TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
Sufficient Statistics
Completeness
Sufficiency and Exponential Families

LV Nr. 105.692

Summer Semester 2021

Sufficiency

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample with pdf from a parametric family $\{f(x|\theta) : \theta \in \Theta\}$.
- Suppose we would like to estimate the parameter value θ from our sample.
- The concept of *sufficient statistic* allows us to separate information contained in the sample into two parts.
  - One part contains all the valuable information as long as we are concerned with parameter θ, while
  - the other part contains pure noise in the sense that it has no valuable information, and can be ignored.
- This concept was introduced by Fisher in 1922.

Sufficiency and Data Reduction

Definition

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

- Let $T(\mathbf{X})$ be a sufficient statistic and consider the pair $(\mathbf{X}, T(\mathbf{X}))$: $(\mathbf{X}, T(\mathbf{X}))$ contains the same information about θ as $\mathbf{X}$ alone, since $T(\mathbf{X})$ is a function of $\mathbf{X}$.
- If we know $T(\mathbf{X})$, then $\mathbf{X}$ itself has no value since its conditional distribution given $T(\mathbf{X})$ does not depend on θ. Hence, by observing $\mathbf{X}$, in addition to $T(\mathbf{X})$, we cannot say whether one particular value of parameter θ is more likely than another.
- Therefore, once we know $T(\mathbf{X})$, we can discard $\mathbf{X}$ completely: A sufficient statistic incorporates all of the information in the data $\mathbf{X}$ about the parameter θ (assuming the correctness of the statistical model).

Example: Normal

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, with known σ^2. We want to show that $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. We know that

$$T(\mathbf{X}) = \bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

The joint pdf of the sample $\mathbf{X}$ is

$$\begin{aligned} f_{\mathbf{X}}(x) &= f_{\mathbf{X}}(x_1, \dots, x_n|\mu) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} f(x_i|\mu) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

and the pdf of T is of the form

$$f_T(T(x)) = f_T(\bar{x}) = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{n}}} e^{-\frac{(\bar{x}-\mu)^2}{2\frac{\sigma^2}{n}}}.$$

Example: Normal (ctd.)

Then the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$ is

$$
\begin{aligned}
f(\mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) &= \frac{f_{\mathbf{X}}}{f_T(T(\mathbf{x}))} \\
&= \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}}{(2\pi)^{\frac{n}{2}}\,\sigma^n} \cdot \frac{(2\pi)^{\frac{1}{2}}\,\frac{\sigma}{\sqrt{n}}}{e^{-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}} \\
&= \frac{1}{(2\pi)^{\frac{n-1}{2}}\,\sigma^{n-1}\,n^{\frac{1}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right)}.
\end{aligned}
$$

Example: Normal (ctd)

Rearranging the terms in the exponent:

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 - n(\bar{x}-\mu)^2 \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2 + 2(\bar{x}-\mu)(\sum_{n=1}^{n} x_i - n\bar{x}) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2
\end{aligned}
$$

Example: Normal (ctd)

Then,

$$
\begin{aligned}
f(\mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) &= \frac{1}{(2\pi)^{\frac{n-1}{2}}\ \sigma^{n-1}\ n^{\frac{1}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2 - n(\bar{x}-\mu)^2\right)} \\
&= \frac{1}{(2\pi)^{\frac{n-1}{2}}\ \sigma^{n-1}\ n^{\frac{1}{2}}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2},
\end{aligned}
$$

which does not depend on μ.

We conclude that $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for the mean μ in the normal family with σ^2 known.

However, $\bar{X}$ is not sufficient if σ^2 is unknown.

Sufficiency Principle

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a pdf f_{θ}.

The sufficiency principle says that all statistical inference should depend on the data only through the sufficient statistic.

The likelihood is

$$L(\theta) = f(\mathbf{X} \mid T(\mathbf{X})) f_{\theta}(T(\mathbf{X}))$$

Since we can drop terms that do not contain the parameter, the likelihood is also

$$L(\theta) = f_{\theta}(T(\mathbf{X}))$$

Sufficiency Principle

$$L(\theta) = f_{\theta}(T(\mathbf{X}))$$

1. Hence, likelihood inference and Bayesian inference automatically obey the sufficiency principle.
2. Non-likelihood frequentist inference, such as the method of moments, does not automatically obey the sufficiency principle.
3. The converse of this is also true. The Neyman-Fisher factorization (next) criterion says that if the likelihood is a function of the data $\mathbf{X}$ only through a statistic T, then T is sufficient.
4. As a result, the whole data are always sufficient, that is, the criterion is trivially satisfied when $T(\mathbf{X}) = \mathbf{X}$.
5. <u>There need not be any non-trivial sufficient statistic.</u>

Sufficient Order Statistics

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a pdf f, where we are unable to specify more information about the pdf.
The sample density is then

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_{(i)})$$

where $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ are the order statistics.

The definition of sufficiency, which leads to checking whether

$$\frac{f(\mathbf{x})}{f_T(T(\mathbf{x}))}$$

is constant as a function of a parameter, gives that the order statistics are a sufficient statistic.

In this case, there is no reduction in the sample.

Sufficient Order Statistics

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from the Cauchy with pdf,

$$f(x \mid \theta) = \frac{1}{\pi(x - \theta)^2}$$

The sufficient statistic for θ is the order statistics:

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

Check that the order statistics is sufficient also for the parameter of the logistic distribution with pdf

$$f(x \mid \theta) = \frac{e^{-(x-\theta)}}{(1 + e^{(x-\theta)})^2}$$

The Factorization Theorem

- The definition of sufficiency is hard to work with, because it does not indicate how to find a sufficient statistic, and given a candidate statistic T it might be very hard to conclude whether it is sufficient statistic because of the difficulty in evaluating the conditional distribution.
- In practice, we use a simple method for finding a sufficient statistic which can be applied in many problems: the Factorization theorem gives a general approach for how to find a sufficient statistic.

The Factorization Theorem

Theorem

(Fisher-Neyman Factorization theorem) Let $f(x|\theta)$ be the pdf of X with unknown parameter θ. Then $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that

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

where the function $g(\cdot)$ depends on θ and the statistic $T(x)$, while the function $h(\cdot)$ does not contain θ.

In particular, this theorem implies that if the likelihood $L(\theta \mid x)$ depends on the data x only through $T(x)$, then $T(x)$ is sufficient for θ.

The Factorization Theorem: Proof

We prove the theorem only for discrete random variables.

Let $\mathbf{x}$ be a realization of $\mathbf{X}$ and $T(\mathbf{x}) = t$. Suppose $T(\mathbf{X})$ is a sufficient statistic for θ. Then,

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

Since $T(\mathbf{X})$ is sufficient for θ, the first probability in the RHS does not depend on θ, while the second depends on the data $\mathbf{X}$ only through $T(\mathbf{X})$.

Set $g(T(\mathbf{x}), \theta) = \mathbb{P}_\theta(T(\mathbf{X}) = T(\mathbf{x}))$ and $h(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t)$, to complete the proof.

The Factorization Theorem: Proof

Conversely, if $L(\theta, \mathbf{x}) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$, we show that $T(\mathbf{x})$ is sufficient for θ:

$$\begin{aligned}
\mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid T(\mathbf{X}) = t) &= \frac{\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t)}{\mathbb{P}_\theta(T(\mathbf{X}) = t)} \\
&= \begin{cases} \frac{\mathbb{P}_\theta(\mathbf{X}=\mathbf{x})}{\sum_{\mathbf{x}_i : T(\mathbf{x}_i) = t} \mathbb{P}_\theta(\mathbf{X}=\mathbf{x}_i)} & T(\mathbf{x}) = t \\ 0 & T(\mathbf{x}) \neq t \end{cases} \\
&= \begin{cases} \frac{g(t,\theta)h(\mathbf{x})}{\sum_{\mathbf{x}_i : T(\mathbf{x}_i) = t} g(t,\theta)h(\mathbf{x}_i)} & T(\mathbf{x}) = t \\ 0 & T(\mathbf{x}) \neq t \end{cases} \\
&= \begin{cases} \frac{h(\mathbf{x})}{\sum_{\mathbf{x}_i : T(\mathbf{x}_i) = t} h(\mathbf{x}_i)} & T(\mathbf{x}) = t \\ 0 & T(\mathbf{x}) \neq t \end{cases}
\end{aligned}$$

which does not depend on θ, so that $T(\mathbf{X})$ is sufficient by definition.

Factorization theorem: Normal

Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2)$. Then,

$$
\begin{aligned}
f(\mathbf{x}|\theta) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2} \\
&= \underbrace{\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right)}}_{\text{g(T(x))}} = g(T(\mathbf{x}))
\end{aligned}
$$

depends on the sample through the functions $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} x_i$.

Factorization theorem: Normal

Thus,

$$T(\mathbf{X}) = \Big(\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i\Big)$$

is a sufficient statistic (here $h(x) = 1$) for (μ, σ^2). In this example we actually have a pair of sufficient statistics.

We can also write

$$\begin{aligned} f(\mathbf{x}|\theta) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\, e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right)} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\, e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\bar{x}-\mu)^2\right)}. \end{aligned}$$

Thus, $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$ is another sufficient statistic for (μ, σ^2) in the normal model.

Another sufficient statistic is $T(\mathbf{X}) = (X_1, \ldots, X_n)$. Note that $\bar{X}$ itself is not a sufficient in this example.

Sufficiency and Data Reduction

- Sufficiency is related to the concept of **data reduction**.
- Suppose $\mathbf{X}$ takes values in $\mathbb{R}^n$. If we can find a sufficient statistic T that takes values in $\mathbb{R}^j$, $j < n$ then we can reduce the original data vector $\mathbf{X}$, whose dimension n is usually very large, to the vector statistic T, whose dimension j is usually much smaller than n, with no loss of information about the parameter θ.
- The lower dimensional T captures all the information about θ contained in the n-dimensional sample.

HW

1. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution for which the value of the mean λ is unknown $(\lambda > 0)$. Show that $T = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ.

2. Let $X_1, \ldots, X_n$ be a random sample with a population with pdf $f(x|\theta)$ with θ an unknown parameter. Then the statistics $T_1 = (X_1, \ldots, X_n)$ and $T_2 = (X_1^2, \ldots, X_n^2)$ are sufficient statistics for θ, while $T_3 = (X_1, \ldots, X_{n-1})$ and $T_4 = X_1$ are not sufficient statistics for θ.

3. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases},$$

   where the parameter θ is unknown $(\theta > 0)$. Check if $T = \prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

4. Let $X_1, \ldots, X_n$ be a random sample from a *Gamma*(α, β) population. Find a sufficient statistic for (α, β).

Theorem

Let $X_1, \ldots, X_n$ be a random sample from a population with pdf $f(x|\theta)$, where the parameter θ is unknown ($\theta \in \Theta$). Suppose that $T(X_1, \ldots, X_n)$ and $T_1(X_1, \ldots, X_n)$ are two statistics and there is a one-to-one map between T and T_1. Then T_1 is a sufficient statistic for θ if and only if T is a sufficient statistic for θ.

Proof: Let the one-to-one mapping between T and T_1 be u, i.e. $T_1 = u(T)$ and $T = u^{-1}(T_1)$ and u^{-1} is also one-to-one. The statistic T is sufficient if and only if the joint pdf $f_{\mathbf{X}}(\mathbf{x}|\theta)$ can be factorized as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta) \cdot h(\mathbf{x}).$$

This can be written in terms of the statistic T_1 as follows

$$g(T(\mathbf{x})|\theta) \cdot h(\mathbf{x}) = g(u^{-1}(T_1(\mathbf{x}))|\theta) \cdot h(\mathbf{x}) = g_1(T_1(\mathbf{x})|\theta) \cdot h(\mathbf{x}).$$

Therefore the joint pdf can be factorized as $f_{\mathbf{X}}(\mathbf{x}|\theta) = g_1(T_1(\mathbf{x})|\theta) \cdot h(\mathbf{x})$ and by Theorem 1, we conclude T_1 is sufficient.

Example

- In a previous example we considered a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from $\mathcal{N}(\mu, \sigma^2)$, with known σ^2.
- We showed that $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ.
- Moreover, the statistic $T_1(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is also sufficient, since there exists one-to-one mapping u between T and T_1 given by $u(T_1(\mathbf{X})) = nT(\mathbf{X})$.
- Other statistics like $T_2(\mathbf{X}) = T_1^3(\mathbf{X}) = \left(\sum_{i=1}^{n} X_i\right)^3$ and $T_3(\mathbf{X}) = e^{T(\mathbf{X})} = e^{\bar{X}}$ are also sufficient statistics.

If $\sum\limits_{i=1}^{n} X_i$ is a sufficient statistic for θ, then $(\sum\limits_{i=1}^{m} X_i, \sum\limits_{i=m+1}^{n} X_i)$ and $(\sum\limits_{i=1}^{n} X_i, \sum\limits_{i=1}^{n} X_i^2)$ are sufficient for θ.

Example: Beta

Let $X_1, \dots, X_n$ be a random sample from a beta distribution with parameters α and β, i.e. with pdf of the form

$$f(x|\beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}, & 0 \leqslant x \leqslant 1 \\ 0, & \text{otherwise} \end{cases},$$

where the value of α is known and the value of β is unknown ($\beta > 0$). Show that

$$T = \frac{1}{n}\Big(\sum_{i=1}^{n} \log \frac{1}{1-X_i}\Big)^3$$

is a sufficient statistic of the parameter β.

Example: Beta (ctd.)

First we represent the joint pdf of $\mathbf{X} = (X_1, \ldots, X_n)$ as

$$f_{\mathbf{X}}(\mathbf{x}|\beta) = \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}$$

$$= \frac{1}{\Gamma(\alpha)}^{n} \Big(\prod_{i=1}^{n} x_i\Big)^{\alpha-1} \cdot \Big(\frac{\Gamma(\alpha+\beta)^n}{\Gamma(\beta)^n}\,\big(\prod_{i=1}^{n}(1-x_i)\big)^{\beta-1}\Big).$$

Denote by T_1 the following statistic

$$T_1(\mathbf{X}) = \prod_{i=1}^{n}(1-X_i).$$

Example: Beta (ctd.)

If we also let

$$h(\mathbf{x}) = \frac{1}{\Gamma(\alpha)}^{n} \Big(\prod_{i=1}^{n} x_i \Big)^{\alpha-1} \qquad \text{and} \qquad g(T_1(\mathbf{x})|\beta) = \frac{\Gamma(\alpha+\beta)^n}{\Gamma(\beta)^n} T_1(\mathbf{x})^{\beta-1},$$

then the joint pdf can be represented in the form

$$f_{\mathbf{X}}(\mathbf{x}|\beta) = g(T_1(\mathbf{x})|\beta) \cdot h(\mathbf{x}).$$

The function h depends only on $\mathbf{x} = (x_1, \ldots, x_n)$, and the function g depends on $\mathbf{x}$ only through T_1. By the Factorization theorem we conclude that T_1 is a sufficient statistic for β.

Example: Beta (ctd.)

Consider now the following statistic

$$T(\mathbf{X}) = u(T_1(\mathbf{X})) = -\frac{1}{n}\log^3(T_1(\mathbf{X}))$$

obtained from T_1 through one-to-one mapping $u(t) = -\frac{1}{n}\log^3 t$. Then by Theorem 2 it follows that T is also a sufficient statistic for β.

1. **[HW]** Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \begin{cases} \frac{\Gamma(2\theta)}{\Gamma(\theta)^2}\left(x(1-x)\right)^{\theta-1}, & 0 \leqslant x \leqslant 1 \\ 0, & \text{otherwise} \end{cases},$$

where $\theta > 0$ is unknown parameter. Find a sufficient statistic for θ.

Minimal sufficient statistics

- A sufficient statistic is not unique!
  - Obvious example: If $T(\mathbf{X})$ is sufficient then for any other statistic $S(\mathbf{X})$, $(T(\mathbf{X}), S(\mathbf{X}))$ is also sufficient.
  - Also, the entire sample is always a trivial sufficient statistic
- Intuitively, we would like to find a *minimal* sufficient statistic that implies the maximal reduction of the data

Minimal sufficient statistics

Suppose we have two statistics: $T(\mathbf{X})$ and $T^*(\mathbf{X})$.

We say that T^* is not bigger than T if there exists some function r such that $T^*(\mathbf{X}) = r(T(\mathbf{X}))$.

That is, we can calculate $T(\mathbf{X})$ whenever we know $T^*(\mathbf{X})$.

Definition

A sufficient statistic $T(\mathbf{X})$ is called *minimal* if for any sufficient statistic $T^*(\mathbf{X})$ there exists some function r such that $T^*(\mathbf{X}) = r(T(\mathbf{X}))$.

Minimal sufficient statistics

- A minimal sufficient statistic is a function of any other sufficient statistic and it contains all the information from a sample relevant to the estimation of unknown parameters of the sample.
- Thus, the minimal sufficient statistic gives us the greatest data reduction without a loss of information about parameters.
- The following theorem gives a characterization of minimal sufficient statistics.

Minimal sufficient statistics

Theorem

Let $f(x|\theta)$ be the pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{X})$ such that, for every sample points $\mathbf{x}$ and $\mathbf{y}$ the ratio

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} \tag{2}$$

is constant as a function of θ if and only if

$$T(\mathbf{x}) = T(\mathbf{y}).$$

Then $T(\mathbf{X})$ is a *minimal sufficient statistic* for θ.

Examples

- Suppose $X_1, \ldots, X_n$ is a random sample from uniform $(\theta, 1+\theta)$ distribution, where θ is unknown. The ratio (2) is a constant as a function of θ, i.e. independent of θ, if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$, which is the case if and only if $T(x) = T(y)$. Therefore $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is minimal sufficient statistic for θ.

- Let X_1, X_2, X_3 be i.i.d. random variables following Bernoulli (p) distribution, with $0 < p < 1$ unknown. Denote by $\mathbf{X} = (X_1, X_2, X_3)$. Then $T_1(\mathbf{X}) = X_1 + X_2 + X_3$ is a minimal sufficient statistic for p, while the statistic $T_1(\mathbf{X}) = (X_1 + X_2, X_3)$ is a sufficient but not a minimal sufficient statistic for p.

Examples: Normal

Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2)$.

- Let $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ be two sample points.
- Let $(\bar{x}, s_x^2)$ and $(\bar{y}, s_y^2)$ be the sample means and sample variances corresponding to the samples $\mathbf{x}$ and $\mathbf{y}$ respectively.

Examples: Normal

- Then, we use the representation (17) to write the ratio

$$\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(\mathbf{y}|\mu,\sigma^2)} = \frac{e^{-\frac{1}{2\sigma^2}\left((n-1)s_x^2+n(\bar{x}-\mu)^2\right)}}{e^{-\frac{1}{2\sigma^2}\left((n-1)s_y^2+n(\bar{y}-\mu)^2\right)}} = e^{\frac{1}{2\sigma^2}\left(-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_x^2-s_y^2)\right)}.$$

  The ratio will be constant as function of μ and σ^2 if and only if $\bar{x}=\bar{y}$ and $s_x^2 = s_y^2$.

- Therefore, by Theorem 3 we conclude that $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

Cauchy

- Let $X_1, \ldots, X_n$ be a random sample from the Cauchy distribution with parameter θ, i.e. with the pdf of the form

$$f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$$

  with the unknown location parameter θ.

- Then

$$f_{\mathbf{X}}(x) = f_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) = \frac{1}{\pi^n \prod_{i=1}^{n}(x_i-\theta)^2}.$$

- By Theorem 2 we conlude that $T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient statistic for θ.

- A minimal sufficient statistic is *not* unique. Any one-to-one function of a minimal sufficient statistic is a minimal sufficient statistic.

- **[HW]** Let $X_1, \ldots, X_n$ be a random sample from from uniform $(\theta, 1+\theta)$ distribution, where θ is unknown. We showed that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic for θ. Show that

$$T_1 = \left(X_{(n)} - X_{(1)}, \frac{X_{(1)} + X_{(n)}}{2} \right)$$

  is also minimal sufficient statistic for θ.

Connection between unbiasedness and sufficient statistics

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a population with pdf $f(x|\theta)$, with unknown parameter θ.
- Let $\widehat{\theta}$ be an estimator of θ and $T(\mathbf{X})$ be a sufficient statistic for θ.
- We say that a statistic $S = S(\mathbf{X})$ is an *efficient* estimator of $\tau(\theta) = \mathbb{E}_{\theta}(S(\mathbf{X}))$ if

  $$MSE_{\tau(\theta)}(S) = \frac{\tau'(\theta)^2}{nI(\theta)}.$$

  In other words, an efficient estimator S is the best possible estimator of $\tau(\theta)$ in the sense that it achieves the smallest possible value for the MSE for all θ.

- Particularly, an unbiased estimator $\hat{\theta}$ of θ which achieves the Cramér-Rao lower bound $\frac{1}{nI(\theta)}$ is *efficient* or *uniformly best unbiased estimator*.
- The Rao-Blackwell theorem states that if an estimator is not a function of a sufficient statistic, it can be improved so that its modification has smaller MSE.
- That is, it shows that for any estimator W there is another estimator which depends on data $\mathbf{X}$ only through $T(\mathbf{X})$ and has smaller variance, i.e. is uniformly better than W.

Rao-Blackwell Theorem

Theorem

Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define

$$\phi(T) = E(W \mid T).$$

Then,

$$\mathbb{E}_{\theta}(\phi(T)) = \tau(\theta)$$

and

$$\mathbb{V}\mathrm{ar}_{\theta}(\phi(T)) \leqslant \mathbb{V}\mathrm{ar}_{\theta}(W)$$

for all θ, i.e. $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Proof

We use the law of total expectation and the conditional variance identity. The estimator $\phi(T)$ is unbiased for $\tau(\theta)$ because

$$\tau(\theta) = \mathbb{E}_\theta(W) = \mathbb{E}_\theta \mathbb{E}_\theta(W \mid T) = \mathbb{E}_\theta(\phi(T)).$$

Also,

$$\begin{aligned} \mathbb{V}\mathrm{ar}_\theta(W) &= \mathbb{V}\mathrm{ar}_\theta(\mathbb{E}_\theta(W \mid T)) + \mathbb{E}_\theta\left(\mathbb{V}\mathrm{ar}_\theta(W \mid T)\right) \\ &= \mathbb{V}\mathrm{ar}_\theta(\phi(T)) + \mathbb{E}_\theta\left(\mathbb{V}\mathrm{ar}_\theta(W \mid T)\right) \\ &\geqslant \mathbb{V}\mathrm{ar}_\theta(\phi(T)) \end{aligned}$$

since $\mathbb{V}\mathrm{ar}_\theta(W \mid T) \geqslant 0$.

Proof (ctd.)

Thus, $\phi(T)$ has smaller variance than W, i.e. smaller MSE.
We conclude, $\phi(T)$ is uniformly better than W.

Since T is sufficient and W is only a function of the sample it follows that the distribution of $W \mid T$ does not depend on θ.

Thus, $\phi = \mathbb{E}_{\theta}(W \mid T)$ is a function of the sample and is independent of θ, i.e. $\phi(T)$ is an estimator.

We conclude, $\Phi(T)$ is a uniformly better unbiased estimator of the parameter $\tau(\theta)$.

It follows from the Rao-Blackwell theorem that when searching for efficient estimators, there is no need to consider estimators that cannot be written as functions of a sufficient statistic.

Binomial: Let $X_1, \dots, X_n$ be a random sample from $bin(p,k)$, i.e.

$$P(X_j = i) = \binom{k}{i} p^i (1-p)^{k-i}, \qquad i > 0.$$

Suppose our parameter of interest is the probability of one success, i.e. $\theta = P(X=1) = kp(1-p)^{k-1}$. One possible estimator of θ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i=1)}.$$

This estimator is unbiased, i.e. $\mathbb{E}(\hat{\theta}) = \theta$.

Example: Binomial (ctd.)

Let us now find a sufficient statistic for θ. The joint pmf is of the form

$$\begin{aligned} f(x) = f(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i} \\ &= a(x_1, \ldots, x_n)\, p^{\sum_{i=1}^{n} x_i} (1-p)^{nk - \sum_{i=1}^{n} x_i}, \end{aligned}$$

where $a(x_1, \ldots, x_n)$ is a function of the sample. Thus $T = \sum_{i=1}^{n} X_i$ is sufficient. In fact it is minimal sufficient, as we show next.

Example: Binomial (ctd.)

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two sample points. The ratio

$$\begin{aligned}\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} &= \frac{a(x_1, \ldots, x_n)}{a(y_1, \ldots, y_n)} \cdot \frac{p^{\sum_{i=1}^n x_i} (1-p)^{nk - \sum_{i=1}^n x_i}}{p^{\sum_{i=1}^n y_i} (1-p)^{nk - \sum_{i=1}^n y_i}} \\ &= \frac{a(\mathbf{x})}{a(\mathbf{y})} \cdot p^{(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)} \cdot (1-p)^{-(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)}\end{aligned}$$

is a constant as function of θ if and only if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Therefore, by Theorem 3 we conclude that $T = \sum_{i=1}^n X_i$ is a minimal sufficient statistic for θ.

Using the Rao-Blackwell theorem, we can improve $\hat{\theta}$ by considering its conditional expectation given T. Let $\phi(T) = E(\hat{\theta}|T)$ denote this estimator. Then, for any nonnegative integer t,

$$\begin{aligned}
\phi(t) = E(\hat{\theta}|T = t) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{(X_i=1)}\,\Big|\,\sum_{i=1}^{n}X_i = t\right) \\
&= \frac{1}{n}\cdot E\left(\sum_{i=1}^{n}P(X_i=1)\,\Big|\,\sum_{j=1}^{n}X_j = t\right) \\
&= P\Big(X_1 = 1\,\Big|\,\sum_{j=1}^{n}X_j = t\Big) \\
&= \frac{P\big(X_1 = 1, \sum_{j=1}^{n}X_j = t\big)}{P\big(\sum_{j=1}^{n}X_j = t\big)} \\
&= \frac{P\big(X_1 = 1, \sum_{j=2}^{n}X_j = t-1\big)}{P\big(\sum_{j=1}^{n}X_j = t\big)}
\end{aligned}$$

$$
\begin{aligned}
\cdots &= \frac{P\big(X_1 = 1\big) \cdot P\big(\sum_{j=2}^{n} X_j = t-1 \big)}{P\big(\sum_{j=1}^{n} X_j = t \big)} \\
&= \frac{kp(1-p)^{k-1} \cdot \binom{k(n-1)}{t-1} p^{t-1} (1-p)^{k(n-1)-(t-1)}}{\binom{kn}{t} p^{t} (1-p)^{kn-t}} \\
&= \frac{k \cdot \binom{k(n-1)}{t-1}}{\binom{kn}{t}} \\
&= \frac{k \cdot (k(n-1))!}{(t-1)!(kn-k-t+1)!} \cdot \frac{t!(kn-t)!}{(kn)!} \\
&= \frac{k \cdot (kn-k)! \cdot t}{(kn)!(kn-k-t+1)!}.
\end{aligned}
$$

- In the previous calculations we used that X_1 and $(X_2, \ldots, X_n)$ are independent, $\sum_{i=1}^n X_i \sim bin(kn, p)$ and $\sum_{i=2}^n X_i \sim bin(k(n-1), p)$.
- Thus, our new estimator is

$$\hat{\theta}_1 = \phi(T) = \phi(\sum_{i=1}^n X_i) = \frac{k \cdot (kn-k)! \cdot \left(\sum_{i=1}^n X_i\right)}{(kn)!(kn-k+1-\sum_{i=1}^n X_i)!}.$$

- By the Rao-Blackwell theorem, the estimator $\hat{\theta}_1$ is unbiased and has smaller variance, i.e. smaller MSE, than $\hat{\theta}$.
- Note that the result does not depend on p.

Sufficient Statistics and Exponential Family

It is easy to find a sufficient statistic from an exponential family of distributions using the Factorization theorem. Let's look at an example.

Binomial sufficient statistic: Consider a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from $bin(p, k)$, with unknown $0 < p < 1$. The joint pmf is of the form

$$\begin{aligned}
P(\mathbf{X} = x) &= \prod_{i=1}^{n} p(X_i = x_i) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i} \\
&= (\prod_{i=1}^{n} \binom{k}{x_i}) \cdot p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{kn - \sum_{i=1}^{n} x_i} \\
&= \left(\prod_{i=1}^{n} \binom{k}{x_i} \right) \cdot (1-p)^{nk} \cdot e^{\left(\sum_{i=1}^{n} x_i \right) \log \frac{p}{1-p}} \\
&= h(x) \cdot g(T(\mathbf{x})|p),
\end{aligned}$$

Sufficient Statistics and Exponential Family

with $h(x) = \prod_{i=1}^{n} \binom{k}{x_i}$, $T_1(x) = \sum_{i=1}^{n} x_i$ and

$$g(T(\mathbf{x})|p) = (1-p)^{nk} \cdot e^{\left(\sum_{i=1}^{n} x_i\right) \log \frac{p}{1-p}} = (1-p)^{nk} \cdot e^{T_1(x) \cdot \log \frac{p}{1-p}}.$$

Thus, by the factorization theorem we conclude $T_1(x) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for the one dimensional unknown parameter p.

Sufficiency and Exponential Families

A generalization of the computations in the binomial example is an important result.

Theorem

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be i.i.d. observations from a pdf or pmf $f(x|\theta)$ that belongs to an exponential family

$$f(x|\theta) = h(x)\, c(\theta)\; e^{\sum_{i=1}^{k} w_i(\theta)\, t_i(x)}, \tag{3}$$

where $\theta = (\theta_1, \ldots, \theta_d)$, $d \leqslant k$. Then,

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right) \tag{4}$$

is a sufficient statistic for θ.

Sufficiency and Exponential Families

1. The theorem is a straightforward application of the factorization criterion
2. The k-dimensional statistic $T(\mathbf{X})$ is called *natural sufficient statistic* of the family.
3. The parameter
$$\boldsymbol{\psi} = (w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta}))$$
is a k-dimensional parameter for the family and is called the *natural parameter*.
4. From now on, we will use $\boldsymbol{\theta}$ instead of $\boldsymbol{\psi}$ to denote the natural parameter

Complete Statistics

- Does "Rao-Blackwellization" necessarily yield an UMVUE?
- Generally not.
- To guarantee UMVUE an additional requirement of *completeness* on a sufficient statistic T is needed.
- Under mild conditions, it can be shown that if the distribution of the data belongs to the k-parameter exponential family and an unbiased estimator is a function of the corresponding sufficient statistic $\left(\sum_{i=1}^{n} T_1(x_i), \ldots, \sum_{i=1}^{n} T_k(x_i)\right)$, it is an UMVUE

Complete Statistics

Definition

Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. This family of probability distributions is called *complete* if

$$E_{\theta}(g(T)) = 0 \quad \text{for all } \theta$$

implies

$$P_{\theta}\left(g(T) = 0\right) = 1 \quad \text{for all } \theta$$

$T(\mathbf{X})$ is called a *complete statistic*.

In other words, $T(\mathbf{X})$ is complete if $\mathbb{E}_{\theta} g(T) = 0$ for all θ, then necessarily $g(T) = 0$.

Binomial complete sufficient statistic

Suppose X_i iid Bernoulli(p). Let $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$.

Then, $T \sim bin(n, p)$, $0 < p < 1$ and let g be a function such that $E\, g(T) = 0$. Then,

$$
\begin{aligned}
0 = E\, g(T) &= \sum_{t=0}^{n} g(t) \binom{n}{t} p^t \, (1-p)^{n-t} \\
&= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t
\end{aligned}
$$

for all p, $0 < p < 1$. The factor $(1-p)^n \neq 0$ for all $0 < p < 1$.

Binomial complete sufficient statistic

Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all r, $r > 0$. The last equation is a polynomial of degree n in r, where the coefficient of r^t is $g(t) \binom{n}{t}$.

For the polynomial to be equal zero for all r, each coefficient must be zero.

Since none of the binomial coefficients $\binom{n}{t}$ is zero, this implies that $g(t) = 0$ for all $t = 0, 1, \ldots, n$.

Since T takes the values $0, 1, \ldots, n$ with probability 1, we conclude $P(g(T) = 0) = 1$ for all p, as desired.

Therefore, T is a complete statistic.

Complete statistics in the exponential family

To verify completeness for a general distribution can be a non-trivial mathematical problem.

Fortunately, it is much simpler for the exponential family of distributions.

Theorem

Let $X_1, \ldots, X_n$ be i.i.d. observations from an exponential family with pdf or pmf of the form (3), where $\theta = (\theta_1, \ldots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j) \right) \tag{5}$$

is complete as long as the parameter space Θ contains an open set in $\mathbb{R}^k$.

- The condition that the parameter space Θ contains an open set is needed to avoid a situation as the following.
- The $\mathcal{N}(\theta, \theta^2)$ can be written in the form (3), but the parameter space (θ, θ^2) does not contain a two-dimensional open set, as it consists of only the points on a parabola.
- Exponential families such as the $\mathcal{N}(\theta, \theta^2)$, where the parameter space is a lower-dimensional curve, are called *curved exponential families*.

- Completeness is often used to prove the uniqueness of various estimators
- Completeness of a sufficient statistic yields its minimal sufficiency:

Theorem

If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.

- In the full rank exponential family (Θ is open), the sufficient statistic is minimal and complete.

- Recall that a statistic $S = S(\mathbf{X})$ is called an *efficient* estimator of $\tau(\theta) = \mathbb{E}_{\theta} S(\mathbf{X})$ if
$$MSE_{\tau(\theta)}(S) = \frac{\tau'(\theta)^2}{nI(\theta)}.$$
- In other words, efficient estimator S is the best possible unbiased estimator of $\tau(\theta)$ in the sense that it achieves the smallest possible value for the MSE for all θ.
- Moreover, let S be any unbiased estimator of $\tau(\theta)$. Then, it is efficient for $\tau(\theta)$ if and only if there exists a function $a(\theta)$ such that the *attainability condition*
$$a(\theta)\Big(S(\mathbf{x}) - \tau(\theta)\Big) = \ell'(\theta|x) \tag{6}$$
holds, where $\ell'(x|\theta) = \frac{\partial}{\partial\theta} \log L(x|\theta)$ is the log likelihood function.

Thus, an unbiased efficient estimator of $\tau(\theta)$ can be represented in the form

$$S = a(\theta) \sum_{i=1}^{n} z(X_i|\theta) \, + \, \tau(\theta),$$

for some function $a(\theta)$.

The statistic $S(\mathbf{X})$ must be a function of the sample only and it can not depend on θ. This means that efficient estimates do not always exist and they exist only if we can represent the derivative of log likelihood $\ell'(\theta)$ as

$$\ell' = \sum_{i=1}^{n} z(X_i, \theta) = \frac{S - \tau(\theta)}{a(\theta)},$$

where S does not depend on θ.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a pdf in the exponential family (3). Then,

$$\log f(x|\theta) = \log h(x) + \log c(\theta) + \sum_{i=1}^{k} w_i(\theta) \cdot t_i(x)$$

and the score function

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{c'(\theta)}{c(\theta)} + \sum_{i=1}^{k} w_i'(\theta) \cdot t_i(x).$$

This implies that

$$\sum_{j=1}^{n} z(X_j, \theta) = n\,\frac{c'(\theta)}{c(\theta)} + \sum_{i=1}^{k} w_i'(\theta) \cdot \sum_{j=1}^{n} t_i(X_j)$$

and

$$\frac{1}{n}\sum_{j=1}^{n} t_i(X_j) = \frac{1}{n\sum_{i=1}^{n} w_i'(\theta)} \sum_{j=1}^{n} z(X_j, \theta) - \frac{c'(\theta)}{c(\theta)\sum_{i=1}^{k} w_i'(\theta)}.$$

If we take

$$S = \frac{1}{n}\sum_{j=1}^{n} t_i(X_j) \quad \text{and} \quad \tau(\theta) = \mathbb{E}_{\theta}(S) = -\frac{c'(\theta)}{c(\theta)\sum_{i=1}^{k} w_i'(\theta)} \tag{7}$$

then S will be an efficient estimate of $\tau(\theta)$. We used the property $\mathbb{E}_{\theta} z(X, \theta) = 0$ proven in Fisher information and the Cramér-Rao lower bound section.

Example: Binomial efficient statistic

Consider a sample $\mathbf{X} = (X_1, \ldots, X_n)$ from $Bin(k, p)$. The statistic

$$S = \frac{1}{n} \sum_{j=1}^{n} t_1(X_j) = \frac{1}{n} \sum_{j=1}^{n} X_j = \bar{X}$$

is an efficient estimator of its expectation

$$\tau(p) = \mathbb{E}(S) = E\,\bar{X} = \mathbb{E}(X_1) = kp.$$

We can also compute its expectation directly using expression (7):

$$\mathbb{E}(S) = -\frac{c'(p)}{c(p)\, w_1'(p)} = -\frac{-k(1-p)^{k-1}}{(1-p)^k} \;\cdot p(1-p) = kp.$$

- In general, method of moments estimators are not functions of sufficient statistics, and therefore can be always improved upon by conditioning on a sufficient statistic (see Rao-Blackwel theorem).
- In the case of exponential families, there can be a correspondence between a modified method of moments strategy and maximum likelihood estimation.

Example

Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample from a pdf in the exponential family (3), where the support of $f(x|\theta)$ is independent of θ. The likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) = h(\mathbf{x}) \cdot \big(c(\theta)\big)^n \, e^{\sum_{i=1}^{k} w_i(\theta) \sum_{j=1}^{n} t_i(x_j)},$$

A *modified method of moments* approach would $w_i(\theta)$, $i = 1, \ldots, k$ by $\hat{w}_i(\theta)$, the solutions of the following system of k equations

$$\sum_{j=1}^{n} t_i(x_j) = \mathbb{E}_{\theta} \left(\sum_{j=1}^{n} t_i(X_j) \right), \quad i = 1, \ldots, k.$$

Moreover, the estimators $\hat{w}_i(\theta)$ are the MLEs of $w_i(\theta)$.

Example

A large telephone company wants to estimate the average number of telephone calls made by its private clients. It is known that on the average women make a times more calls than men. It is reasonable to assume that the number of daily calls has a Poisson distribution. During a certain day, the company registered the numbers of daily calls made by n female and m male randomly chosen clients.

1. What is the statistical model for the data? Does the distribution of the data belong to the exponential family?
2. Find the MLE for the average daily number of calls made by female and male clients.
3. Are the MLEs unbiased?

Example: Answer

Let $X_1, \ldots, X_n$ be the number of daily calls made by female clients, and $Y_1, \ldots, Y_m$ those made by male clients.

Then X_i are iid Poisson($a\lambda$) and Y_j are iid Poisson(λ).

Their joint distribution is

$$f_\lambda(\mathbf{x}, \mathbf{y}) = e^{(an+m)\lambda} \lambda^{\sum_{i=1}^n x_i + \sum_{j=1}^m y_j} \frac{a^{\sum_i x_i}}{\prod_{i=1}^n x_i! \prod_{j=1}^m y_j!}$$

Obviously the joint belongs to the exponential family and

$$T(\mathbf{X}, \mathbf{Y}) = \sum_i X_i + \sum_j Y_j$$

is the *minimal sufficient* and *complete* statistic for λ.

Example: Answer

Since $L(\lambda \mid \mathbf{x}, \mathbf{y}) = f_\lambda(\mathbf{x}, \mathbf{y})$, the MLEs for the average numbers of daily calls are

$$\hat{\lambda} = \frac{\sum_i X_i + \sum_j Y_j}{an + m} \quad \text{males}$$

$$a\hat{\lambda} = a\frac{\sum_i X_i + \sum_j Y_j}{an + m} \quad \text{females}$$

Easy to verify that both are unbiased for λ and $a\lambda$, respectively.