MATH 5010 –Foundations of Statistical Theory and Probability

❖ Sufficient Statistics

Instructor: He Wang
Department of Mathematics
Northeastern University

Sufficient Statistics

The Likelihood Principle

❖ Statistic of Random variables

Let $\{X_1, \dots, X_n\}$ be a random sample from a population**.**

Let $Y = T(X_1, \dots, X_n)$ be a **statistic** of the random sample.

We say a statistic T is an **estimator** of a population parameter if T is usually close to θ.

Examples

- The sample mean is an estimator for the population mean.
- The sample variance is an estimator for the population variation.

Heuristic definition of Sufficient Statistics

A **sufficient statistic** is a function of the data that captures *all* the information needed to estimate a parameter θ.

Once you know the sufficient statistic, **the data provides no additional information** about the parameter.

$T = T(X_1, X_2, \cdots, X_n)$ is a **sufficient statistic** if the statistician who knows the value of T can do just as good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample $\{X_1, X_2, \cdots, X_n\}$.

Mathematical definition of Sufficient Statistics

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with parameter θ.

Let $T = T(X_1, X_2, \cdots, X_n)$ be a **statistic.**

$T = T(X_1, X_2, \cdots, X_n)$ is a **sufficient for $\boldsymbol{\theta}$** if for each t, the conditional distribution of $X_1, X_2, \cdots, X_n$ given value $T = t$ does not depend on θ.

Factorization theorem

Theorem:

Let $X_1, X_2, \cdots, X_n$ be a random sample with joint density $f(x_1, \ldots, x_n|\theta)$ with parameter θ.

A statistics $T = T(X_1, X_2, \cdots, X_n)$ is a **sufficient for θ** if and only if there exist functions $g(t|\theta)$ and $h(\vec{x})$ such that for all sample points x and all parameter points θ

$$f(\vec{x}\,|\theta) = g(T(\vec{x})|\theta)\, h(\vec{x})$$

Remark:

- $g(T(\vec{x})|\theta)$ depends on the data only through $T(x)$ and on θ
- $h(\vec{x})$: depends only on the data, not on θ

Example: 1. Bernoulli Distribution

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

- The joint PMF:

$$f(x_1, \ldots, x_n \mid p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

- So, $T(X) = \sum X_i$ is a sufficient statistic for p.

Example 2. Normal Distribution with Known Variance

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, known σ^2

- The likelihood:

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- This simplifies in terms of $\sum x_i$
- So, $T(X) = \sum X_i$ or $\bar{X}$ is sufficient for μ

Example. Normal Distribution with Known Variance

Let $X_1, \ldots, X_n \sim Normal(\mu, \sigma^2)$ with known σ^2.

The joint density is

$$f(x_1, \cdots, x_n|\mu) = (2\pi)^{-n/2}\,\sigma^{-n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= (2\pi)^{-n/2}\,\sigma^{-n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2}\right)$$

Since σ^2 is known, let

$$u(x_1, \cdots, x_n) = (2\pi)^{-n/2}\,\sigma^{-n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right)$$

$$v(r(x_1, x_2, \cdots, x_n), \mu) = \exp\left(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}r(x_1, x_2, \cdots, x_n)\right)$$

By the factorization theorem this shows that $r(X_1, \ldots, X_n) = X_1 + \cdots + X_n$ is a sufficient statistics.

Example *(Uniform population)* Now suppose the X_i are uniformly distributed on $[0, \theta]$ where θ is unknown. Then the joint density is

$$f(x_1, \cdots, x_n|\theta) = \theta^{-n}\, 1(x_i \leq \theta,\, i = 1, 2, \cdots, n)$$

Here $1(E)$ is an indicator function. It is 1 if the event E holds, 0 if it does not. Now $x_i \leq \theta$ for $i = 1, 2, \cdots, n$ if and only if $\max\{x_1, x_2, \cdots, x_n\} \leq \theta$. So we have

$$f(x_1, \cdots, x_n|\theta) = \theta^{-n}\, 1(\max\{x_1, x_2, \cdots, x_n\} \leq \theta)$$

By the factorization theorem this shows that

$$T = \max\{X_1, X_2, \cdots, X_n\}$$

is a sufficient statistic.

What about the sample mean?

Is it a sufficient statistic in this example?

By the factorization theorem it is a sufficient statistic only if we can write

$$1(\max\{x_1, x_2, \cdots, x_n\} \leq \theta)$$

as a function of just the sample mean and θ.

This is impossible, so the sample mean is not a sufficient statistic in this setting.

Example *(Gamma population, α unknown, β known)* Now suppose the population has a gamma distribution and we know β but α is unknown. Then the joint density is

$$f(x_1, \cdots, x_n|\alpha) = \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \left(\prod_{i=1}^{n} x_i^{\alpha-1} \right) \exp(-\beta \sum_{i=1}^{n} x_i)$$

We can write

$$\prod_{i=1}^{n} x_i^{\alpha-1} = \exp\left((\alpha - 1) \sum_{i=1}^{n} \ln(x_i) \right)$$

By the factorization theorem this shows that

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

is a sufficient statistic. Note that $\exp(T) = \prod_{i=1}^{n} X_i$ is also a sufficient statistic. But the sample mean is not a sufficient statistic.

A **minimal sufficient statistic** is the "smallest" sufficient statistic in the sense that all other sufficient statistics are functions of it.

A statistic $T(\vec{X})$ is **complete and sufficient** if it is sufficient and has the property that:

$$E[g(T)] = 0 \; for \; all \; \theta \;\; \Rightarrow g(T) = 0 \; almost \; surely.$$

Why It Matters?

- **Data Reduction**: Simplifies estimation without losing information.
- **MLE and Bayesian Estimation** often depend only on sufficient statistics.
- **Statistical Inference** becomes more efficient and interpretable.

The Likelihood Principle

Likelihood function is a specific and important statistics used to summarize data.

Let $f(\vec{x} \mid \theta)$ be the joint pdf or pmf of the sample $\vec{X} = (X_1, \dots, X_n)$.

Given $\vec{X} = \vec{x} = (x_1, \dots, x_n)$ is observed, the function of θ defined by

$$L(\theta|\vec{x}) \coloneqq f(\vec{x} \mid \theta)$$

is called the **likelihood function.**

Remark: Likelihood function and the joint pdf are the same but by considering different variables.

Likelihood Principle: If two different experiments yield the **same likelihood function** (up to a constant),

$$L(\theta|\vec{x}) = C(\vec{x}, \vec{y})L(\theta|\vec{y}) \text{ for all } \theta$$

then:

- They provide **the same evidence** about θ.
- And all inferences about θ should be **identical**.

The Likelihood Principle says: once you've observed the data, how it was collected doesn't matter — only the likelihood matters.

It's a **foundational idea** in **Bayesian statistics**, and often **ignored in frequentist procedures**. (uses the likelihood + prior.)

Coin Tossing Example:

Scenario 1 (Binomial):

- Toss a coin $n = 10$ times.
- Observe 7 heads: $X = 7$
- Likelihood: $L(p) = \binom{10}{7} p^7 (1-p)^3$

Scenario 2 (Negative Binomial):

- Toss a coin until 7 heads appear.
- It takes 10 tosses: $T = 10$
- Likelihood: $L(p) = \binom{9}{6} p^7 (1-p)^3$

Both have likelihoods proportional to $p^7(1-p)^3$, so the **Likelihood Principle** says:

These two datasets should lead to the same inference about p,
even though they arise from different experiments.

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**
-

Online books:

https://www.probabilitycourse.com/