Lecture 2
Some Principles in Statistical Inference

5. Data Reduction in Statistical 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 $x_1, x_2, .., x_n$ to make inference about θ. His 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. The statistic summarizes the data in that, rather than reporting the entire sample $\mathbf{x}$, it reports only that $\mathbf{T}(\mathbf{x})$=$\mathbf{t}$**.** The ultimate goal in the data reduction is, when only using the value of the statistic $\mathbf{T}(\mathbf{x})$ 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 then and, in particular, we would be inclined to 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. In such a way we arrive at the definition of sufficiency.

Definition 5.1.(sufficient statistic)

We say that T is an $m-$dimensional sufficient statistic for the parameter θ of the family $\mathcal{P} = \{\mathcal{P}_\theta(\mathbf{X})\}$ if the conditional distribution of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ given $\mathbf{T} = \mathbf{t}$ does not depend on θ for all values of $\mathbf{t}$.

5.2. 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 statistic $T(X) = \sum_{i=1}^n X_i$ is sufficient for θ.

Proof: Since T is the sum of n i.i.d. Bernoulli random variables, $T \sim Bin(n, \theta)$. The possible realizations of T are $t = 0, 1, \ldots, n$ and

$$P(T = t) = \begin{pmatrix} n \\ t \end{pmatrix} \theta^t (1-\theta)^{n-t}, t = 0, 1, \ldots, n.$$

It holds:

$$P(X = x | \sum_{i=1}^n X_i = t) = \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, \sum_{i=1}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)}$$

If $\sum_{i=1}^n x_i \neq t$ then $P(X = x | \sum_{i=1}^n X_i = t) = 0$. On the other hand, if it happens that $\sum_{i=1}^n x_i = t$ then

$$P(X = x | \sum_{i=1}^n X_i = t) = \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}}{\begin{pmatrix} n \\ t \end{pmatrix} \theta^t (1-\theta)^{n-t}} = \frac{1}{\begin{pmatrix} n \\ t \end{pmatrix}}.$$

In both cases, the conditional probability $P(X = x | \sum_{i=1}^n X_i = t)$ does not depend on θ.$\blacksquare$

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:**

5.3. Definition (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:

5.4. 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 if and only if
$L(X, \theta) = f_\theta(X_1, X_2, \ldots, X_n) = g(T(X), \theta)h(X)$

(Note: X, T, θ may all be vectors, $g \geq 0, h \geq 0$).

Proof: at lecture.

Examples : at lecture.

i) Bernoulli; (Show that if $\sum_{i=1}^n x_i = t$ then $L(x, \theta) = \theta^t(1-\theta)^{n-t}$. Hence $T = \sum_{i=1}^n X_i$ can be taken.)

ii) univariate normal distribution with unknown mean μ and variance σ^2; (Show that $L(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]\}$ and hence T can be taken to be a vector statistic with 2 components: $T_1 = \sum_{i=1}^n X_i, T_2 = \sum_{i=1}^n (X_i - \bar{X})^2$.)

iii) uniform distribution in $[0, \theta)$; (Show that if $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are the order statistics then $L(x; \theta) = \frac{1}{\theta^n} I_{(x_{(n)}, \infty)}(\theta)$ and hence $T = X_{(n)}$ can be chosen.)

iv) multivariate normal. (If $X_i \sim N(\mu, C)$ with $\mu \in R^p$, and C being a positive definite $p \times p$ matrix, show that

$$L(x; \mu, C) = (2\pi)^{-np/2}|C|^{-n/2} \exp\{-\frac{1}{2} tr[C^{-1}(\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})' + n(\bar{x} - \mu)(\bar{x} - \mu)')]\}$$

where tr denotes the trace operator. Hence T can be taken to be with two components: the vector $\bar{X} \in R^p$ and the matrix $\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$.)

5.5. Minimal sufficient statistic. 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.

Using sufficient statistic, we want to "condense" the data without loosing any information about the parameter. However a sufficient statistic is not uniquely defined. For example, in sampling from a normal distribution with both μ and σ^2 unknown, the sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ itself, the ordered sample $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$, and the pair

$(\bar{X}, s^2)$ are sufficient statistics for $\theta = (\mu, \sigma^2)'$. We naturally prefer the pair $(\bar{X}, s^2)$ since it condenses the data more than either of the other two. We are always looking at a statistic that allows **the greatest data reduction without loss of information** on θ. From the above, we see that a statistic that allows the greatest data reduction will be a function of any other sufficient statistics. Such a statistic is called the **minimal sufficient** statistic.

Before we move on further we would like to summarize some properties of sufficient statistics that we discussed until now:

i) If T is sufficient, so is any one-to-one function of T;

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. In fact, it divides the sample space $\mathcal{X}$ into *disjoint* subsets A_t, each containing all possible observations $\mathbf{x}$ with the same value $\mathbf{t}$.

5.6. Lehmann and Scheffe′s method for finding a minimal sufficient statistic

Very 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. In the following theorem, an easy-to-use criterion for a minimal sufficent statistic will be formulated:

Theorem (simplified version of Theorem by Lehmann- Scheffe) Suppose there exists a function $\mathbf{T}(\mathbf{x})$ such that, for two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $\frac{L(\mathbf{x},\theta)}{L(\mathbf{y},\theta)}$ is constant as a function of θ if and only if $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$. Then $\mathbf{T}(\mathbf{X})$ is a minimal sufficient statistic for θ.

5.7. Examples (at lecture)

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, and could even equal the sample size n itself.

5.8.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:

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

which is *not* a function of θ if and only if $\sum_{i=1}^{n} d(x_i) = \sum_{i=1}^{n} d(y_i)$. Hence, $T = \sum_{i=1}^{n} d(X_i)$ is minimal sufficient.

Note: Quite a lot of the standard distributions considered in your introductory Statistics 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.

5.9. Generalization of 5.8.(k- parameter exponential family) (at lecture).

6. Maximum Likelihood Inference

6.1. 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. Note 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 for a minimal sufficient statistic $\mathbf{T}(\mathbf{X})$ we have $L(\mathbf{y}, \theta) = h(\mathbf{y}, \mathbf{x})L(\mathbf{x}, \theta)$, which means that the values $\mathbf{x}$ and $\mathbf{y}$ for which $T(x) = T(y)$ must lead to the same inference about θ. An even stronger version of this requirement is the *likelihood principle:*

"Data sets with proportional likelihood functions should 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.

6.2. 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 $L_B(\mathbf{y}, \theta) = \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k}$ 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 $P_\theta(\mathbf{Z} = z) = \left(\begin{array}{c} z-1 \\ k-1 \end{array} \right) \theta^k (1-\theta)^{z-k}, z = k, k+1, \ldots$ Here the number of trials is random but if it happened that $z = n$ then $L_C(n, \theta) = \left(\begin{array}{c} n-1 \\ k-1 \end{array} \right) \theta^k (1-\theta)^{n-k}$.

In all three cases considered above, if the values of k and n were the same, the corresponding likelihood funcitons $L_A(x, \theta), L_B(y, \theta)$ and $L_C(n, \theta)$ would be proportional. Then, according to the likelihood principle, for the specific realizations of the variables $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ that lead to the same k and n values, the conclusions about θ in all these three cases must be identical.