

# **SDS7102: Linear Models and Extensions**

## Point Estimation

---

Qiang Sun, Ph.D. <[qiang.sun@mbzuai.ac.ae](mailto:qiang.sun@mbzuai.ac.ae)>

These slides are due to Eric Moulines.

August 19, 2025

MBZUAI

# Statistical Model

- Let  $X_1, \dots, X_n$  be random variables (or random vectors) and suppose that we observe  $x_1, \dots, x_n$ , which can be thought of as outcomes of the random variables  $X_1, \dots, X_n$ .
- Suppose that the joint distribution of  $\mathbf{X} = (X_1, \dots, X_n)$  is unknown but belongs to some particular family of distributions. Such a family of distributions is called a *statistical model*.
- It is convenient to index the distributions belonging to a statistical model by a parameter  $\theta$ ;  $\theta$  typically represents the unknown or unspecified part of the model. We can then write

$$\mathbf{X} = (X_1, \dots, X_n) \sim F_\theta \quad \text{for} \quad \theta \in \Theta,$$

where  $F_\theta$  is the joint distribution function of  $\mathbf{X}$  and  $\Theta$  is the set of possible values for the parameter  $\theta$ ; we will call the set  $\Theta$  the parameter space.

## Statistical Model

- In general,  $\theta$  can be either a single real-valued parameter or a vector of parameters; in this latter case, we will often write  $\theta = (\theta_1, \dots, \theta_p)$  to emphasize that we have a vector-valued parameter.
- We write  $P_\theta(A)$ ,  $E_\theta(X)$ , and  $\text{Var}_\theta(X)$  to denote (respectively) probability, expected value, and variance with respect to a distribution with unknown parameter  $\theta$ .
- We usually assume that  $\Theta$  is a subset of some Euclidean space; such a model is often called a **parametric model**.
- Models whose distributions cannot be indexed by a finite dimensional parameter are often (somewhat misleadingly) called **non-parametric models**.

# Identifiability

- For a given parameter  $\theta$  corresponds to a single distribution  $F_\theta$ . However, this does not rule out the possibility that there may exist distinct parameter values  $\theta_1$  and  $\theta_2$  such that  $F_{\theta_1} = F_{\theta_2}$ .
- We often require that a given model, or more precisely, its parametrization be **identifiable**; a model is said to have an **identifiable parametrization** (or to be an identifiable model) if  $F_{\theta_1} = F_{\theta_2}$  implies that  $\theta_1 = \theta_2$ .
- A **nonidentifiable parametrization** can lead to problems in estimation of the parameters in the model.

## Example: Poisson Model

- Suppose that  $X_1, \dots, X_n$  are i.i.d. Poisson random variables with mean  $\lambda$ .
- The joint frequency function of  $\mathbf{X} = (X_1, \dots, X_n)$  is

$$f(\mathbf{x}; \lambda) = \prod_{i=1}^n \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!}$$

for  $x_1, \dots, x_n = 0, 1, 2, \dots$ .

- The parameter space for this parametric model is  $\{\lambda : \lambda > 0\}$ .

## Example: non-parametric and semi-parametric model

- Suppose that  $X_1, \dots, X_n$  are i.i.d. random variables with a continuous distribution function  $F$  that is unknown.
- The parameter space for this model consists of all possible continuous distributions. These distributions cannot be indexed by a finite dimensional parameter and so this model is **non-parametric**.
- We may also assume that  $F(x)$  has a density  $f(x - \theta)$  where  $\theta$  is an unknown parameter and  $f$  is an unknown density function satisfying  $f(x) = f(-x)$ .
- This model is also non-parametric but depends on the real-valued parameter  $\theta$ . (This might be considered a semiparametric model because of the presence of  $\theta$ .)

## Example: linear Gaussian regression

- Suppose that  $X_1, \dots, X_n$  are independent Normal random variables with  $E_{\theta}(X_i) = \beta_0 + \beta_1 t_i + \beta_2 s_i$  (where  $t_1, \dots, t_n$  and  $s_1, \dots, s_n$  are known constants) and  $\text{Var}_{\theta}(X_i) = \sigma^2$ ; the parameter space is

$$\{\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \sigma) : -\infty < \beta_0, \beta_1, \beta_2 < \infty, \sigma > 0\}.$$

- The parametrization for this model is identifiable if, and only if, the vectors

$$z_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, z_1 = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}, \quad \text{and} \quad z_2 = \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}$$

are linearly independent.

# Exponential families

- Suppose that  $X_1, \dots, X_n$  have a joint distribution  $F_\theta$  where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$  is an unknown parameter.
- We say that the family of distributions  $\{F_\theta\}$  is a  $k$ -parameter exponential family if the joint density or joint frequency function of  $(X_1, \dots, X_n)$  is of the form

$$f(\mathbf{x}; \boldsymbol{\theta}) = \exp \left[ \sum_{i=1}^k c_i(\boldsymbol{\theta}) T_i(\mathbf{x}) - d(\boldsymbol{\theta}) + S(\mathbf{x}) \right]$$

for  $\mathbf{x} = (x_1, \dots, x_n) \in A$  where  $A$  does not depend on the parameter  $\boldsymbol{\theta}$ .

- It is important to note that  $k$  need not equal  $p$ , the dimension of  $\boldsymbol{\theta}$ , although, in many cases, they are equal.

## Binomial distribution

- Suppose that  $X$  has a Binomial distribution with parameters  $n$  and  $\theta$  where  $\theta$  is unknown.
- The frequency function of  $X$  is

$$\begin{aligned}f(x; \theta) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \\&= \exp \left[ \ln \left( \frac{\theta}{1 - \theta} \right) x + n \ln(1 - \theta) + \ln \binom{n}{x} \right]\end{aligned}$$

for  $x \in A = \{0, 1, \dots, n\}$ .

- The distribution of  $X$  has a one-parameter exponential family.

## Gamma Distribution

- Suppose that  $X_1, \dots, X_n$  are i.i.d. Gamma random variables with unknown shape parameter  $\alpha$  and unknown scale parameter  $\lambda$ .
- The joint density function of  $\mathbf{X} = (X_1, \dots, X_n)$  is

$$f(\mathbf{x}; \alpha, \lambda)$$

$$\begin{aligned}&= \prod_{i=1}^n \left[ \frac{\lambda^\alpha x_i^{\alpha-1} \exp(-\lambda x_i)}{\Gamma(\alpha)} \right] \\&= \exp \left[ (\alpha - 1) \sum_{i=1}^n \ln(x_i) - \lambda \sum_{i=1}^n x_i + n\alpha \ln(\lambda) - n \ln(\Gamma(\alpha)) \right]\end{aligned}$$

(for  $x_1, \dots, x_n > 0$ ) and so the distribution of  $\mathbf{X}$  is a two-parameter exponential family.

# Gaussian distribution

- Suppose that  $X_1, \dots, X_n$  are i.i.d. Normal random variables with mean  $\theta$  and variance  $\theta^2$  where  $\theta > 0$ .
- The joint density function of  $(X_1, \dots, X_n)$  is

$$f(\mathbf{x}; \theta)$$

$$\begin{aligned} &= \prod_{i=1}^n \left[ \frac{1}{\theta \sqrt{2\pi}} \exp \left( -\frac{1}{2\theta^2} (x_i - \theta)^2 \right) \right] \\ &= \exp \left[ -\frac{1}{2\theta^2} \sum_{i=1}^n x_i^2 + \frac{1}{\theta} \sum_{i=1}^n x_i - \frac{n}{2} (1 + \ln(\theta^2) + \ln(2\pi)) \right], \end{aligned}$$

and so  $A = \mathbb{R}^n$ . Note that this is a two-parameter exponential family despite the fact that the parameter space is one-dimensional.

## Poisson distribution

Suppose that  $X_1, \dots, X_n$  are independent Poisson random variables with  $E(X_i) = \exp(\alpha + \beta t_i)$  where  $t_1, \dots, t_n$  are known constants. Setting  $\mathbf{X} = (X_1, \dots, X_n)$ , the joint frequency function of  $\mathbf{X}$  is

$$f(\mathbf{x}; \alpha, \beta)$$

$$\begin{aligned} &= \prod_{i=1}^n \left[ \frac{\exp(-\exp(\alpha + \beta t_i)) \exp(\alpha x_i + \beta x_i t_i)}{x_i!} \right] \\ &= \exp \left[ \alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i t_i + \sum_{i=1}^n \exp(\alpha + \beta t_i) - \sum_{i=1}^n \ln(x_i!) \right]. \end{aligned}$$

This is a two-parameter exponential family model; the set  $A$  is simply  $\{0, 1, 2, 3, \dots\}^n$ .

## Uniform distribution

Suppose that  $X_1, \dots, X_n$  are i.i.d. Uniform random variables on the interval  $[0, \theta]$ . The joint density function of  $\mathbf{X} = (X_1, \dots, X_n)$  is

$$f(\mathbf{x}; \theta) = \frac{1}{\theta^n} \quad \text{for } 0 \leq x_1, \dots, x_n \leq \theta$$

The region on which  $f(\mathbf{x}; \theta)$  is positive clearly depends on  $\theta$  and so this model is not an exponential family model.

# Mean and variance of exponential distribution

## Proposition

Suppose that  $\mathbf{X} = (X_1, \dots, X_n)$  has a one-parameter exponential family distribution with density or frequency function

$$f(\mathbf{x}; \theta) = \exp[c(\theta)T(\mathbf{x}) - d(\theta) + S(\mathbf{x})]$$

for  $\mathbf{x} \in A$  where

- (a) the parameter space  $\Theta$  is open,
- (b)  $c(\theta)$  is a one-to-one function on  $\Theta$ ,
- (c)  $c(\theta), d(\theta)$  are twice differentiable functions on  $\Theta$ .

Then

$$E_\theta[T(\mathbf{X})] = \frac{d'(\theta)}{c'(\theta)}$$

and  $\text{Var}_\theta[T(\mathbf{X})] = \frac{d''(\theta)c'(\theta) - d'(\theta)c''(\theta)}{[c'(\theta)]^3}$

# Statistics

- Suppose that the model for  $\mathbf{X} = (X_1, \dots, X_n)$  has a parameter space  $\Theta$ .
- Since the true value of the parameter  $\theta$  (or, equivalently, the true distribution of  $\mathbf{X}$ ) is unknown, we would like to summarize the available information in  $\mathbf{X}$  without losing too much information about the unknown parameter  $\theta$ .
- At this point, we are not interested in estimating  $\theta$  per se but rather in determining how to best use the information in  $\mathbf{X}$ .

# Statistics

- Define a **statistic**  $T = T(\mathbf{X})$  to be a function of  $\mathbf{X}$  that does not depend on any unknown parameter; that is, the statistic  $T$  depends only on observable random variables and known constants.
- A **statistic** can be real- or vector-valued.

## Example

$T(\mathbf{X}) = \bar{X} = n^{-1} \sum_{i=1}^n X_i$ . Since  $n$  (the sample size) is known,  $T$  is a statistic.

## Example

$T(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)})$  where  $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$  are the order statistics of  $\mathbf{X}$ . Since  $T$  depends only on the values of  $\mathbf{X}$ ,  $T$  is a statistic.

# Statistics

- It is important to note that any statistic is itself a random variable and so has its own probability distribution; this distribution **may or may not** depend on the parameter  $\theta$ .
- Ideally, a statistic  $T = T(\mathbf{X})$  should contain as much information about  $\theta$  as  $\mathbf{X}$  does.
- However, this raises several questions.
  - For example, how does one determine if  $T$  and  $\mathbf{X}$  contain the same information about  $\theta$  ?

# Ancillary Statistics

## **Definition (Ancillary statistics)**

A statistic  $T$  is an ancillary statistic (for  $\theta$ ) if its distribution is independent of  $\theta$ ; that is, for all  $\theta \in \Theta$ ,  $T$  has the same distribution.

## Example: ancillary statistics in normal sample

- Suppose that  $X_1$  and  $X_2$  are independent Normal random variables each with mean  $\mu$  and variance  $\sigma^2$  (where  $\sigma^2$  is known).
- Let  $T = X_1 - X_2$ ; then  $T$  has a Normal distribution with mean 0 and variance  $2\sigma^2$ . Thus  $T$  is ancillary for the unknown parameter  $\mu$ .
- However, if both  $\mu$  and  $\sigma^2$  were unknown,  $T$  would not be ancillary for  $\theta = (\mu, \sigma^2)$ . (The distribution of  $T$  depends on  $\sigma^2$  so  $T$  contains some information about  $\sigma^2$ .)

## Example: ancillarity range w.r.t translation parameter

- Suppose that  $X_1, \dots, X_n$  are i.i.d. random variables with density function

$$f(x; \mu, \eta) = \frac{1}{2\eta} \quad \text{for } \mu - \eta \leq x \leq \mu + \eta.$$

- Define a statistic  $R = X_{(n)} - X_{(1)}$ , which is the sample range of  $X_1, \dots, X_n$ .
- The density function of  $R$  is

$$f_R(r) = \frac{n(n-1)r^{n-2}}{(2\eta)^{n-1}} \left(1 - \frac{r}{2\eta}\right) \quad \text{for } 0 \leq r \leq 2\eta$$

which depends on  $\eta$  but not  $\mu$ . Thus  $R$  is ancillary for  $\mu$ .

## Uniform distribution

Suppose that  $X_1, \dots, X_n$  are i.i.d. Uniform random variables on the interval  $[0, \theta]$  where  $\theta > 0$  is an unknown parameter. Define two statistics,  $S = \min(X_1, \dots, X_n)$  and  $T = \max(X_1, \dots, X_n)$ . The density of  $S$  is

$$f_S(x; \theta) = \frac{n}{\theta} \left(1 - \frac{x}{\theta}\right)^{n-1} \quad \text{for } 0 \leq x \leq \theta,$$

while the density of  $T$  is

$$f_T(x; \theta) = \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} \quad \text{for } 0 \leq x \leq \theta.$$

- Note that the densities of both  $S$  and  $T$  depend on  $\theta$  and so neither is ancillary for  $\theta$ . However, as  $n$  increases, it becomes clear that the density of  $S$  is concentrated around 0 for all possible values of  $\theta$  while the density of  $T$  is concentrated around  $\theta$ .

# Sufficiency

- The first mention of sufficiency was made by Fisher (1920) in which he considered the estimation of the variance  $\sigma^2$  of a Normal distribution based on i.i.d. observations  $X_1, \dots, X_n$ .
- In particular, he considered estimating  $\sigma^2$  based on the statistics

$$T_1 = \sum_{i=1}^n |X_i - \bar{X}| \quad \text{and} \quad T_2 = \sum_{i=1}^n (X_i - \bar{X})^2$$

where  $\bar{X}$  is the average of  $X_1, \dots, X_n$ .

- Fisher showed that the distribution of  $T_1$  conditional on  $T_2 = t$  does not depend on the parameter  $\sigma$  while the distribution of  $T_2$  conditional on  $T_1 = t$  does depend on  $\sigma$ .
- He concluded that all the information about  $\sigma^2$  in the sample was contained in the statistic  $T_2$  and that any estimate of  $\sigma^2$  should be based on  $T_2$ ;
- Any estimate of  $\sigma^2$  based on  $T_1$  could be improved by using the information in  $T_2$  while  $T_2$  could not be improved by using  $T_1$ .

# Sufficient statistics

## Definition (Sufficient statistics)

A statistic  $T = T(\mathbf{X})$  is a sufficient statistic for a parameter  $\theta$  if for all sets  $A$ ,  $P_\theta[\mathbf{X} \in A | T = t]$  is independent of  $\theta$  for all  $t$  in the range of  $T$ .

- Sufficient statistics are not unique; from the definition of sufficiency, it follows that if  $g$  is a one-to-one function over the range of the statistic  $T$  then  $g(T)$  is also sufficient.
- It also follows that if  $T$  is sufficient for  $\theta$  then the distribution of any other statistic  $S = S(\mathbf{X})$  conditional on  $T$  is independent of  $\theta$ .

## Sufficient statistics in binomial model

- Suppose that  $X_1, \dots, X_k$  are independent random variables where  $X_i$  has a Binomial distribution with parameters  $n_i$  (known) and  $\theta$  (unknown).
- Let  $T = X_1 + \dots + X_k$ ;  $T$  will also have a Binomial distribution with parameters  $m = n_1 + \dots + n_k$  and  $\theta$ .
- Show that  $T$  is sufficient.

# Neyman factorization Lemma

## Theorem (Neyman Factorization Criterion)

Suppose that  $\mathbf{X} = (X_1, \dots, X_n)$  has a joint density or frequency function  $f(\mathbf{x}; \theta)$  ( $\theta \in \Theta$ ). Then  $T = T(\mathbf{X})$  is sufficient for  $\theta$  if, and only if,

$$f(\mathbf{x}; \theta) = g(T(\mathbf{x}); \theta)h(\mathbf{x}).$$

(Both  $T$  and  $\theta$  can be vector-valued.)

## Sufficiency in uniform model

Suppose that  $X_1, \dots, X_n$  are i.i.d. random variables with density function

$$f(x; \theta) = \frac{1}{\theta} \quad \text{for } 0 \leq x \leq \theta$$

- Show that  $X_{(n)}$  is sufficient.

## Sufficient statistics in exponential model

Suppose that  $\mathbf{X} = (X_1, \dots, X_n)$  have a distribution belonging to a  $k$ -parameter exponential family with joint density or frequency function satisfying

$$f(\mathbf{x}; \theta) = \exp \left[ \sum_{i=1}^k c_i(\theta) T_i(\mathbf{x}) - d(\theta) + S(\mathbf{x}) \right] I(\mathbf{x} \in A)$$

- Show that the statistic

$$T = (T_1(\mathbf{X}), \dots, T_k(\mathbf{X}))$$

is sufficient for  $\theta$ .

## Minimal sufficient statistics

- There are two notions of what is meant by the "best possible" reduction of the data.
- The first of these is **minimal sufficiency**; a sufficient statistic  $T$  is minimal sufficient if for any other sufficient statistic  $S$ , there exists a function  $g$  such that  $T = g(S)$ .
- Thus a minimal sufficient statistic is the sufficient statistic that represents the maximal reduction of the data that contains as much information about the unknown parameter as the data itself.

## Complete sufficient statistics

- A second (and stronger) notion is completeness. If  $\mathbf{X} \sim F_\theta$  then a statistic  $T = T(\mathbf{X})$  is **complete** if  $E_\theta(g(T)) = 0$  for all  $\theta \in \Theta$  implies that  $P_\theta(g(T) = 0) = 1$  for all  $\theta \in \Theta$ .
- In particular, if  $T$  is complete then  $g(T)$  is ancillary for  $\theta$  only if  $g(T)$  is constant; thus a complete statistic  $T$  contains no ancillary information.