Math 4501 - Probability and Statistics II

6.7 - Sufficient Statistics

Sufficient Statistic

STATISTIC: function of random sample that DOES NOT depend on any unknown parameter.

Definition

Let $X_1, \ldots, X_n$ be a random sample from a distribution with pdf/pmf $f(\cdot;\theta)$ depending on a parameter θ. ← θ is the unknown parameter

A statistic $Y = u(X_1, \ldots, X_n)$ is said to be *sufficient for* θ if the conditional distribution of $X_1, \ldots, X_n$ given $Y = y$ does not depend on θ for any value y of Y.

Interpretation:

Y summarizes ALL the relevant info about θ in the random sample

- A sufficient statistic is a function of the random sample whose value contains all the information needed to compute any estimate of the parameter, i.e. there is no additional information about the unknown parameter left in the remaining (conditional) distribution.
- The joint probability distribution of the data is conditionally independent of the parameter given the value of the sufficient statistic for the parameter.

$\boxed{Y = u(X_1, \ldots, X_n)}$ is sufficient for θ if $P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y)$ does not depend on θ

Factorization Theorem

Theorem (Fisher-Neyman Factorization Theorem)

given

Let $X_1, X_2, \ldots, X_n$ denote random variables with joint pdf/pmf $f(x_1, x_2, \ldots, x_n; \theta)$ depending on the parameter θ.

The statistic $\boxed{Y = u(X_1, X_2, \ldots, X_n)}$ is sufficient for θ if and only if

$$f(x_1, x_2, \ldots, x_n; \theta) = \phi(\underbrace{u(x_1, x_2, \ldots, x_n)}_{y}; \theta)\, h(x_1, x_2, \ldots, x_n),$$

where ϕ depends on $x_1, x_2, \ldots, x_n$ only through $\underbrace{u(x_1, \ldots, x_n)}_{y}$ and $h(x_1, \ldots, x_n)$ does not depend on θ.

Note: it is often easier to check sufficiency using the Factorization Theorem than it is using the definition.

If $x_1, \ldots, x_n$ is a random sample $f_{joint}(x_1, \ldots, x_n; \theta) = \overbrace{\prod_{i=1}^{n} f(x_i, \theta)}$

Remark

If Y is sufficient for a parameter θ, then every function of Y not involving θ, but with a single-valued inverse, is also a sufficient statistic for θ.

Interpretation: knowing either Y or $W = \nu(Y)$, we know the other (provided ν^{-1} is single valued).

Reasoning: Let Y be a sufficient statistic for θ. By the factorization theorem, we have that

$$f_{joint}(x_1, \dots, x_n; \theta) = \phi(\underbrace{u(x_1, \dots, x_n)}_{Y}, \theta) \cdot h(x_1, \dots, x_n)$$

Let W be another statistic of the form $W = \nu(Y) = \nu(u(x_1, x_2, \dots, x_n))$

As long as ν has an inverse, we can write:

$$Y = \nu^{-1}(W) = \nu^{-1}(\nu(u(x_1, x_2, \dots, x_n)))$$ and so, we get

$$\Rightarrow f_{joint}(x_1, \dots, x_n, \theta) = \phi(\nu^{-1}(\underbrace{\nu(\overbrace{u(x_1, x_2, \dots, x_n)}^{Y})}_{W}), \theta) \cdot h(x_1, \dots, x_n)$$

that is, W is also a sufficient statistic (by the factorization theorem!)

EXAMPLE 1 :

Last time we've seen that $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for the parameter p of Bernoulli (p)

From the previous remark, we can conclude that

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$ is also sufficient

because $\bar{X} = \nu(Y)$ where $\nu(y) = \frac{1}{n} y$ is invertible

EXAMPLE 2:

Last time, we've also seen that $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a sufficient statistic for the parameter λ of a Poisson distribution

Using the previous remark, we can immediately conclude that

$$Y = \sum_{i=1}^{n} X_i \quad \text{is also sufficient}$$

since $\quad Y = v(\bar{X}) \quad$ where $\quad v(x) = n \cdot x \quad$ is invertible

Example

mean μ is the only unknown parameter!

$\sigma^2 = 1$ is known

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, 1)$, $\mu \in \mathbb{R}$.

Show that the sample mean $\bar{X}$ is a sufficient statistic for μ.

pdf of $N(\mu,1)$ is

$$f(x,\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

We will use the factorization theorem:

$$f_{joint}(x_1, x_2, \ldots, x_n, \mu) = \prod_{i=1}^{n} f(x_i, \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \exp\left(\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2}\right)$$

$$= (2\pi)^{-n/2} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\underbrace{(x_i - \bar{x}) + (\bar{x}}_{=0} - \mu)\right)^2\right)$$

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

$$= (2\pi)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(x_i - \bar{x})^2 - \frac{1}{2}\sum_{i=1}^{m} 2(x_i - \bar{x})(\bar{x} - \mu) - \frac{1}{2}\sum_{i=1}^{m}(\bar{x} - \mu)^2 \right)$$

$$= (2\pi)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\bar{x} - \mu)^2 - (\bar{x} - \mu)\underbrace{\sum_{i=1}^{m}(x_i - \bar{x})}_{0} - \frac{1}{2}\sum_{i=1}^{m}(x_i - \bar{x})^2 \right)$$

$$= (2\pi)^{-m/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\bar{x} - \mu)^2 - \frac{1}{2}\sum_{i=1}^{m}(x_i - \bar{x})^2 \right)$$

$$= (2\pi)^{-m/2} \underbrace{\exp\left(-\frac{1}{2}\sum_{i=1}^{m}(\bar{x} - \mu)^2 \right)}_{\phi(y,\mu),\ y = \bar{x}} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \right)$$

$\exp\left(-\frac{m}{2}(\bar{x} - \mu)^2 \right)$

$h(x_1, x_2, \ldots, x_m$

We conclude that the joint pdf of $X_1, \dots, X_n$ is of the form:

$$f_{joint}(x_1, \dots, x_n, \mu) = \phi(y, \mu) \cdot h(x_1, \dots, x_n)$$

where $\phi(y, \mu) = \exp\left(-\frac{n}{2}(y - \mu)^2\right)$ where $y = \bar{x}$

and $h(x_1, \dots, x_n) = (2\pi)^{-n/2} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$ (does not depend on μ)

Factorization theorem guarantees that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a sufficient statistic for μ

(as a consequence for any invertible function v, $W = v(\bar{X})$ is also sufficient
For instance, if we take $v(x) = n \cdot x \Rightarrow Y = \sum_{i=1}^{n} X_i$ is also sufficient)

Exponential family of distributions

→ includes nearly all of the distributions studied in 3501

Definition

A random variable X (discrete or continuous) has a distribution from the *exponential family* if its pdf/pmf may be parameterized in terms of a single parameter θ and expressed in the form

$$f(x;\theta) = \exp[K(x)p(\theta) + S(x) + q(\theta)]$$

with support not depending on θ.

simplification comes from

$X_1, \ldots, X_n$ random sample

$$f_{joint}(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = \prod_{i=1}^{n} \exp\left(K(x_i)p(\theta) + S(x_i) + q(\theta)\right)$$

$$= \exp\left(\sum_{i=1}^{n}\left(K(x_i)p(\theta) + S(x_i) + q(\theta)\right)\right)$$

EXAMPLE 1 :

Bernoulli (p) is an element of the exponential family of distributions since its pmf is

$$f(x) = p^{x}(1-p)^{1-x} \quad , \quad x = 0,1$$

(discrete distribution)

may be written as:

$$f(x) = p^{x}(1-p)^{1-x} = \underbrace{e^{x \ln p}}_{p^{x}} \cdot \underbrace{e^{(1-x)\ln(1-p)}}_{(1-p)^{1-x}}$$

$x = e^{\ln x}$
for $x > 0$

$$= e^{x \ln p + (1-x)\ln(1-p)} = e^{x \ln p + \ln(1-p) - x \ln(1-p)}$$

$$= \exp\left(x(\ln p - \ln(1-p)) + \ln(1-p) \right) = \exp\left(\underbrace{x}_{K(x)} \cdot \underbrace{\ln\left(\frac{p}{1-p}\right)}_{P(p)} + \underbrace{\ln(1-p)}_{Q(p)} \right)$$

is of the form $\exp\left(K(x) \cdot P(p) + S(x) + Q(p) \right)$ where $K(x) = x$, $P(p) = \ln\left(\frac{p}{1-p}\right)$, $S(x) = 0$, $Q(p) = \ln(1-p)$

EXAMPLE 2: Poisson (λ) is also an element of the exponential family of distributions

since its pmf $\boxed{f(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}}$, $x = 0, 1, 2, \ldots.$, $\underline{\underline{\lambda > 0}}$

$$\boxed{x = e^{\ln x} \quad \text{for } x > 0}$$

since

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{e^{x \ln \lambda} \cdot e^{-\lambda}}{e^{\ln(x!)}} = e^{x \ln \lambda - \lambda - \ln x!}$$

$$= \exp\left(x \cdot \ln \lambda \underbrace{- \ln(x!)} \underbrace{- \lambda}\right)$$

is of the form $\exp\left(K(x) \cdot p(\lambda) + S(x) + q(\lambda)\right)$

with $K(x) = x$, $p(\lambda) = \ln \lambda$, $S(x) = -\ln(x!)$, $q(\lambda) = -\lambda$

EXAMPLE 3 : $N(\mu, 1)$ is also an element of the exponential family since its pdf $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right)$, $x \in \mathbb{R}$

may be written as:

$$f(x) = \exp\left(-\ln\sqrt{2\pi}\right) \cdot \exp\left(-\frac{1}{2}\left(x^2 - 2\mu x + \mu^2\right)\right)$$

$$= \exp\left(-\frac{1}{2}\ln(2\pi)\right) \cdot \exp\left(-\frac{1}{2}x^2 + x\mu - \frac{1}{2}\mu^2\right)$$

$$= \exp\left(x \cdot \mu \quad -\frac{1}{2}x^2 \quad -\frac{1}{2}\mu^2 - \frac{1}{2}\ln 2\pi\right)$$

which is of the form $\exp\left(K(x) \cdot p(\mu) + S(x) + q(\mu)\right)$

with $K(x) = x$, $p(\mu) = \mu$, $S(x) = -\frac{1}{2}x^2$ and $q(\mu) = -\frac{1}{2}\mu^2 - \frac{1}{2}\ln 2\pi$

Example (all of the previous 3 examples put together)

$Y = \sum_{i=1}^{n} X_i$

1) The Bernoulli distribution is an element of the exponential family since its pmf is of the form

$K(x) = x$

$$p^x(1-p)^{1-x} = \exp\left\{ x \ln\left(\frac{p}{1-p}\right) + \ln(1-p) \right\} , \quad x = 0, 1 .$$

$\overline{X}$

$Y = \sum_{i=1}^{n} X_i$

2) The Poisson distribution is an element of the exponential family since its pmf is of the form

$K(x) = x$

$$\frac{e^{-\lambda}\lambda^x}{x!} = \exp\{ x \ln\lambda - \ln x! - \lambda \} , \quad x = 0, 1, 2, \ldots .$$

$Y = \sum_{i=1}^{n} X_i$

3) The $N(\mu, 1)$ distribution is an element of the exponential family since its pdf is of the form

$K(x) = x$

$$\frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} = \exp\left\{ x\mu - \frac{x^2}{2} - \frac{\mu^2}{2} - \frac{1}{2}\ln(2\pi) \right\} , \quad x \in \mathbb{R} .$$

Note: For all of these, the sum $\sum_{i=1}^{n} X_i$ of the observations of the random sample was a sufficient statistic for the parameter.

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution belonging to the exponential family.

Then the statistic $\boxed{\sum_{i=1}^{n} K(X_i)}$ is sufficient for θ.

Notes: This theorem provides a simple strategy to find a sufficient statistic for a parameter: just identify $K(x)$.

For instance, in all the previous examples we had $K(x) = x$. As a consequence, the statistic

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

was a sufficient statistic for the distribution parameter.

Proof : We will use the factorization theorem:

Since $X_1, \dots, X_n$ is a random sample from a distribution belonging to the exponential family, then

$$f_{joint}(x_1,\dots,x_n,\theta) = \prod_{i=1}^{n} f(x_i,\theta) \xrightarrow{\text{exponential}} \prod_{i=1}^{n} \exp\big(K(x_i).p(\theta) + S(x_i) + q(\theta)\big)$$

$$= \exp\Big(\sum_{i=1}^{n}\big(K(x_i)p(\theta) + S(x_i) + q(\theta)\big)\Big)$$

$$= \exp\Big(p(\theta).\sum_{i=1}^{n} K(x_i) + nq(\theta) + \sum_{i=1}^{n} S(x_i)\Big)$$

$$= \underbrace{\exp\Big(p(\theta).\sum_{i=1}^{n} K(x_i) + n\,q(\theta)\Big)}_{\phi(y,\theta),\ \ y=\sum_{i=1}^{n} K(x_i)} . \underbrace{\exp\Big(\sum_{i=1}^{n} S(x_i)\Big)}_{h(x_1,x_2,\dots,x_n)}$$

We conclude that f_{joint} may be written as

$$f_{joint}(x_1, x_2, ..., x_n) = \phi(y, \theta) \cdot h(x_1, x_2, ..., x_n)$$

with $\phi(y, \theta) = \exp\left(p(\theta) \cdot y + q(\theta)\right)$, $y = \sum_{i=1}^{n} K(x_i)$

and $h(x_1, x_2, ..., x_n) = \exp\left(\sum_{i=1}^{n} S(x_i)\right)$

The factorization theorem guarantees that $Y = \sum_{i=1}^{n} K(X_i)$

is a sufficient statistic

Example

Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with pdf

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \ , \quad x \in \mathbb{R} \ , \quad \theta > 0 \ .$$

Identify a sufficient statistic for θ.

One approach would be to use the factorization theorem. (we've already seen examples like that)

Another approach is to check that the exponential distribution is an element of the exponential family of distributions:

Observe that $f(x,\theta)$ may be written as:

$$f(x;\theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) = \underbrace{\exp(-\ln\theta)}_{1/\theta} \cdot \exp\left(-\frac{x}{\theta}\right)$$

$$= \exp\left(-\frac{x}{\theta} - \ln\theta\right) = \exp\Big(\underbrace{x}_{K(x)} \cdot \underbrace{\left(-\frac{1}{\theta}\right)}_{p(\theta)} \underbrace{- \ln\theta}_{q(\theta)}\Big) \qquad S(x) = 0$$

that is,

$f(x,\theta)$ is of the form $\exp\left(K(x) . p(\theta) + S(x) + q(\theta) \right)$

with $K(x) = x$, $p(\theta) = -\frac{1}{\theta}$, $S(x) = 0$, $q(\theta) = -\ln\theta$.

By the previous theorem, we know that

$$Y = \sum_{i=1}^{n} K(X_i) = \sum_{i=1}^{n} X_i$$ is a sufficient statistic for θ

(As a consequence, since $v(y) = \frac{1}{n} y$ is invertible, then

$$\overline{X} = v(Y) = \frac{1}{n} \sum_{i=1}^{n} X_i$$ is also sufficient)