STAT 135, Concepts of Statistics

Helmut Pitters

Sufficiency

Department of Statistics
University of California, Berkeley

February 23, 2017

Sufficiency.

Example (Coin tossing)

Flip coin 10 times and record pattern HHTHHHTHTT.

1. Natural guess for probability p for heads?

$$\frac{\text{\# of heads in HHTHHHTHTT}}{10}?$$

2. Imagine we throw the coin 10^6 times

$$\text{HTTHHHHHHTTT}\cdots$$

Pointless to analyze details of corresponding pattern of heads and tails.
To estimate p, seems sufficient to know number (statistic)

$$h(\text{HTTHHHHHHTTT}\cdots)$$

of heads observed. $h(\cdots)$ is said to be a *sufficient statistic* for p.

Sufficiency.

From sample

$$(X_1, X_2, \ldots, X_n) \sim \mathbb{P}_\theta$$

want to learn θ.

If sample size n is large, may be hard to interpret list of numbers $x_1, x_2, \ldots, x_n$. Instead, might be enough to consider some key features, e.g.

mean, standard deviation, $x_{(1)} = \min_i x_i$, $x_{(n)} = \max_i x_i$,

etc. that are functions of the data ("statistics" in statistical jargon).

Statistics reduce/compress the data.

Natural questions:

- How can we compress data without compromising quality of inference?
- Is there an "optimal" method to compress? If so, how can we find it?

Sufficiency.

More generally: A statistic $T(X_1, \dots, X_n)$ is called *sufficient for* θ if any inference about θ depends on $X_1, \dots, X_n$ only via $T(X_1, \dots, X_n)$.

Definition (Sufficient statistic)

Statistic $T(X_1, \dots, X_n)$ is called *sufficient statistic* for θ if conditional distribution of $X_1, \dots, X_n$ given $T = t$, i.e.

$$\mathbb{P}_\theta\{X_1 \in \cdot, \dots, X_n \in \cdot | T = t\}$$

does not depend on θ for any value of t.

In other words: Inference of θ is not improved by gaining more information about $X_1, \dots, X_n$ than is contained in $T(X_1, \dots, X_n)$.

Sufficiency.

Example (Coin tossing)

Consider again n independent tosses of a coin that shows up heads w.p. p. Let

$$X_i := \begin{cases} 1 & \text{coin shows heads in } i\text{th toss} \\ 0 & \text{otherwise.} \end{cases}$$

Argued earlier that, intuitively,

$$H := H(X_1, \ldots, X_n) := \sum_{i=1}^{n} X_i = \#\text{ of heads}$$

should be sufficient statistic for p.

Sufficiency.

Example (Coin tossing)

Does H satisfy definition of sufficiency?

$$\mathbb{P}\left\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | H = h\right\}$$

$$= \frac{\mathbb{P}\left\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right\}}{\mathbb{P}\left\{H = h\right\}} = \frac{p^h(1-p)^{n-h}}{\binom{n}{h}p^h(1-p)^{n-h}} = \binom{n}{h}^{-1}$$

does not depend on p, therefore H is sufficient stat. for p.

Remark

Notice that sufficient statistic need not be unique, e.g. statistic $2H$ would do just as well as H.

Sufficiency.

Theorem (Factorization theorem)

Statistic T is sufficient for θ, if and only if $f(x|\theta)$ can be written as

$$f(x|\theta) = g(T(x), \theta)h(x). \tag{1}$$

Remark

Recall that MLE for θ is the value $\hat{\theta}$ that maximizes $f(x|\theta)$.

Suppose T is sufficient for θ. Because of the factorization theorem, $\hat{\theta}$ maximizes $f(x|\theta)$ if and only if it maximizes $g(T(x), \theta)$, in other words, MLE is a function of the sufficient statistic $T(X)$.

Sufficiency.

Proof of factorization theorem.

We prove this theorem only for the discrete case. Suppose $f(x|\theta) = \mathbb{P}_\theta\{X = x\}$ satisfies above factorization and $T(X) = t$. Then

$$\mathbb{P}_\theta\{X = x|T(X) = t\} = \frac{\mathbb{P}_\theta\{X = x\}}{\mathbb{P}_\theta\{T(X) = t\}}$$
$$= \frac{g(T(x), \theta)h(x)}{\sum_{x:\, T(x)=t} g(T(x), \theta)h(x)} = \frac{g(t, \theta)h(x)}{\sum_{x:\, T(x)=t} g(t, \theta)h(x)}$$
$$= \frac{h(x)}{\sum_{x:\, T(x)=t} h(x)},$$ and this quantity does not depend on θ.

□

Sufficiency.

Proof.

Suppose now that T is sufficient and $T(X) = t$. Then

$$\mathbb{P}_\theta\{X = x\} = \mathbb{P}_\theta\{X = x | T(X) = t\}\mathbb{P}_\theta\{T(X) = t\},$$

where the first factor does not depend on θ (by sufficiency), hence factorization is given by

$$h(x) \coloneqq \mathbb{P}_\theta\{X = x | T(X) = t\}$$

and

$$g(T(X), \theta) \coloneqq \mathbb{P}_\theta\{T(X) = t\}.$$

□

Sufficiency.

Example (uniform, one parameter). Let $X_1, \dots, X_n$ be i.i.d. with uniform distribution on $[0, \theta]$. Want to estimate unknown θ.

Write $x = (x_1, \dots, x_n)$.

$$f(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(x_i) = \theta^{-n} \mathbf{1}_{[0,\theta]}(\max_i x_i) \qquad \text{for } x_1, \dots, x_n \geq 0,$$

where

$$\mathbf{1}_A(z) := \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{otherwise} \end{cases} \quad \text{denotes the indicator of } A.$$

$T(x) := \max_i x_i$ is sufficient statistic for θ, since

$$g(T(x), \theta) := \theta^{-n} \mathbf{1}_{[0,\theta]}(T(x))$$
$$h(x) := 1.$$

Moreover, $\max_i x_i$ is the MLE for θ, since it maximizes $f(x|\theta)$.

Sufficiency.

Example (Poisson). Let $X_1, \ldots, X_n$ be i.i.d. with $\mathrm{Poisson}(\lambda)$ distribution. Want to estimate unknown λ.

$$f(x|\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_i x_i}}{\prod_i x_i!} = g(T(x), \lambda) h(x),$$

where

$$\begin{aligned} T(x) &:= \sum_i x_i \\ g(T(x), \lambda) &:= e^{-\lambda n} \lambda^{T(x)} \\ h(x) &:= \frac{1}{\prod_i x_i!}. \end{aligned}$$

By the factorization theorem, $\sum_i X_i$ is a sufficient statistic for λ.

Sufficiency.

Definition

If X is an estimator for θ, its *mean squared error* is defined by

$$\boxed{\mathrm{MSE}(X) := \mathbb{E}(X-\theta)^2}$$

and often used to measure the accuracy of an estimate.

If $\mathbb{E}(X-\theta)^2 < \infty$ we have

$$\mathbb{E}(X-\theta)^2 = \mathrm{Var}(X) + b^2(\theta, X),$$

where

$$b(\theta, X) := \mathbb{E}[X] - \theta$$

is the *bias* of X.

Sufficiency.

The next theorem shows that if we look for an estimator with small MSE, it is enough to consider estimators that are functions of sufficient statistics.

Theorem (Rao-Blackwell)

Let $\hat{\theta}$ be an estimator of θ such that $\mathbb{E}\hat{\theta}^2 = \mathbb{E}_\theta\hat{\theta}^2 < \infty$ for all θ. Suppose that T is sufficient for θ, and define $\tilde{\theta} := \mathbb{E}[\hat{\theta}|T]$ to be the conditional expectation of $\tilde{\theta}$ given T. Then, for all θ

$$\mathrm{MSE}(\tilde{\theta}) = \mathbb{E}(\tilde{\theta} - \theta)^2 \leq \mathbb{E}(\hat{\theta} - \theta)^2 = \mathrm{MSE}(\hat{\theta}).$$

The inequality is strict unless $\tilde{\theta} = \hat{\theta}$.

Sufficiency.

Proof.

(of Rao-Blackwell thm.) From the tower property of conditional expectation (i.e. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$),

$$\mathbb{E}\hat{\theta} = \mathbb{E}[\mathbb{E}[\hat{\theta}|T]] = \mathbb{E}\tilde{\theta}.$$

Consequently, $\tilde{\theta}$ and $\hat{\theta}$, have the same bias, and

$$\mathrm{MSE}(\tilde{\theta}) - \mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\tilde{\theta}) - \mathrm{Var}(\hat{\theta}).$$

Recall the conditional variance formula

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(\mathbb{E}[\hat{\theta}|T]) + \mathbb{E}[\mathrm{Var}(\hat{\theta}|T)] \\ &= \mathrm{Var}(\tilde{\theta}) + \mathbb{E}[\mathrm{Var}(\tilde{\theta}|T)] \geq \mathrm{Var}(\hat{\theta}),\end{aligned}$$

with equality if and only if $\mathrm{Var}(\tilde{\theta}) = 0$, in other words, $\hat{\theta}$ is a function of T. $\square$