X} = \bigcup_{t\in\tau} A_t$. The ultimate goal in the data reduction is, when only using the value of the statistic $T(\mathbf{x})$ instead of the whole vector $\mathbf{x}$,"not to lose information" about the parameter of interest θ. The whole information about θ will be contained in the statistic and, in particular, we will treat as equal *any* two samples $\mathbf{x}$ and $\mathbf{y}$ that satisfy $\mathbf{T}(\mathbf{x})=\mathbf{T}(\mathbf{y})$ even though the actual sample values may be different. That way 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 1(**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.

Note: We have seen above that the partition is defined through a suitable statistic. If the statistic $\mathbf{T}$ is such that it generates a sufficient partition of the sample space then the statistic itself is sufficient.

3.2 Example:

$\mathbf{X}=(X_1, X_2, .., X_n)$ i.i.d. Bernoulli with parameter θ, i.e.
$P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}, x_i = 0, 1$. 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.

Proof: At lecture.

Note that given the 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 $P(X = x \mid T = t)$ is a function of x and t **only**

(i.e., is **not** a function of θ). Thus once having observed the particular realization t of 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

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.

The following is a very useful criterion that helps us to check whether a statistic is sufficient or not by just looking at the joint density:

3.4 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
$L(X, \theta) = f_\theta(X_1, X_2, \ldots, X_n) = g(T(X), \theta)h(X)$(Note: X, T, θ may all be vectors, $g \geq 0, h \geq 0$).

Proof: at the lecture.

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, applying the factorization criterion, we have:

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

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 one will be looking at 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.

Here we summarize some 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 Examples

At lecture.

i) Bernoulli with probability of success $\theta \in (0,1)$. Sufficient statistic: $T = \sum_{i=1}^{n} X_i$.

ii) Univariate normal distribution with unknown μ and σ^2; Sufficient statistic for $\theta = (\mu, \sigma^2)'$: with two components $T_1 = \bar{X}, T_2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

iii)i.i.d. uniform in $(0,\theta)$: $f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta), x > 0, \theta > 0$. Sufficient statistic is $X_{(n)}-$ the maximal of the n observations.

iv) multivariate normal with mean vector μ and a covariance matrix Σ. Sufficient statistic: the vector $\bar{X}$ and the matrix $\sum_{i=1}^{n}(X_1 - \bar{X})(X_i - \bar{X})'$.

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. Yet to find a general method for constructing a minimal sufficient statistic is a difficult task.

Consider a partition $\mathcal{A}$ of $\mathcal{X}$ by defining for any $x \in \mathcal{X}$: $A(x) = \{y : \frac{L(y,\theta)}{L(x,\theta)}$ does not depend on θ, i.e. is a function of the type $h(y,x)\}$. (Note that here we do not define explicitly the statistic that produces this partition, but in particular cases we shall always try to find a statistic whose contours produce the partition).

Theorem 3.1. *(Lehmann- Scheffe's method) The above defined sets $\{A(x), x \in \mathcal{X}\}$ indeed form a partition of $\mathcal{X}$ and this partition is minimal sufficient.*

Proof: (This proof is included for completeness only and for students who are curious to see how it really works. However the details will NOT be discussed at the lecture and are NOT required! Hence you can as well safely skip this proof.)

Proof (for simplicity- the discrete case only).

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. Indeed, this is true:

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

ii) Then $L(z,\theta)/L(x_0,\theta) = \frac{L(z,\theta)}{L(x,\theta)} / \frac{L(x_0,\theta)}{L(x,\theta)}$ is *not* a function of θ because each of the two ratios on the RHS are not.

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

iv) But then $L(x_0,\theta)/L(u,\theta) = \frac{L(x_0,\theta)}{L(x,\theta)} \cdot \frac{L(x,\theta)}{L(z,\theta)} \cdot \frac{L(z,\theta)}{L(u,\theta)}$ is *not* a function of θ since the RHS is *not*.

The conclusion from this chain of statements is that an *arbitrary* $x_0 \in A(x)$ belongs to $A(u)$, too.

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 $P(Y = y \mid Y \in A(x))$:

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

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, i.e. $\mathcal{A}$ 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. $L(z, \theta)/L(y, \theta)$ 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 Examples

(at lecture).

i) Let $X_i, i = 1, 2, \ldots, n$ be i.i.d. Bernoulli with parameter $\theta \in [0, 1]$ as a probability of success. Consider the $n-$tupels $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$. Then

$$\frac{L(Y, \theta)}{L(X, \theta)} = \left\{ \frac{\theta}{1 - \theta} \right\}^{\sum_{i=1}^n Y_i - \sum_{i=1}^n X_i} .$$

Hence, given $x = (x_1, \ldots, x_n)$, the sets in the minimal sufficient partition $A(x)$ are given as $A(x) = \{y = (y_1, \ldots, y_n) : \sum_{i=1}^n x_i = \sum_{i=1}^n y_i\}$. Hence $T = \sum_{i=1}^n X_i$ is minimal sufficient for θ.

Note: Of course, in this simple example we could have argued that $T = \sum_{i=1}^n X_i$ must be minimal sufficient simply by dimension considerations (*we know that T is sufficient and is one-dimensional and you cannot further reduce the dimension that is already equal to one*). However we went directly through the original definition of minimal sufficiency, as well, to confirm out findings.

ii) i.i.d.normal with unknown μ and σ^2 : $(N(\mu, \sigma^2))$. Minimal sufficient statistic for $\theta = (\mu, \sigma^2)'$ is the vector statistic $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)'$.;

iii) i.i.d. uniform in $(0,\theta)$: $f(x,\theta) = \frac{1}{\theta} I_{(x,\infty)}(\theta), x > 0, \theta > 0$. Minimal sufficient statistic is $X_{(n)}-$ the maximal of the n observations.

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. The minimal sufficient statistics is the vector of ordered observations

$$T = (X_{(1)}, X_{(2)}, \ldots, X_{(n)})'$$

and its dimension n cannot be reduced any further.

3.8 A very important general example: 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. Obviously, in this case we have:

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

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

Note: Quite a lot of the standard distributions considered in your undergraduate courses can be seen to belong to the one parameter exponential family. Try to convince yourself that each of the following distributions is such:

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

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

3.9 Generalization to a k- parameter exponential family)

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

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

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 requested that the $k \times k$ matrix of partial derivatives $\left\{\frac{\partial c_j}{\partial \theta_{j'}}\right\}, j = 1,\ldots,k; j' = 1,\ldots,k$ has a non-zero determinant. Minimal sufficient (vector) statistic: $T = (\sum_{i=1}^n d_1(X_i),\ldots,\sum_{i=1}^n d_k(X_i))'$.

Example of a two-parameter exponential family:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{\mu^2}{2\sigma^2})\exp(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2}).$$

We have $d_1(x) = x, d_2(x) = x^2$ and $\bar{T} = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)'$ is minimal sufficient for $\theta = (\mu,\sigma^2)'$.

More details, examples and discussions-at lecture.

3.10 Ancillary Statistic. Ancillarity principle

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

Definition 2. A statistic is called ancillary if its distribution does not depend on θ.

Intuitively, *alone,* the knowledge of an ancillary statistic should not help in inference about θ. It is therefore even more interesting that an ancillary statistic, when used *in conjunction* with another statistic, sometimes *does help* in inference about θ. Inference for θ could be improved in general, if it is done *conditionally* on the ancillary statistic.

The ancillarity principle. The most important case where the above situation can occur is when a statistic $\mathbf{T}$ is minimal sufficient for θ but its dimension is *greater* than that of θ. Sometimes, 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* $\mathbf{T}_2$:

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

Then ancillarity principle postulates that inference about θ then should be based on $L_1(\mathbf{t}_1 \mid \mathbf{t}_2,\theta)$, ie., the inference about θ should be based on the conditional distribution of $\mathbf{T_1}$ given $\mathbf{T_2} = \mathbf{t_2}$.

Often, the precision of the inference about θ, as provided by such conditionality, varies with values of $\mathbf{T}_2$. In a sense, this is similar to the varying precision of samples of different sizes, n.

3.11 Examples

3.11.1

Assume that n i.i.d. $X_1, \ldots, X_n$ are given from the uniform in $(\theta, 1+\theta)$ distribution. You can easily show that the minimal sufficient statistic for θ is $T = (X_{(1)}, X_{(n)})$. Since any 1-to -1 transformation is also minimal sufficient then $T^* = (X_{(n)} - X_{(1)}, X_{(n)} + X_{(1)})$ is also minimal sufficient.

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. We see that

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

Therefore only the distribution of the largest and of the smallest order statistic from *uniform in (0,1)* distribution is involved with no dependence on θ whatsoever. Hence the first component $X_{(n)} - X_{(1)}$ of the minimal sufficient statistic turns out to be ancillary statistic.

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.

3.11.2

Inference in case of normal mixture. Assume that the density of Y is fiven by $f_Y(y) = \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_1}e^{-0.5(y-\mu)^2/\sigma_1^2} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma_2}e^{-0.5(y-\mu)^2/\sigma_2^2}$. 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.

3.11.3

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 $\mathbf{T}_2 = X_2 - X_1$. This statistic is ancillary. First, note that if $X \sim F_\theta$, then $X - \theta \sim F_0$ since:

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

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

But $F_{\mathbf{T}_2}(y, \theta) = P_\theta(\mathbf{T}_2 < y) = P\{[(X_2 - \theta) - (X_1 - \theta)] < y\}$ and the latter expression obviously does not depend on θ. Hence $\mathbf{T}_2$ is ancillary.

Along the same lines, $\tilde{T_2} = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ is also ancillary.

Definition 3. 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 illustrated in many statements about efficiency of estimators based on conditional inference. We will only formulate one of these famous results (due to an Australian Statistician).

Theorem 3.2. *If $\tilde{\theta}$ is* **any** *equivariant estimator with $E_0\tilde{\theta} < \infty$ then the estimator*

$$\hat{\theta}_P = \tilde{\theta} - E_0(\tilde{\theta}|\tilde{T_2}), \tilde{T_2} = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$$

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. We see that the Pitman estimator has a form of a Bayes estimator with respect to an (*improper*) prior on $(-\infty, \infty)$.

Exercise: Show that when $X_i, i = 1, 2, \ldots, n$ are in addition normally distributed, the Pitman estimator coincides with $\bar{X}$.

3.12 Maximum Likelihood Inference

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. Not that if $T(\mathbf{X})$ is sufficient for θ then $L(\mathbf{x},\theta) = g(T(\mathbf{x}),\theta)h(\mathbf{x})$ holds (Factorization criterion) and thus the *maximum likelihood estimator* $\hat{\theta}$(that maximizes L or, equivalently, g w.r. θ) will be a function of every sufficient statistic. In particular, the Maximum Likelihood Estimator will be a function of the *minimal sufficient statistic* when the latter exists.

Let us remember now that 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. One could have *different sampling processes* A and B that lead to likelihood functions $L_A(\mathbf{x},\theta)$ and $L_B(\mathbf{y},\theta)$. As long as $\frac{L_A(\mathbf{x},\theta)}{L_B(\mathbf{y},\theta)}$ does not depend on θ, inference about θ should be the same.

3.12.2 Example

i) In an experiment A, we observe $\mathbf{x} = (x_1, x_2, .., x_n) : n$ 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 $\mathbf{x}$.

ii) In an experiment B, we only observe one realization $\mathbf{y}$ of a single random variable $\mathbf{Y}$ =number of successes in n i.i.d. Bernoulli trials. Then $L_B(\mathbf{y},\theta) = \begin{pmatrix} n \\ k \end{pmatrix} \theta^k(1-\theta)^{n-k}$ if it happened that $\mathbf{y}$ =k.

iii) In an experiment C we observe realization of a random variable $\mathbf{Z}$ - number of trials until k successes occurred. It is known that $P_\theta(\mathbf{Z} = z) = \begin{pmatrix} z-1 \\ k-1 \end{pmatrix} \theta^k(1-\theta)^{z-k}, z = k, k+1, \ldots$ 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 considered 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-introduction

We have discussed some important principles: *sufficiency, ancillarity, weak likelihood principle* on which inference should be based. Each of these looks reasonable as a principle but it does not give us a *constructive procedure* for finding reasonable estimators of the

parameter of interest. Such well known procedure is the Maximum Likelihood Estimation Method. It is defined as $\hat{\theta} = \arg[\sup_{\theta\in\Theta} L(\mathbf{x},\theta)]$. 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 quantity $R(\mathbf{x},\theta) = \frac{L(\mathbf{x},\theta)}{L(\mathbf{x},\hat{\theta})}$ which has a range of [0,1]. This quantity is called **normed likelihood.** For a fixed $\mathbf{x}$, it is just a function of θ and we shall sometimes denote it by $R(\theta)$. For example, 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 $\hat{\theta} = .32$ and $R(\theta) = \frac{\theta^{32}}{.32^{32}}.\frac{(1-\theta)^{68}}{.68^{68}}$. An even more often used measure is the *deviance* $D(\theta)$ which is defined as $D(\theta) = -2\text{lnR}(\theta) = -2[\text{lnL}(\mathbf{x},\theta)\text{-lnL}(\mathbf{x},\hat{\theta})]$. The deviance is a non-negative number. The larger the deviance, the further the model under consideration from the most likely model, in the set under study, given the observed data. This observation can be used to construct confidence intervals for the parameter. We shall come back to this later.

3.14 Information and Likelihood

Now, 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. Having done this in a proper way, we will be able to demonstarte quantitatively (with numbers) that indeed when we are using 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 a sufficient statistic, we are losing some of the information that is contained in the whole sample.

3.14.1 Score function

We define $V(\mathbf{X},\theta) = \frac{\partial}{\partial\theta}\text{logL}(\mathbf{X},\theta)$ to be the *score function*. Generally speaking, it can be defined in the above way even for non-i.i.d. random variables where $L(\mathbf{X},\theta)$ is the joint density. In the case where $\mathbf{X} = (X_1, X_2, .., X_n)$ and the X_i are i.i.d. with a density $f(x,\theta)$ then $V(\mathbf{X},\theta) = \sum_{i=1}^{n} \frac{(\partial/\partial\theta)f(X_i,\theta)}{f(X_i,\theta)}$. Note also that obviously, for the maximum likelihood estimator (MLE)$\hat{\theta}$: $V(\mathbf{x},\hat{\theta}) = 0$ holds. Also, the property $E_\theta(V(\mathbf{X},\theta)) = 0$ holds under suitable regularity conditions (proof: at lecture).

3.14.2 Expected Fisher Information about θ contained in the vector X

It is denoted by $I_{\mathbf{X}}(\theta)$ and is defined as $I_{\mathbf{X}}(\theta)$ =$\text{Var}_\theta(V(\mathbf{X},\theta)) = E_\theta\{\frac{\partial}{\partial\theta}\text{lnL}(\mathbf{X},\theta)\}^2$(where we utilized 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 $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n})$ then

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

ii) If $T(X)$ is sufficient for θ then $I_T(\theta) = I_X(\theta)$

iii) Under regularity conditions: $I_X(\theta) = -E(\frac{\partial^2}{\partial\theta^2} \ln L(X, \theta))$

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:

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

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) (for the discrete case) First, let us note that because of the sufficiency,

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

holds and hence $E_\theta[\frac{\partial}{\partial\theta} \log f_T(T; \theta)]^2 = E_\theta[\frac{\partial}{\partial\theta} \log g_T(T; \theta)]^2$ holds. Then

$$I_T(\theta) = E_\theta[\frac{\partial}{\partial\theta} \log f_T(T; \theta)]^2 = E_\theta[\frac{\partial}{\partial\theta} \log g(T; \theta)]^2 =$$

$$E_\theta[\frac{\partial}{\partial\theta}(\log g(T; \theta) + \log h(X))]^2 = E_\theta[\frac{\partial}{\partial\theta} \log L(X; \theta)]^2 = I_X(\theta).$$

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

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

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

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

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

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

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

iv) To show this property, we need two properties of conditional expected values which we now state first. 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, we know that:

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

If we were able to show in (3) that

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

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

$$E[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}\log f_T(T,\theta)] = E_T[\frac{\partial}{\partial\theta}\log f_T(t,\theta)E(\frac{\partial}{\partial\theta}\log L(X,\theta)|T = t)] \tag{5}$$

Try to show now as an exercise that $E(\frac{\partial}{\partial\theta}\log L(X,\theta)|T = t) = \frac{\partial}{\partial\theta}\log f_T(t,\theta)$ holds. Then substitution in (5) shows that (4) holds.

We can also see from the derivations that the only way in which we may end up with a true equality $I_T(\theta) = I_X(\theta)$ is if we had equality in (4). But this is only possible if $\frac{\partial}{\partial\theta}\log L(X,\theta) = \frac{\partial}{\partial\theta}\log f_T(T,\theta)$ holds. In turn, this means that the difference $\log L(X,\theta) - \log f_T(T,\theta)$ does **not** depend on θ. If we denote this difference by $\log h(X)$, say, then we see that $\log L(X,\theta) = \log f_T(T,\theta) + \log h(X)$ holds. This also means that $L(X,\theta)$ can be factorized as in the Neyman Fisher criterion and T must be sufficient. That is, the only way in which the information in the statistic T can equal the information in the whole sample is if T is sufficient.