stats6254suff 625.3 - Sufficiency

Gunnar StefÃąnsson

December 1, 2015

Contents

1 Sufficient statistics

1.1 Data Reduction

Let $\{X\}_n$ be i.i.d.
If $T : \mathbb{R}^n \to \mathbb{R}$ is a function such that $T(\mathbf{X})$ is a random variable then $T(\mathbf{X})$ is a statistic.

1.1.1 Handout

Data reduction

Let $X_1, ..., X_n$ be i.i.d. random variables with a common c.d.f., F_θ, where the parameter θ is unknown, but in some parameter set $\theta \in \Theta$. We commonly have $\theta \in \mathbb{R}$, sometimes $\theta \in \mathbb{R}^p$ and Θ may even be a discrete set. Write $\mathbf{X}$ for the random vector

$$\mathbf{X} = (X_1, .., X_n)^T : \Omega \to \mathbb{R}^n.$$

If $t : \mathbb{R}^n \to \mathbb{R}$ is a function such that $T = t \circ \mathbf{X} = t(\mathbf{X})$ is also a random variable, then $T = t(\mathbf{X})$ is called a *statistic*.

Note that we may be sloppy with the notation, alternatively using T, $t(\mathbf{X})$ or $T(\mathbf{X})$ for the same thing.

For a given set of data $x = (x_1, ..., x_n)^T$ one might consider just using $T(\mathbf{x})$ and then "forgetting" the original values, thus reducing the data set. To do this one needs to know that the resulting number $T(\mathbf{x})$ in some sense contains all the information about the parameter that is in the original data set. This section will make these concepts specific.

1.2 Sufficiency

$T(\mathbf{X})$ is called a sufficient statistic if the distribution of X, conditionally on $T(\mathbf{X})$ is a constant function of Θ
The definition implies that if $T = T(\mathbf{X})$ is sufficient then $f_{X|T}(x|t)$ does not contain Θ.

1.2.1 Handout

We want to define a concept to represent the notion that $T(\mathbf{X})$ is a *sufficient statistic* for θ. This concept should mean that information about θ is completely contained in $T(\mathbf{X})$, i.e. $\mathbf{X}$ does not give any information once we know $T(\mathbf{X})$.

Note that the only link between the data and the parameter is through the probability distribution. Thus, for a given data set, all the information about θ is contained in the joint density (or p.m.f.) of the data set, i.e. in $f_\theta(x)$.

Definition 1 $T(\mathbf{X})$ is a sufficient statistic if the distribution of $\mathbf{X}$, conditionally on $T(\mathbf{X})$, is a constant function of θ.

Remark 1.1. Note the following

- Basically the definition implies that if $T = T(\mathbf{X})$ is sufficient, then the function

  $$f_{\mathbf{X}|T}(x|t)$$

  does not contain θ. In other words, $P[\mathbf{X} \in A | T(\mathbf{X}) = t]$ is a constant in θ.

- The probability measure P_θ is indexed by $\theta \in \Theta$. Let $P_\theta[T(\mathbf{X}) = t] > 0$ for a discrete r.v.$\mathbf{X}$, to obtain

$$P_\theta[\mathbf{X} = x | T(\mathbf{X}) = t] = \frac{P_\theta[\mathbf{X} = x, T(\mathbf{X}) = t]}{P_\theta[T(\mathbf{X}) = t]}$$

(Remember that $\{X = x\}$ is a subset of $\{T(X) = T(x)\}$ see p. 273 in Casella, Berger). Assume $t = T(x)$ and we want to investigate whether

$$P_\theta[\mathbf{X} = x | T(\mathbf{X}) = T(x)] = \frac{P_\theta[\mathbf{X} = x]}{P_\theta[T(\mathbf{X}) = T(x)]}$$

is a constant in θ.
For a discrete r.v. $\mathbf{X}$ this is given by

$$P_\theta[\mathbf{X} = x | T(\mathbf{X}) = T(x)] = \frac{p_\theta(x)}{q_\theta(T(x))}$$

where p_θ is the p.m.f. of $\mathbf{X}$ and q_θ is the p.m.f. of $T(\mathbf{X})$

$$q_\theta(T) = \sum_{x:T(x)=t} p_\theta(x)$$

Theorem 1.1 If f_θ is the (joint) p.d.f. of $\mathbf{X}$ and q_θ is the p.d.f. of $T(\mathbf{X})$, then $T(\mathbf{X})$ is sufficient for θ if $\frac{p_\theta(x)}{q_\theta(T(x))}$ is a constant in θ for $x \in \mathbb{R}^n$ (or $x \in \mathbf{X}(\Omega)$).

Example 1 Consider random variables $X_1, ..., X_n \sim b(1, p)$ iid; $\theta = p$
An obvious candidate for a sufficient statistic is $T(\mathbf{X}) := \sum_{i=1}^n X_i \sim b(n, p)$ then $P[X_i = x_i] = p(1-p)$ and since we get

$$\frac{p_\theta(x)}{q_\theta(T(x))} = \frac{\prod_{i=1}^n p_i^{x_i}(1-p)^{1-x_i}}{\binom{n}{T(x)} p^{T(x)}(1-p)^{n-T(x)}} = \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{\binom{n}{\sum x_i} p^{\sum x_i}(1-p)^{n-\sum x_i}} = \frac{1}{\binom{n}{\sum x_i}}$$

We now know that $T(\mathbf{X})$ is a sufficient statistic since the last fraction does not involve θ and is thus a constant in θ

Example 2 Consider Gaussian random variables, $X_1, ..., X_n \sim n(\mu, \sigma^2)$, with known σ^2 but unknown location parameter $\theta = \mu$.

The obvious candidate for a sufficient statistic is $T(\mathbf{X}) := \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n X_i$.

The joint p.d.f. is given by

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

The density function for $T(\mathbf{X})$ is easy to obtain since it is known that $\mathbf{X} \sim n(\mu, \frac{\sigma^2}{n})$ and thus

$$g_\mu(T(\mathbf{X})) = g_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma/\sqrt{n}} e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}}$$

Note that the quadratic term involving the x and the unknown can be rewritten:

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

which implies

$$\frac{f_\mu(x)}{g_\mu(T(x))}=\frac{\frac{1}{(2\pi)^{n/2}\sigma^n}e^{-\frac{1}{2\sigma^2}\sum(x_i-\bar{x})^2}-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2}{\frac{1}{(2\pi)^{1/2}\sigma/\sqrt{n}}e^{-(\bar{x}-\mu)^2/2\sigma^2/n}}=\frac{(2\pi)^{1/2}\sigma}{\sqrt{n}(2\pi)^{n/2}\sigma^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Since this ratio does not involve μ, T is a sufficient statistic.

Example 3 Let Θ be the collection of all c.d.f.s of continuous random variables and let $X_1,...,X_n \sim F \in \Theta$ be i.i.d. Then the order statistic, $(X_{(1)},...,X_{(n)})$, is sufficient.

The search for sufficient statistics is made easier by the following theorem.

Theorem 1.2 $T(\mathbf{X})$ is a sufficient statistic if and only if there exist functions g_θ and h such that the joint p.d.f. of $\mathbf{X}$ can be written in the form

$$f_\theta(x)=g_\theta(T(x))h(x)$$

Proof. Suppose $\mathbf{X}$ is discrete.

(1) Let $T(\mathbf{X})$ be sufficient. Then we can define

$$g_\theta(t):=p_\theta[T(\mathbf{X})=t]$$

$$h(x):=p_\theta[\mathbf{X}=x|T(\mathbf{X})=t]$$

and these functions satisfy the conditions.

(2) Next assume that the functions g_θ and h exist and let q_θ be the mass function of $T(\mathbf{X})$. Take an arbitrary point $x\in\mathbb{R}^n$ and let $t=T(\mathbf{X})$. Consider

$$\frac{f_\theta(x)}{q_\theta(T(x))}=\frac{g_\theta(T(x))h(x)}{q_\theta(T(x))}=\frac{g_\theta(T(x))h(x)}{q_\theta(t)}=\frac{g_\theta(T(x))h(x)}{\sum_{y:T(y)=t}f_\theta(y)}=\frac{g_\theta(T(x))h(x)}{\sum_{y:T(y)=t}g_\theta(T(y))h(y)}=$$

to obtain

$$\frac{g_\theta(T(x))h(x)}{g_\theta(t)\sum_{y:T(y)=t}h(y)}=\frac{h(x)}{\sum_{y:T(y)=t}h(y)}$$

which is a constant in θ and hence $T(\mathbf{X})$ is sufficient. $\square$

Example 4 $X_1, ..., X_n \sim n(\mu, \sigma^2)$ iid, $\theta = (\mu, \sigma^2)$
$T(\mathbf{X}) := (\bar{\mathbf{X}}, S^2)$ is sufficient:

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

$$\underbrace{\frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{(n-1)S^2}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}}}_{=: g_\theta(T(x))}$$

1.3 Minimal Sufficient Statistics

1.3.1 Handout

Definition 2 Let $X_n \sim F_\theta$ be independent and $T\colon \mathbb{R}^n \to \mathbb{R}$ such that $T(\mathbf{X})$ is a random variable. $T(\mathbf{X})$ is a minimal sufficient statistic if for every sufficient statistic T' there exists a function k such that $T(\mathbf{x}) = k(T'(\mathbf{x})),\ \mathbf{x} \in \mathbb{R}^n$ $(\mathbf{x} \in \mathbf{X}(\Omega))$.

Theorem 1.3 If T is such that, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the ratio $\frac{f_\theta(\mathbf{x})}{f_\theta(\mathbf{y})}$ 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.

Proof: Define the sets $A_t = \{x : T(x) = t\}$ Thus if $T(x) = T(y) = t$ then x and y are both elements in A_t. Define a function $\gamma(t)$ that picks the smallest element of A_t. Note that $\gamma(T(x))$ is in the same set A_t as x but is not necessarily equal to x. The fraction $K = \frac{f_\theta(x)}{f_\theta(\gamma(T(x))}$ does not depend on θ because of how we have defined γ. We can now write the density as

$$f_\theta(x) = f_\theta(\gamma(T(x))[\frac{f_\theta(x)}{f_\theta(\gamma(T(x))}]$$

now we choose $g(T, \theta) = f_\theta(\gamma(T(x))$ and $h(x) = K$ from above (which does not depend upon θ) Obtaining by theorem 1.2. that T is a sufficient statistic.

Now let S(X) be another sufficient statistic. By theorem 1.2. we obtain $f_\theta(x) = g_2(S, \theta)h_2(x)$.

Then, if $S(x) = S(y)$,

$$\frac{f_\theta(x)}{f_\theta(y)} = \frac{g_2(S,\theta)h_2(x)}{g_2(S,\theta)h_2(y)} = \frac{h_2(x)}{h_2(y)}$$

which does not depend on θ implying $T(x) = T(y)$ by assumption.
If $T(x) = T(y)$ whenever $S(x) = S(y)$, then T is a function of S. Therefore, T is a function of any sufficient statistic S.

Now we have shown that T is both a sufficient statistic and a function of any other sufficient statistic. Thus T is a minimal sufficient statistic. q.e.d.

Example 5 $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2) in a normal distribution (both unknown).

1.4 Ancillary statistics

1.4.1 Handout

Definition 3 $S(\mathbf{x})$ is an *ancillary statistic* if the distribution of $S(\mathbf{X})$ is a constant in θ ("free of θ").

Example 6 If $X_1, \ldots, X_n \sim N(\theta, 1)$ are i.i.d., we know that

$$\begin{aligned} Z_i &= X_i - \theta \sim N(0,1) \\ \text{and} \quad \bar{X} &= \bar{Z} + \theta \sim N(\theta, 1/n) \end{aligned}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

And we know that if we define

$$\begin{aligned} \tilde{X} &= \text{median}(X_1, \ldots, X_n) \\ \tilde{Z} &= \text{median}(Z_1, \ldots, Z_n) \end{aligned}$$

then $\tilde{X}$ has a distribution with parameter θ, but the distribution of $\tilde{Z}$ has nothing to do with θ.

On the other hand, if $R = \bar{X} - \tilde{X}$ then

$$R = \bar{Z} - \tilde{Z}$$

since

$$\begin{aligned} \tilde{Z} &= \text{median}(Z_1, \ldots, Z_n) \\ &= \text{median}(X_1 - \theta, \ldots, X_n - \theta) \\ &= \text{median}(X_1, \ldots, X_n) - \theta \end{aligned}$$

But since the distribution of $\bar{Z}$ and $\tilde{Z}$ is "free of θ", so is the distribution of R. R is a random variable and is therefore an ancillary statistic.

Note that Z_i are not proper random variables: The X_i are of course random variables so they are of the form $X_i : \Omega \longrightarrow \mathbb{R}$ whereas Z_i is a function of both ω and θ, i.e. is a function of the form $Z_i : \Omega \times \Theta \longrightarrow \mathbb{R}$.

Example 7 Assume that $X_1, \ldots, X_n$ are independent random variables with a c.d.f. of the form

$$P_\theta[X_i \leq x] = I(x - \theta),$$

i.e.

$$X_1, \ldots, X_n \sim F_\theta \quad \text{with} \quad F_\theta(x) = F(x - \theta).$$

Such a family is called a **location family**.

If we write $Z_i = X_i - \theta$, then the c.d.f. of Z_i is given by:

$$\begin{aligned} P(Z_i \le z) &= P(X_i - \theta \le z) \\ &= P(X_i \le z + \theta) \\ &= F\left((z + \theta) - \theta\right) \\ &= I(z) \end{aligned}$$

which is a constant in θ.

We thus see that $R = \bar{X} - \tilde{X} = \bar{Z} - \tilde{Z}$ is an ancillary statistic.

Example 8 Let $X_1, \ldots, X_n \sim U(\theta, \theta + 1)$. Then $X_{(n)} - X_{(1)} = Z_{(n)} - Z_{(1)}$ is ancillary.

Example 9 Suppose $X_1, \ldots, X_n \sim F_\sigma$ where $F_\sigma(x) = F\left(\frac{X_i}{\sigma}\right), \quad \sigma > 0$, a **scale family**. Statistics of interest in relation to σ include

$$\begin{aligned} S &= \sqrt{\tfrac{1}{n-1} \sum \left(X_i - \bar{X}\right)^2} \\ M &= median(X_i - \tilde{X}) \end{aligned}$$

Note that M/S is an ancillary statistic [Write $V_i = \frac{X_i}{\sigma}$ etc.]

Example 10 (Location scale family) $X_1, \ldots, X_n \sim F_{\mu,\sigma}$ iid, $F_{\mu,\sigma}(x) = F\left(\frac{x-\mu}{\sigma}\right)$ and show in each of the following cases that the random variable is ancillary.

1. $$\frac{\bar{X} - \tilde{X}}{S}$$

2. $$\frac{\bar{X} - \tilde{X}}{M}$$

3. $$\frac{X_{(n)} - X_{(1)}}{\bar{X} - \tilde{X}}$$

Solution:

1. Let $Z_1, \ldots, Z_n \sim F$. We get:

$$P_{\mu,\sigma}\left[\frac{X_i - \mu}{\sigma} \le w\right] = P_{\mu,\sigma}\left[X_i \le \sigma w + \mu\right] = F_{\mu,\sigma}(\sigma w + \mu) = F(w) = P[Z_i \le w]$$

and thus

$$(Z_1, \ldots, Z_n) = \left(\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma}\right)$$

in distribution. Therefore:

$$\frac{\bar{X} - \tilde{X}}{S_X} = \frac{\sigma \bar{Z} + \mu - \sigma \tilde{Z} - \mu}{S_{\sigma Z + \mu}} = \frac{\sigma \bar{Z} - \sigma \tilde{Z}}{\sigma S_Z} = \frac{\bar{Z} - \tilde{Z}}{M_Z}$$

where

$$S = \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}.$$

2. Let $Z_1, \ldots, Z_n \sim F$. We get:

$$P_{\mu,\sigma}\left[\frac{X_i - \mu}{\sigma} \leq w\right] = P_{\mu,\sigma}\left[X_i \leq \sigma w + \mu\right] = F_{\mu,\sigma}(\sigma w + \mu) = F(w) = P[Z_i \leq w]$$

and thus

$$(Z_1, \ldots, Z_n) = \left(\frac{X_1 - \mu}{\sigma}, \ldots, \frac{X_n - \mu}{\sigma}\right)$$

in distribution. Therefore:

$$\frac{\bar{X} - \tilde{X}}{M} = \frac{\sigma \bar{Z} + \mu - \sigma \tilde{Z} - \mu}{M_{\sigma Z + \mu}} = \frac{\sigma \bar{Z} - \sigma \tilde{Z}}{\sigma M_Z} = \frac{\bar{Z} - \tilde{Z}}{M_Z}$$

where $M_X = median|X_i - \bar{X}|$.

3. Let Z_i be as in 1) and 2). We get:

$$\frac{X_{(n)} - X_{(1)}}{\bar{X} - \tilde{X}} = \frac{Z_{(n)} - Z_{(1)}}{\bar{Z} - \tilde{Z}}$$

Definition 4 A statistic $T(\mathbf{X})$ is *complete* if the following holds for all functions g:

$$\begin{aligned} &E_\theta[g(T)] = 0 \quad \text{for all} \quad \theta \in \Theta \\ \Rightarrow &P_\theta[g(T) = 0] = 1 \quad \text{for all} \quad \theta \in \Theta \end{aligned}$$

Theorem 1.4 (Basu) If $T(\mathbf{X})$ is a complete and minimal sufficient statistic and $S(\mathbf{X})$ is an ancillary statistic, then $T(\mathbf{X})$ and $S(\mathbf{X})$ are **independent**.

Proof. We give the proof only for discrete distributions.
Let $S(\mathbf{X})$ be any ancillary statistic. Then $P(S(\mathbf{X}) = s)$ does not depend on θ since $S(\mathbf{X})$ is ancillary. Also the conditional probability,

$$P(S(\mathbf{X}) = s \mid T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x} : S(\mathbf{x}) = s\} \mid T(\mathbf{X} = t)$$

does not depend on θ because $T(\mathbf{X})$ is a sufficient statistic. Thus to show that $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that that

$$P(S(\mathbf{X}) = s \mid T(\mathbf{X} = t) = P(S(\mathbf{X} = s)$$

for all possible value $t \in \tau$. Now,

$$P(S(\mathbf{X}) = s) = \sum_{t\in\tau} = P(S(\mathbf{X}) = s \mid T(\mathbf{X}) = t)P_\theta(T(\mathbf{X}) = t)$$

Furthermore, since $\sum_{t\in\tau} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(S(\mathbf{X}) = s) = \sum_{t\in\tau} = P(S(\mathbf{X}) = s)P_\theta(T(\mathbf{X}) = t)$$

Therefore, if we define the statistic

$$g(t) = P(\mathbf{X}) = s \mid T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s)$$

the above two equations show that

$$E_\theta g(T) = \sum_{t\in\tau} g(t)P_\theta(T(\mathbf{X}) = t) = 0 \quad \text{for all } \theta$$

Since $T(\mathbf{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \tau$ □

Example 11 $X_1, \ldots, X_n \sim N(\mu, 1)$. If $E_\mu[g(\bar{X})] = 0 \quad \forall\mu$

$$\int_{-\infty}^{\infty} g(x)\frac{1}{\sqrt{2\pi n}}e^{-1/2n(x-\mu)^2}dx \quad \forall\mu \quad \text{since} \quad \bar{X} \sim N(\mu, 1/n) \tag{1}$$

If g is a step function then it is easy to see that (1) implies $g = 0$ and one can then draw the conclusion that the result follows for all functions which can be approximated by step functions.

Example 12 Let $X \sim P(\lambda)$. If

$$\begin{aligned} &E_\lambda[g(X)] = 0 \quad \forall\lambda \\ \Rightarrow &\sum_{x=0}^{\infty} g(x)\frac{e^{-\lambda}\lambda^x}{x!} = 0 \quad \forall\lambda \\ \Rightarrow &\sum_{x=0}^{\infty}\left(\frac{g(x)}{x!}\right)\lambda^x = 0 \quad \forall\lambda \end{aligned}$$

i.e. a function of the form $h(\lambda) = \sum_0^\infty au\lambda^u$ is the constant $0 \quad \forall\lambda$ which implies $au = 0 \quad \forall u$ og thus $g(x) = 0$ for $x \in \mathbb{N}$ and therefore $P_\lambda[g(X) = 0] = 1$.

1.5 The Likelihood Principle

1.5.1 Handout

Likelihood functions

Definition 5 Let $X_1, \ldots, X_n$ be random variables with a joint probability density function f_θ, so that $f_\theta(\mathbf{x})$ is defined for $\mathbf{x} \in \mathbf{X}(\Omega) \subset \mathbb{R}^n$. Write $\mathbf{X} = (X_1, \ldots, X_n)' \sim f_\theta$. Given a data vector, $\mathbf{x}$, the **likelihood function** is the function $L_\mathbf{x}(\theta) := f_\theta(\mathbf{x}), \quad \theta \in \Theta$.

Remark 1.2. Note that L and f are "the same" in the sense that if we write $g(\mathbf{x}, \theta) := f_\theta(\mathbf{x})$ then of course $f_\theta(\mathbf{x}) = L_\mathbf{x}(\theta) = g(\mathbf{x}, \theta)$, i.e. both can be viewed as functions with two arguments. However, the point of the definition is to emphasize that the **likelihood is a function of the parameters for a fixed data set**.

Example 13 $X_1, \ldots, X_n \sim U(0, \theta)$ iid.

$$f_\theta(\mathbf{x}) = h_\theta(x_1) \cdots h_\theta(x_n) = \begin{cases} \frac{1}{\theta^n} & 0 \le x_i \le \theta, \quad i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

$$h_\theta(t) = \begin{cases} \frac{1}{\theta} & 0 \le x_i \le \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\text{note} \quad h_\theta(t) = \frac{1}{\theta} I_{[0,\theta]}(t)$$

$$\text{so} \quad f_\theta(\mathbf{x}) = \frac{1}{\theta^n} \prod_{i=1}^n I_{[0,\theta]}(x_i)$$

$$\Rightarrow \quad f_\theta(\mathbf{x}) = \frac{1}{\theta^n} I_{[0,\theta]}(x_{(n)}) I_{[0,\infty[}(x_{(1)})$$

[$0 \le x_i \le \theta$ for all $i \Leftrightarrow x_{(1)} \ge 0$ og $x_{(n)} \le \theta$]

$$L_\mathbf{x}(\theta) = \frac{1}{\theta^n} I_{[0,\theta]}(x_{(n)}) I_{[0,\infty[}(x_{(1)})$$

If $x_{(1)} > 0$ then $x_{(n)} > 0$

$$L_\mathbf{x}(\theta) = \frac{1}{\theta^n} I_{[x_{(n)},\infty[}(\theta)$$

Example 14 Let $\mathbf{X} = (X_1, ..., X_n)$ be a sample of n i.i.d. Poisson random variables with joint pdf $f(\mathbf{x}|\lambda)$. The likelihood function of λ given $\mathbf{X} = \mathbf{x}$ is

$$L(\lambda|\mathbf{x}) = f(\mathbf{x}|\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{\lambda} = \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!} e^{n\lambda}$$

Likelihood principle

The *likelihood principle* states that inference on θ should only be based on the relative value of the likelihood function. In other words, if

$$L_\mathbf{x}(\theta) = \kappa L_\mathbf{y}(\theta), \quad \forall \theta \in \Theta \quad (\kappa \text{ is a constant})$$

then $\mathbf{x}$ og $\mathbf{y}$ should lead to the same inference on θ.

Example 15 The likelihood function provides information on how "likely" a parameter value is, given a set of data.

$$\begin{aligned} X &\sim Bin(n,p), & \theta &= p \\ P[X=x] &= \binom{n}{x} p^x (1-p)^{n-x}, & x &= 0,\dots,n \\ L(p) &= \binom{n}{x} p^x (1-p)^{n-x}, & 0 &\le p \le 1 \end{aligned}$$

$$\begin{aligned} \ln(L(p)) &= \ln\binom{n}{p} + x\ln p + (n-x)\ln(1-p) \\ \frac{d\ln(L(p))}{dp} &= \frac{x}{p} - \frac{n-x}{1-p} = 0 \\ \Rightarrow \quad x(1-p) &= p(n-x) \\ \Rightarrow \quad x - px &= np - xp \\ \Rightarrow \quad p &= \frac{x}{n} \end{aligned}$$

As is typical for the discrete case we can interpret this as the value of p which gives the maximum probability to the measurements which were obtained. This interpretation is **not** correct in the continuous case.

Example 16 Let $X_1,...,X_n \sim n(\theta,\sigma^2)$, iid. Both parameters are unknown and we would like to find maximum likelihood estimators for θ and σ^2. The likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}) &= \prod_{i=1}^{n} f_{x_i}(x_i;\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right) \end{aligned}$$

(The following material is covered in more detail in the next section).

We take notice that it is more convenient to maximize the natural logarithm (written here as log due to convention) of the function instead since

$$\begin{aligned} \log L(\boldsymbol{\theta};\mathbf{x}) &= \log\left((2\pi\sigma^2)^{-\frac{n}{2}}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 \end{aligned}$$

Necessary conditions for a maximum of $\log L$ w.r.t. θ and σ^2 are

$$\frac{\partial \log L(\boldsymbol{\theta};\mathbf{x})}{\partial\theta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) = 0 \tag{2}$$

and

$$\frac{\partial \log L(\boldsymbol{\theta};\mathbf{x})}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\theta)^2 = 0 \tag{3}$$

Using (1) and (2) we can find MLE candidates. From (1) we get

$$\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$

so a MLE candidate for θ is $\hat{\theta} = \bar{x}$ which is the sample mean. Likewise (2) gives

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2$$

thus a MLE candidate for σ^2 is $\hat{\sigma}^2 = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}S^2$ where we have used the MLE candidate for θ. All that is now left to prove is that $\log L$ achieves its maximum at $\hat{\theta}$ and $\hat{\sigma}^2$.

Remember that $\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2 \ \forall a \in \mathbb{R}$ so $\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) \geq \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - a)^2\right) \ \forall a \in \mathbb{R}$. So now we only have to confirm that $\log L$ achieves its maximum w.r.t. σ^2. We look at the second derivative

$$\begin{aligned}\frac{\partial^2 \log L(\boldsymbol{\theta}; \mathbf{x})}{\partial(\sigma^2)^2} &= \frac{n}{2}\frac{n^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - \frac{n^3}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^3}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{2}n^3K - n^3K = -\frac{1}{2}n^3K \leq 0\end{aligned}$$

where $K = (\sum_{i=1}^{n}(x_i - \bar{x})^2)^{-2}$. Thus proving that $\log L$ indeed achieves its maximum at $(\hat{\theta}, \hat{\sigma}^2)$ and it is a global maximum since it's the only critical point of $\log L$ which goes to 0 at the $\pm\infty$ limits.