University of California, Los Angeles
Department of Statistics

Statistics 100B **Instructor: Nicolas Christou**

Data reduction and the sufficiency principle

Introduction

We have seen some examples of estimators, some based on intuition or the method of maximum likelihood or the method of moments. For example $\bar{X}$ is an unbiased estimator of μ, S^2 is an unbiased estimator of σ^2 etc. Are these estimators sufficient, in the sense that we no longer need the actual data values? Such estimators exist and are called sufficient statistics. An important result is that sufficient statistics can be used to find minimum variance unbiased estimators.

Sufficient statistics

Definition: A statistic $T(\mathbf{X})$ is a sufficient statistic for a parameter θ if the conditional distribution of $\mathbf{X}$ given the value $T(\mathbf{X})$ does not depend on θ.

Theorem

If $p(\mathbf{X}|\theta)$ is the joint probability density or joint probability mass function of $\mathbf{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, we say that $T(\mathbf{X})$ is a sufficient statistic of θ if the ratio $\frac{p(\mathbf{X}|\theta)}{q(T(\mathbf{X})|\theta)}$ is constant as a function of θ.

Example 1

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli random variables with parameter p. Show that $T(\mathbf{X}) = X_1 + \ldots + X_n$ is a sufficient statistic for p.

Example 2

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma)$ random variables with σ^2 known. Show that $\bar{X}$ is a sufficient statistic of μ.

Factorization theorem
Let L be the joint pdf or pmf of a random sample. Then U is a sufficient statistic for the estimation of a parameter θ iff the likelihood function can be expressed as the product of two nonnegative functions:

$$L = g(u, \theta)h(\mathbf{X}).$$

Example 2 (Revisit)

Example
Let $X_1, \ldots, X_n$ be i.i.d. with $X_i \sim exp(\frac{1}{\lambda})$. Show that $\bar{X}$ is a sufficient statistic for the estimation of λ.

Example
Let $X_1, \dots, X_n$ be i.i.d. $N(\mu, \sigma)$, with both μ and σ^2 unknown. Show that $(\bar{X}, S^2)$ is a sufficient statistic for (μ, σ^2).

Theorem
Let $X_1, \ldots, X_n$ be i.i.d. random variables from a pdf or pmf that belongs in an exponential family:

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right).$$

Then $T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$ is a sufficient statistic of $\boldsymbol{\theta}$.

Minimal sufficiency and MVUE

Find a sufficient statistic that reduces the data as much as possible. This statistic is called minimal sufficient statistic and can be used to find minimum variance unbiased estimators (MVUE) for a parameter θ. How do we find minimal sufficient statistics?

Definition

A sufficient statistic $T(\mathbf{X})$ is called minimal sufficient statistic if, for an other competitor sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$.

Example

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma)$ random variables with σ^2 known. Show that $T(\mathbf{X}) = \bar{X}$ achieves a greater data reduction than $T'(\mathbf{X}) = (\bar{X}, S^2)$.

Theorem - Lehmann and Scheffé (A method for deriving minimal sufficient statistics)
Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be two samples from the same pdf or pmf. Then, if $\frac{L(x_1,\ldots,x_n|\theta)}{L(y_1,\ldots,y_n|\theta)}$ is free of the unknown parameter θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$ then $T(\mathbf{X})$ is a minimal sufficient statistic of the unknown parameter θ.

In addition: If $T(\mathbf{X})$ is an unbiased estimator of a parameter θ and it is a function of a minimal sufficient statistic it will be a minimum variance unbiased estimator (MVUE). (It will have the smallest possible variance among all the unbiased estimators of θ).

Example 1
Let $Y_1, \ldots, Y_n$ be a random sample from the Bernoulli distribution with parameter p. Find the minimal sufficient statistic for p and use it to find an MVUE of p.

Example 2

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables from the Weibull distribution $f(y|\theta) = (\frac{2y}{\theta})exp(-\frac{y^2}{\theta}), y > 0$. Find the minimum variance unbiased estimator for θ.

Example 3

Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma)$, with both μ and σ^2 unknown. Find the minimum variance unbiased estimators for (μ, σ^2).

Example 4
Let $Y_1, \ldots, Y_n$ be i.i.d. random variables from the exponential density function with parameter $\frac{1}{\lambda}$. Find the MVUE of $var(Y_i)$.

Theorem.
Minimal sufficient statistics in the exponential family case:
Let $X_1, \ldots, X_n$ be i.i.d. random variables from a pdf or pmf that belongs in an exponential family:

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right).$$

We say that $T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$ is a complete statistic of $\boldsymbol{\theta}$ and therefore minimal sufficient statistic.

Definition
A statistic $S(\mathbf{X})$ is called ancillary if its distribution does not depend on θ.

Rao-Blackwell theorem - Minimum variance unbiased estimators
Let $\hat{\theta}$ be an unbiased estimator for θ. If U is a sufficient statistics for θ, define $\hat{\theta}^* = E(\hat{\theta}|U)$. Then, $E(\hat{\theta}^*) = \theta$ and $var(\hat{\theta}^*) \leq var(\hat{\theta})$.