

# **IB Statistics**

Ishan Nath, Lent 2023

Based on Lectures by Dr. Sergio Bacallado

February 4, 2023

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Probability Review . . . . .	2
1.2	Moment Generating Functions . . . . .	4
1.3	Limit Theorems . . . . .	4
1.4	Conditioning . . . . .	5
1.5	Change of Variables . . . . .	6
1.6	Important Distributions . . . . .	6
<b>2</b>	<b>Estimation</b>	<b>7</b>
2.1	Bias-Variance Decomposition . . . . .	8
2.2	Sufficiency . . . . .	9
2.3	Minimal Sufficiency . . . . .	11
2.4	Rao-Blackwell Theorem . . . . .	13
2.5	Maximum Likelihood Estimation . . . . .	15
2.6	Confidence Intervals . . . . .	18
2.7	Interpreting Confidence Intervals . . . . .	21
<b>3</b>	<b>Bayesian Inference</b>	<b>22</b>
	<b>Index</b>	<b>23</b>

# 1 Introduction

*Statistics* is the science of making informed decisions. It can include:

- Design of experiments,
- Graphical exploration of data,
- Formal statistical inference (part of Decision theory),
- Communication of results.

Let  $X_1, X_2, \dots, X_n$  be independent observations from a distribution  $f(x \mid \theta)$ , with parameter  $\theta$ . We wish to make inferences about the value of  $\theta$  from  $X_1, X_2, \dots, X_n$ . Such inference can include:

- Estimating  $\theta$ ,
- Quantifying uncertainty in estimates,
- Testing a hypothesis about  $\theta$ .

## 1.1 Probability Review

Let  $\Omega$  be the *sample space* of outcomes in an experiment. A measurable subset of  $\Omega$  is called an *event*. We denote the set of events as  $\mathcal{F}$ .

A function  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$  is called a *probability measure* if:

- $\mathbb{P}(\emptyset) = 0$ ,
- $\mathbb{P}(\Omega) = 1$ ,
- $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$ , if  $(A_i)$  are disjoint and countable.

A *random variable* is a (measurable) function  $X : \Omega \rightarrow \mathbb{R}$ .

The *distribution function* of  $X$  is

$$F_X(x) = \mathbb{P}(X \leq x).$$

A *discrete random variable* takes values in a countable subset  $E \subset \mathbb{R}$ , and its *probability mass function* or pmf is  $p_X(x) = \mathbb{P}(X = x)$ .

We say  $X$  has *continuous* distribution if it has a *probability density function* or pdf, satisfying

$$\mathbb{P}(X \in A) = \int_A f_X(x) \, dx,$$

for any measurable  $A$ . The *expectation* of  $X$  is defined

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in X} x \cdot p_X(x) & X \text{ discrete,} \\ \int x \cdot f_X(x) dx & X \text{ continuous.} \end{cases}$$

If  $g : \mathbb{R} \rightarrow \mathbb{R}$ , then

$$\mathbb{E}[g(x)] = \int g(x) f_X(x) dx.$$

The *variance* of  $X$  is

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

We say that  $X_1, X_2, \dots, X_n$  are *independent* if for all  $x_1, x_2, \dots, x_n$ ,

$$\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n).$$

If the variables have probability density functions, then

$$f_X(x) = \prod_{i=1}^n f_{X_i}(x_i),$$

where  $X$  is the vector of variables  $(X_1, \dots, X_n)$  and  $x$  is the vector  $(x_1, \dots, x_n)$ .

Importantly, if  $a_1, \dots, a_n \in \mathbb{R}$ ,

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n].$$

Moreover,

$$\text{Var}(a_1 X_1 + \cdots + a_n X_n) = \sum_{i,j} a_i a_j \text{Cov}(X_i, X_j).$$

Here the *covariance* of  $X_i$  and  $X_j$  is

$$\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])].$$

If  $X = (X_1, \dots, X_n)^T$  and  $\mathbb{E}[X] = (\mathbb{E}[X_1], \dots, \mathbb{E}[X_n])$ , then the linearity of expectation can be rewritten as

$$\mathbb{E}[a^T X] = a^T \mathbb{E}[X],$$

and moreover

$$\text{Var}(a^T X) = a^T \text{Var}(X) a,$$

where  $\text{Var}(X)$  is the *covariance matrix*:  $(\text{Var}(X))_{ij} = \text{Cov}(X_i, X_j)$ .

## 1.2 Moment Generating Functions

The *moment generating function* of a variable  $X$  is

$$M_X(t) = \mathbb{E}[e^{tx}].$$

This may only exist for  $t$  in some neighbourhood of 0. The important properties of MGFs is that

$$\mathbb{E}[X^n] = \frac{d^n}{dt^n} M_X(0),$$

and from this we obtain  $M_X = M_Y \iff F_x = F_y$ .

MGFs also make it easy to find the distribution function of sums of iid variables.

### Example 1.1.

Let  $X_1, \dots, X_n$  be iid Poisson( $\mu$ ). Then

$$\begin{aligned} M_{X_1}(t) &= \mathbb{E}[e^{tX_1}] = \sum_{x=0}^{\infty} e^{tx} \cdot \frac{e^{-\mu} \mu^x}{x!} \\ &= e^{-\mu} \sum_{x=0}^{\infty} \frac{(e^t \mu)^x}{x!} = e^{-\mu} e^{\mu \exp(t)} = e^{-\mu(1-e^t)}. \end{aligned}$$

If  $S_n = X_1 + \dots + X_n$ , then

$$\begin{aligned} M_{S_n}(t) &= \mathbb{E}[e^{t(X_1 + \dots + X_n)}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] \\ &= e^{-\mu(1-e^t)n} \end{aligned}$$

This is the same as a Poisson( $\mu n$ ) MGF, so  $S_n \sim \text{Poisson}(\mu \cdot n)$ .

## 1.3 Limit Theorems

We list some important limit theorems, starting with the *weak law of large numbers* (WLLN). This says if  $X_1, \dots, X_n$  are iid with  $\mathbb{E}[X_1] = \mu$ , then let  $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  be the sample mean. WLLN says that for all  $\varepsilon > 0$ ,

$$\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \rightarrow 0,$$

as  $n \rightarrow \infty$ .

The *strong law of large numbers* (SLLN) says a stronger result, namely

$$\mathbb{P}(\overline{X}_n \rightarrow \mu) = 1,$$

i.e.  $\overline{X_n}$  converges to  $\mu$  almost surely.

The *central limit theorem* is another important limit theorem. If we take

$$Z_n = \frac{\sqrt{n}(\overline{X_n} - \mu)}{\sigma},$$

where  $\sigma^2 = \text{Var}(X_i)$ , then  $Z_n$  is “approximately”  $N(0, 1)$  as  $n \rightarrow \infty$ .

What this means is that  $\mathbb{P}(Z_n \leq z) \rightarrow \Phi(z)$  as  $n \rightarrow \infty$  for all  $z \in \mathbb{R}$ , where  $\Phi$  is the distribution function of a  $N(0, 1)$  variable.

## 1.4 Conditioning

Let  $X$  and  $Y$  be discrete random variables. Their *joint pmf* is

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

The *marginal pmf* is

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in Y} p_{X,Y}(x, y).$$

The *conditional pmf* of  $X$  given  $Y = y$  is

$$p_{X|Y}(x | y) = \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

This is defined to be 0 if  $p_Y(y) = 0$ .

For continuous random variables  $X, Y$ , the *joint pdf*  $f_{X,Y}$  has

$$\mathbb{P}(X \leq x', y \leq y') = \int_{-\infty}^{x'} \int_{-\infty}^{y'} f_{X,Y}(x, y) \, dy \, dx.$$

The *marginal pdf* of  $Y$  is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx.$$

The *conditional pdf* of  $X$  given  $Y$  is

$$f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

The *conditional expectation* is given by

$$\mathbb{E}[X | Y] = \begin{cases} \sum_x x \cdot p_{X|Y}(x | y) & X, Y \text{ discrete,} \\ \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x | y) dx & X, Y \text{ continuous.} \end{cases}$$

This is a random variable, which is a function of  $Y$ . The *tower property* says that

$$\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X].$$

Hence we can write the variance of  $X$  as follows:

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \mathbb{E}[\mathbb{E}[X^2 | Y]] - (\mathbb{E}[\mathbb{E}[X | Y]])^2 \\ &= \mathbb{E}[\mathbb{E}[X^2 | Y] - (\mathbb{E}[X | Y])^2] + \mathbb{E}[\mathbb{E}[X | Y]^2] - \mathbb{E}[\mathbb{E}[X | Y]]^2 \\ &= \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]). \end{aligned}$$

## 1.5 Change of Variables

The *change of variables* formula is as follows:

Let  $(x, y) \mapsto (u, v)$  be a differentiable bijection. Then,

$$\begin{aligned} f_{U,V}(u, v) &= f_{X,Y}(x(u, v), y(u, v)) \cdot |\det J|, \\ J &= \frac{\partial(x, y)}{\partial(u, v)} = \begin{pmatrix} \partial x / \partial u & \partial x / \partial v \\ \partial y / \partial u & \partial y / \partial v \end{pmatrix}. \end{aligned}$$

## 1.6 Important Distributions

$X \sim \text{Negbin}(k, p)$  if  $X$  models the time in successive iid  $\text{Ber}(p)$  trials to achieve  $k$  successes. If  $k = 1$ , this is the same as a geometric distribution.

$X \sim \text{Poisson}(\lambda)$  is the limit of  $\text{Bin}(n, \lambda/n)$  random variables, as  $n \rightarrow \infty$ .

If  $X_i \sim \Gamma(\alpha_i, \lambda)$  for  $i = 1, \dots, n$  with  $X_1, \dots, X_n$  independent, then if  $S_n = X_1 + \dots + X_n$ ,

$$M_{S_n}(t) = \prod_{i=1}^n M_{X_i}(t) = \left( \frac{\lambda}{\lambda - 1} \right)^{\alpha_1 + \dots + \alpha_n}$$

which is the mgf of a  $\Gamma(\sum \alpha_i, \lambda)$  random variable. Hence  $S_n \sim \Gamma(\sum \alpha_i, \lambda)$ .

Also, if  $X \sim \Gamma(a, \lambda)$ , then for any  $b \in (0, \infty)$ ,  $bX \sim \Gamma(a, \lambda/b)$ .

Special cases of the Gamma distribution include  $\Gamma(1, \lambda) = \text{Exp}(\lambda)$ , and  $\Gamma(\frac{k}{2}, \frac{1}{2}) = \chi_k^2$ , the Chi-squared distribution with  $k$  degrees of freedom. This can be thought of as the sum of  $k$  independent squared  $N(0, 1)$  random variables.

## 2 Estimation

Suppose we observe data  $X_1, X_2, \dots, X_n$ , which are iid from some pdf (or pmf)  $f_X(x | \theta)$ , with  $\theta$  unknown. We let  $X = (X_1, \dots, X_n)$ .

**Definition 2.1.** An *estimator* is a statistic or a function of the data  $T(X) = \hat{\theta}$ , which we use to approximate the true parameter  $\theta$ . The distribution of  $T(X)$  is called the *sampling distribution*.

### Example 2.1.

If  $X_1, \dots, X_n$  are iid  $N(\mu, 1)$ , we can define an estimator for the mean as

$$\hat{\mu} = T(X) = \frac{1}{n} \sum_{i=1}^n X_i.$$

The sampling distribution of  $\hat{\mu}$  is  $N(\mu, \frac{1}{n})$ .

**Definition 2.2.** The *bias* of  $\hat{\theta} = T(X)$  is

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta.$$

*Remark.* In general, the bias is a function of  $\theta$ , even if the notation  $\text{bias}(\hat{\theta})$  does not make that explicit.

**Definition 2.3.** We say that  $\hat{\theta}$  is *unbiased* if  $\text{bias}(\hat{\theta}) = 0$  for all  $\theta \in \Theta$ .

### Example 2.2.

Our previous estimator

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

is unbiased because  $\mathbb{E}_\mu[\hat{\mu}] = \mu$  for all  $\mu \in \mathbb{R}$ .

**Definition 2.4.** The *mean squared error* (mse) of  $\hat{\theta}$  is

$$\text{mse}(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2].$$

Like the bias, the mean squared error of  $\hat{\theta}$  is a function of  $\theta$ .



## 2.1 Bias-Variance Decomposition

We can write the mean squared error as

$$\begin{aligned}\text{mse}(\hat{\theta}) &= \mathbb{E}_{\theta}[(\hat{\theta} - \theta)^2] = \mathbb{E}_{\theta}[(\hat{\theta} - \mathbb{E}_{\theta}[\hat{\theta}] + \mathbb{E}_{\theta}[\hat{\theta}] - \theta)^2] \\ &= \text{Var}_{\theta}(\hat{\theta}) + \text{bias}^2(\hat{\theta}) + \underbrace{2[\mathbb{E}_{\theta}[(\hat{\theta} - \mathbb{E}_{\theta}[\hat{\theta}])]]}_{0}(\mathbb{E}_{\theta}[\hat{\theta}] - \theta).\end{aligned}$$

The two terms on the right hand side are non-negative, so there is a trade off between bias and variance.

### Example 2.3.

Let  $X \sim \text{Bin}(n, \theta)$ , where  $n$  is known, and we wish to estimate  $\theta$ . The standard estimator is

$$T_u = \frac{X}{n}, \quad \mathbb{E}_{\theta}[T_u] = \frac{\mathbb{E}_{\theta}[X]}{n} = \theta.$$

Hence  $T_u$  is unbiased. We can also calculate the mean squared error as

$$\text{mse}(T_u) = \text{Var}_{\theta}(T_u) = \frac{\text{Var}_{\theta}(X)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}.$$

Consider a second estimator

$$T_B = \frac{X+1}{n+2} = w\frac{X}{n} + (1-w)\frac{1}{2},$$

for  $w = \frac{n}{n+2}$ . In this case  $T_B$  is interpolating between our unbiased estimator, and the constant estimator. The bias of  $T_B$  is

$$\text{bias}(T_B) = \mathbb{E}_{\theta}[T_B] - \theta = \mathbb{E}\left[\frac{X+1}{n+2}\right] - \theta = \frac{1}{n+2} - \frac{2}{n+2}\theta.$$

This is not equal to zero for all but one value of  $\theta$ . Hence,  $T_B$  is biased. We can also calculate the variance

$$\begin{aligned}\text{Var}_{\theta}(T_B) &= \frac{1}{(n+2)^2}n\theta(1-\theta) - w^2\frac{\theta(1-\theta)}{n}, \\ \text{mse}(T_B) &= \text{Var}_{\theta}(T_B) + \text{bias}^2(T_B) \\ &= w^2\frac{\theta(1-\theta)}{n} + (1-w)^2\left(\frac{1}{2} - \theta\right)^2.\end{aligned}$$

Hence the mse of the biased estimator is a weighted average of the mse of the unbiased estimator, and a parabola. For  $\theta$  around  $1/2$ , the biased estimator has a lower mse than the unbiased estimator.

The message here is that our prior judgements about  $\theta$  affect our choice of estimator, and unbiasedness is not always desirable.

#### Example 2.4.

Suppose  $X \sim \text{Poisson}(\lambda)$ . We wish the estimate  $\theta = \mathbb{P}(X = 0)^2 = e^{-2\lambda}$ . For an estimator  $T(X)$  to be unbiased, we must have for all  $\lambda$ ,

$$\begin{aligned}\mathbb{E}_\lambda[\hat{\theta}] &= \sum_{x=0}^{\infty} T(x) \frac{e^{-\lambda} \lambda^x}{x!} = e^{-2\lambda} = \theta \\ \iff \sum_{x=0}^{\infty} T(x) \frac{\lambda^x}{x!} &= e^{-\lambda} = \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!}.\end{aligned}$$

For this to hold for all  $\lambda \geq 0$ , we should take  $T(X) = (-1)^X$ . But this estimator makes no sense.

## 2.2 Sufficiency

Suppose  $X_1, \dots, X_n$  are iid random variables from a distribution with pdf (or pmf)  $f_X(\cdot | \theta)$ . Let  $X = (X_1, \dots, X_n)$ .

The question is: is there a statistic  $T(X)$  which contains all the information in  $X$  needed to estimate  $\theta$ ?

**Definition 2.5.** A statistic  $T$  is *sufficient* for  $\theta$  if the conditional distribution of  $X$  given  $T(X)$  does not depend on  $\theta$ .

Note  $\theta$  and  $T(X)$  may be vector-valued.

#### Example 2.5.

Let  $X_1, \dots, X_n$  be iid  $\text{Ber}(\theta)$  for  $\theta \in [0, 1]$ . Then,

$$f_X(\cdot | \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{x_1 + \dots + x_n} (1 - \theta)^{n - x_1 - \dots - x_n}.$$

This only depends on  $X$  through

$$T(X) = \sum_{i=1}^n x_i.$$

Indeed, for  $x$  with  $x_1 + \cdots + x_n = t$ ,

$$\begin{aligned} f_{X|T=t}(x | T(x) = t) &= \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(X) = t)} = \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T(x) = t)} \\ &= \frac{\theta^{x_1 + \cdots + x_n} (1 - \theta)^{n - x_1 - \cdots - x_n}}{\binom{n}{t} \theta^t (1 - \theta)^{n - t}} = \binom{n}{t}^{-1}, \end{aligned}$$

and otherwise this probability is 0. As this doesn't depend on  $\theta$ ,  $T(X)$  is sufficient for  $\theta$ .

**Theorem 2.1** (Factorization criterion).  *$T$  is sufficient for  $\theta$  if and only if*

$$f_X(x | \theta) = g(T(x), \theta) \cdot h(x),$$

for suitable functions  $g, h$ .

**Proof:** We only do the discrete case.

Suppose that  $f_X(x | \theta) = g(T(x), \theta)h(x)$ . If  $T(x) = t$ , then

$$\begin{aligned} f_{X|T=t}(x | T = t) &= \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(X) = t)} \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{T(x')=t} g(T(x'), \theta)h(x')} \\ &= \frac{g(t, \theta)}{g(t, \theta)} \cdot \frac{h(x)}{\sum_{T(x')=t} h(x')}. \end{aligned}$$

This doesn't depend on  $\theta$ , so  $T(X)$  is sufficient. Conversely, if  $T(X)$  is sufficient, then

$$\begin{aligned} \mathbb{P}_\theta(X = x) &= \mathbb{P}_\theta(X = x, T(X) = t) \\ &= \underbrace{\mathbb{P}_\theta(T(X) = t)}_{g(t, \theta)} \cdot \underbrace{\mathbb{P}_\theta(X = x | T(X) = t)}_{h(x)}. \end{aligned}$$

Therefore the pmf of  $X$  factorizes.

### Example 2.6.

Return to our example from before, where  $X_1, \dots, X_n$  are iid  $\text{Ber}(\theta)$ . Then

$$f_X(x | \theta) = \theta^{x_1 + \cdots + x_n} (1 - \theta)^{n - x_1 - \cdots - x_n}.$$

Hence if we take  $g(t, \theta) = \theta^t(1 - \theta)^{n-t}$ , and  $h(x) = 1$ , we immediately get that  $T(X) = \sum x_i$  is sufficient.

**Example 2.7.**

Let  $X_1, \dots, X_n$  be iid  $U([0, \theta])$ , for  $\theta > 0$ . Then,

$$\begin{aligned} f_X(x \mid \theta) &= \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}(X_i \in [0, \theta]) \\ &= \frac{1}{\theta^n} \underbrace{\mathbb{1}(\max_i x_i \leq \theta)}_{g(T(x), \theta)} \underbrace{\mathbb{1}(\min_i x_i \geq 0)}_{h(x)}. \end{aligned}$$

Hence  $T(x) = \max_i x_i$  is a sufficient statistic for  $\theta$ .

### 2.3 Minimal Sufficiency

Sufficient statistics are not unique. Indeed, any one-to-one function of a sufficient statistic is also sufficient. Also  $T(X) = X$  is always sufficient, but not very useful.

**Definition 2.6.** A sufficient statistic  $T$  is *minimal sufficient* if it is a function of any other sufficient statistic, so if  $T'$  is also sufficient, then

$$T'(x) = T'(y) \implies T(x) = T(y),$$

for all  $x, y$  in our space.

By this definition, any two minimal sufficient statistics  $T, T'$  are in bijection with each other, so

$$T(x) = T(y) \iff T'(x) = T'(y).$$

**Theorem 2.2.** Suppose that  $T(X)$  is a statistic such that

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)}$$

is constant as a function of  $\theta$ , if and only if  $T(x) = T(y)$ . Then  $T$  is minimal sufficient.

Let  $x \stackrel{1}{\sim} y$  if

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)}$$

is constant in  $\theta$ . It is easy to check that  $\sim^1$  is an equivalence relation.

Similarly, for a given statistic  $T$ ,  $x \sim^2 y$  if  $T(x) = T(y)$  defines another equivalence relation.

The condition of the theorem says that  $\sim^1$  and  $\sim^2$  are the same for minimal sufficient statistics.

*Remark.* We can always construct a statistic  $T$  which is constant on the equivalence classes of  $\sim^1$ , which by the theorem is minimal sufficient.

**Proof:** For any value of  $T$ , let  $z_t$  be a representative from the equivalence class

$$\{x \mid T(x) = t\}.$$

Then,

$$f_X(x \mid \theta) = f_X(z_{T(x)} \mid \theta) \frac{f_X(x, \theta)}{f_X(z_{T(x)}, \theta)}.$$

This is exactly in the form  $g(T(x), \theta)h(x)$ , so by the factorization criterion  $T$  is sufficient.

To prove that  $T$  is minimal, take any other sufficient statistic  $S$ . We want to show that if  $S(x) = S(y)$ , then  $T(x) = T(y)$ .

By the factorization criterion, there are functions  $g_s, h_s$  such that

$$f_X(x, \theta) = g_s(S(x), \theta)h_s(x).$$

Suppose  $S(x) = S(y)$ . Then the ratio

$$\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)} = \frac{g_s(S(x), \theta)h_s(x)}{g_s(S(y), \theta)h_s(y)} = \frac{h_s(x)}{h_s(y)},$$

is independent of  $\theta$ . Hence  $x \sim^1 y$ . By the hypothesis, we get that  $T(x) = T(y)$ .

*Remark.* Sometimes the range of  $X$  depends on  $\theta$ . In this case we can interpret

$$\frac{f_X(x \mid \theta)}{f_Y(y \mid \theta)} \text{ constant in } \theta,$$

to mean that

$$f_X(x \mid \theta) = c(x, y)f_Y(y \mid \theta),$$

for some function  $c$  which does not depend on  $\theta$ .

**Example 2.8.**

Suppose that  $X_1, \dots, X_n$  are iid  $N(\mu, \sigma^2)$ , with parameters  $(\mu, \sigma^2)$  unknown. Then,

$$\begin{aligned} \frac{f_X(x \mid t)}{f_X(y \mid t)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2)}{(2\pi\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2} \sum (y_i - \mu)^2)} \\ &= \exp \left[ -\frac{1}{2\sigma^2} \left( \sum x_i^2 - \sum y_i^2 \right) + \frac{\mu}{\sigma^2} \left( \sum x_i - \sum y_i \right) \right]. \end{aligned}$$

Hence if  $\sum x_i^2 = \sum y_i^2$  and  $\sum x_i = \sum y_i$ , this ratio does not depend on  $(\mu, \sigma^2)$ . The converse is also true: if the ratio does not depend on  $(\mu, \sigma^2)$ , then we must have  $\sum x_i^2 = \sum y_i^2$  and  $\sum x_i = \sum y_i$ . By the theorem,  $T(x) = (\sum x_i^2, \sum x_i)$  is minimal sufficient.

Recall that bijections of  $T$  are also minimal sufficient. A more common way of expressing a minimal sufficient statistic in this model is  $S(X) = (\bar{X}, S_{xx})$ , where

$$\bar{X} = \frac{1}{n} \sum_i X_i, \quad S_{xx} = \sum_i (X_i - \bar{X})^2.$$

In this example,  $(\mu, \sigma^2)$  and  $T(X)$  are both 2-dimensional. In general, the parameter and sufficient statistic can have different dimensions.

For example, if  $X_1, \dots, X_n$  are iid  $N(\mu, \mu^2)$ , where  $\mu \geq 0$ , then the minimal sufficient statistic is  $S(X) = (\bar{X}, S_{xx})$ .

**2.4 Rao-Blackwell Theorem**

So far we have written  $\mathbb{E}_\theta$  and  $\mathbb{P}_\theta$  to denote the expectations and probabilities in the model where  $X_1, \dots, X_n$  are iid drawn from  $f_X(\cdot \mid \theta)$ . From now on, we drop the subscript  $\theta$ .

**Theorem 2.3** (Rao-Blackwell Theorem). *Let  $T$  be a sufficient statistic for  $\theta$ . Let  $\tilde{\theta}$  be some estimator for  $\theta$ , with  $\mathbb{E}[\tilde{\theta}^2] < \infty$  for all  $\theta$ . Define a new estimator  $\hat{\theta} = \mathbb{E}[\tilde{\theta} \mid T(X)]$ . Then, for all  $\theta$ ,*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] \leq \mathbb{E}[(\tilde{\theta} - \theta)^2],$$

*with equality if and only if  $\tilde{\theta}$  is a function of  $T(X)$ .*

*Remark.*  $\hat{\theta}$  is a valid estimator, as it does not depend on  $\theta$ , only on  $X$ , as  $T$  is sufficient:

$$\hat{\theta}(T(x)) = \int \tilde{\theta}(x) f_{X|T}(x|T) dx,$$

where neither  $\tilde{\theta}$  nor the conditional distribution depend on  $\theta$ .

The message is that we can improve the mean squared error of any estimator  $\tilde{\theta}$  by taking a conditional expectation given  $T(X)$ .

**Proof:** By the tower property,

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\mathbb{E}[\tilde{\theta} \mid T]] = \mathbb{E}[\tilde{\theta}].$$

So  $\text{bias}(\hat{\theta}) = \text{bias}(\tilde{\theta})$  for all  $\theta$ . By the conditional variance formula,

$$\begin{aligned} \text{Var}(\tilde{\theta}) &= \mathbb{E}[\text{Var}(\tilde{\theta} \mid T)] + \text{Var}(\mathbb{E}[\tilde{\theta} \mid T]) \\ &= \mathbb{E}[\text{Var}(\tilde{\theta} \mid T)] + \text{Var}(\hat{\theta}). \end{aligned}$$

Hence  $\text{Var}(\tilde{\theta}) \geq \text{Var}(\hat{\theta})$  for all  $\theta$ . Hence  $\text{mse}(\tilde{\theta}) \geq \text{mse}(\hat{\theta})$ .

Note that  $\text{Var}(\tilde{\theta} \mid T) > 0$  with some positive probability unless  $\tilde{\theta}$  is a function of  $T(X)$ . So  $\text{mse}(\tilde{\theta}) > \text{mse}(\hat{\theta})$  unless  $\tilde{\theta}$  is a function of  $T(X)$ .

### Example 2.9.

Say  $X_1, \dots, X_n$  are iid  $\text{Poisson}(\lambda)$ . We wish to estimate  $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$ . Then

$$\begin{aligned} f_X(x \mid \lambda) &= \frac{e^{-n\lambda} \lambda^{x_1 + \dots + x_n}}{x_1! \dots x_n!} \\ &= \frac{\theta^n (-\log \theta)^{x_1 + \dots + x_n}}{x_1! \dots x_n!} \end{aligned}$$

Letting  $h(x) = 1/(x_1! \dots x_n!)$ ,  $g(T(x), \theta) = \theta^n (-\log \theta)^{T(x)}$ , by the factorization criterion,  $T(x) = \sum x_i$  is a sufficient statistic. Let  $\tilde{\theta} = \mathbb{1}(X_1 = 0)$ . This is unbiased, but only uses one observation  $X_1$ . Using Rao-Blackwell, we can find

$$\begin{aligned} \hat{\theta} &= \mathbb{E}[\tilde{\theta} \mid T = t] = \mathbb{P}\left(X_1 = 0 \mid \sum_{i=1}^n X_i = t\right) \\ &= \frac{\mathbb{P}(X_1 = 0, X_1 + \dots + X_n = t)}{\mathbb{P}(X_1 + \dots + X_n = t)} = \frac{\mathbb{P}(X_1 = 0, X_2 + \dots + X_n = t)}{\mathbb{P}(X_1 + \dots + X_n = t)} \\ &= \frac{\mathbb{P}(X_1 = 0) \mathbb{P}(X_2 + \dots + X_n = t)}{\mathbb{P}(X_1 + \dots + X_n = t)} = \frac{e^{-\lambda} \mathbb{P}(\text{Poisson}((n-1)\lambda) = t)}{\mathbb{P}(\text{Poisson}(n\lambda) = t)} \\ &= \frac{e^{-n\lambda} ((n-1)\lambda)^t / t!}{e^{-n\lambda} (n\lambda)^t / t!} = \left(1 - \frac{1}{n}\right)^t. \end{aligned}$$

So  $\hat{\theta} = (1 - \frac{1}{n})^{x_1 + \dots + x_n}$  is an estimator which by the Rao-Blackwell theorem has  $\text{mse}(\hat{\theta}) < \text{mse}(\tilde{\theta})$ .

As  $n \rightarrow \infty$ ,

$$\hat{\theta} = \left(1 - \frac{1}{n}\right)^{n\bar{x}} \xrightarrow{n \rightarrow \infty} e^{-\bar{x}},$$

and by the strong law of large numbers

$$\bar{x} \rightarrow \mathbb{E}[X_1] = \lambda.$$

so  $\hat{\theta} \rightarrow e^{-\lambda}$ .

### Example 2.10.

Let  $X_1, \dots, X_n$  be iid  $U([0, \theta])$  where  $\theta$  is unknown and  $\theta \geq 0$ . Then recall  $T(X) = \max_i X_i$  is sufficient for  $\theta$ .

Let  $\tilde{\theta} = 2X_1$ , which is unbiased. Then,

$$\begin{aligned} \hat{\theta} &= \mathbb{E}[\tilde{\theta} \mid T = t] = 2\mathbb{E}[X_1 \mid \max_i X_i = t] \\ &= 2\mathbb{E}[X_1 \mid \max_i X_i = t, \max_i X_i = X_1] \mathbb{P}(\max_i X_i = X_1 \mid \max_i X_i = t) \\ &\quad + 2\mathbb{E}[X_1 \mid \max_i X_i = t, \max_i X_i \neq X_1] \mathbb{P}(\max_i X_i \neq X_1 \mid \max_i X_i = t) \\ &= \frac{2t}{n} + \frac{2(n-1)}{n} \mathbb{E}[X_1 \mid X_1 \leq t, \max_{i>1} X_i = t] = \frac{2t}{n} + \frac{2(n-1)}{n} \frac{t}{2} = \frac{n+1}{n} t. \end{aligned}$$

So  $\hat{\theta} = \frac{n+1}{n} \max_i X_i$  is a valid estimator with  $\text{mse}(\hat{\theta}) < \text{mse}(\tilde{\theta})$ .

## 2.5 Maximum Likelihood Estimation

Let  $X = (X_1, \dots, X_n)$  have joint pdf (or pmf)  $f_X(X \mid \theta)$ .

**Definition 2.7.** The likelihood function is

$$L : \theta \mapsto f_X(X \mid \theta).$$

The *maximum likelihood estimator* is any value of  $\theta$  maximizing  $L(\theta)$ .

If  $X_1, \dots, X_n$  are iid each with pdf (or pmf)  $f_X(\cdot \mid \theta)$ , then

$$L(\theta) = \prod_{i=1}^n f_X(x_i \mid \theta).$$



We will denote the logarithm

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n \log f_X(x_i | \theta).$$

### Example 2.11.

If  $X_1, \dots, X_n$  are iid  $\text{Ber}(\theta)$ , then

$$\ell(\theta) = \left( \sum x_i \right) \log \theta = \left( n - \sum x_i \right) \log(1 - \theta),$$

and the derivative

$$\frac{\partial \ell}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{n - \sum x_i}{1 - \theta}.$$

This is zero if and only if  $\theta = \frac{1}{n} \sum x_i = \bar{X}$ .

Hence  $\bar{X}$  is the maximum likelihood estimator for  $\theta$ , and is unbiased as  $\mathbb{E}[\bar{X}] = \theta$ .

### Example 2.12.

If  $X_1, \dots, X_n$  are iid  $N(\mu, \sigma^2)$ , then

$$\log(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

This is maximized when  $\partial \ell / \partial \mu = \partial \ell / \partial \sigma^2 = 0$ . First, we get

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu),$$

which is equal to zero when  $\mu = \bar{X}$ . Then

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2.$$

This is zero when

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} S_{xx}.$$

Hence  $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, S_{xx}/n)$  give the maximum likelihood estimator in this model.

Note that  $\hat{\mu} = \bar{X}$  is unbiased. Now we want to see if  $\hat{\sigma}^2$  is biased. We could compute it directly, but later in the course we will show that

$$\frac{S_{xx}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-1}^2,$$

hence

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}[\chi_{n-1}^2] \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \neq \sigma^2,$$

which is biased, but asymptotically unbiased.

### Example 2.13.

Let  $X_1, \dots, X_n$  be iid  $U([0, \theta])$ . Then

$$L(\theta) = \frac{1}{\theta^n} \mathbb{1}(\max_i X_i \leq \theta).$$

We can see from the plot that  $\hat{\theta}_{\text{mle}} = \max_i X_i$  is the maximum likelihood estimator for  $\theta$ . We also started from an unbiased estimator, and using Rao-Blackwellization we found an estimator

$$\hat{\theta} = \frac{n+1}{n} \max_i X_i.$$

This is also unbiased. So in this model the mle is biased as

$$\mathbb{E}[\hat{\theta}_{\text{mle}}] = \mathbb{E}\left[\frac{n+1}{n} \hat{\theta}\right] = \frac{n}{n+1} \theta,$$

however it is asymptotically unbiased.

The maximum likelihood estimator has the following properties:

1. If  $T$  is a sufficient statistic, then the maximum likelihood estimator is a function of  $T(X)$ . By the factorization criterion,

$$L(\theta) = g(T(X), \theta)h(X).$$

If  $T(x) = T(y)$ , then the likelihood function with data  $x$  and  $y$  is the same up to a multiplicative constant. Hence the maximum likelihood estimator in each case is the same.

2. If  $\phi = h(\theta)$  where  $h$  is a bijection, then the maximum likelihood estimator of

$\phi$  is

$$\hat{\phi} = h(\hat{\theta}),$$

where  $\hat{\theta}$  is the maximum likelihood estimator of  $\theta$ .

3. Asymptotically, we have normality. This says  $\sqrt{n}(\hat{\theta} - \theta)$  is approximately normal with mean 0 when  $n$  is large. Under some regularity conditions, for a measurable set  $A$ ,

$$\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \in A) \xrightarrow{n \rightarrow \infty} \mathbb{P}(Z \in A),$$

where  $Z \sim N(0, \Sigma)$ . This holds for all regular values of  $\theta$ .

Here  $\Sigma$  is some function of  $\ell$ , and there is a theorem (Cramer-Rao) which says this is the smallest variance attainable.

4. Sometimes, if the maximum likelihood estimator is not available analytically, we can find it numerically.

## 2.6 Confidence Intervals

**Definition 2.8.** A  $(100 \cdot \gamma)\%$  *confidence interval* for a parameter  $\theta$  is a random interval  $(A(X), B(X))$  such that

$$\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma,$$

for all values of  $\theta$ .

The frequentist interpretation of the confidence interval is:

There exists some fixed true parameter  $\theta$ . We repeat the experiment many times.

On average,  $100 \cdot \gamma\%$  of the time the interval  $(A(X), B(X))$  contains  $\theta$ .

The incorrect interpretation is:

Having observed  $X = x$ , there is a probability  $\gamma$  that  $\theta$  is in  $(A(x), B(x))$ .

### Example 2.14.

Let  $X_1, \dots, X_n$  be iid  $N(\theta, 1)$ . To find a 95% confidence interval for  $\theta$ , we know that

$$\bar{X} = \frac{1}{n} \sum_{x_i} \sim N\left(\theta, \frac{1}{n}\right).$$

Hence

$$Z = \sqrt{n}(\bar{X} - \theta) \sim N(0, 1).$$

$Z$  has this distribution for all  $\theta$ . Then let  $z_1, z_2$  be any two numbers such

that  $\Phi(z_2) - \Phi(z_1) = 0.95$ . Then,

$$\mathbb{P}(z_1 \leq \sqrt{n}(\bar{X} - \theta) \leq z_2) = 0.95.$$

Rearranging,

$$\mathbb{P}\left(\bar{X} - \frac{z_2}{\sqrt{n}} \leq \theta \leq \bar{X} - \frac{z_1}{\sqrt{n}}\right) = 0.95.$$

Therefore  $(\bar{X} - \frac{z_2}{\sqrt{n}}, \bar{X} - \frac{z_1}{\sqrt{n}})$  is a 95% confidence interval.

There are multiple ways to choose  $z_1, z_2$ . Usually we minimise the width of the interval, which is achieved by  $z_1 = \Phi^{-1}(0.025)$ ,  $z_2 = \Phi^{-1}(0.975)$ .

To find a confidence interval, we can do the following:

1. Find some quantity  $R(X, \theta)$  such that the  $\mathbb{P}_\theta$  distribution of  $R(X, \theta)$  does not depend on  $\theta$ . This is called a *pivot*.

For example, we chose  $Z = \sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$ .

2. Write down a probabilistic statement about the pivot of the form

$$\mathbb{P}(c_1 \leq R(x, \theta) \leq c_2) = \gamma,$$

by using quantiles  $c_1, c_2$  of the distribution of  $R(X, \theta)$  (typically  $N(0, 1)$  or  $\chi_p^2$ ).

3. Rearrange the inequalities to leave  $\theta$  in the middle.

**Proposition 2.1.** *If  $T$  is a monotone increasing function  $T : \mathbb{R} \rightarrow \mathbb{R}$ , and  $(A(X), B(X))$  is a  $100 \cdot \gamma\%$  confidence interval for  $\theta$ , then  $(T(A(X)), T(B(X)))$  is a confidence interval for  $T(\theta)$ .*

*Remark.* When  $\theta$  is a vector, we talk about confidence sets.

### Example 2.15.

Let  $X_1, \dots, X_n$  be iid  $N(0, \sigma^2)$ . We want to find a 95% confidence interval for  $\sigma^2$ .

Note that  $\frac{X_i}{\sigma} \sim N(0, 1)$ , so using all the data points,

$$R(X, \sigma^2) = \sum_{i=1}^n \frac{X_i^2}{\sigma^2} \sim \chi_n^2$$

is a pivot. Let  $c_1 = F_{\chi_n^2}^{-1}(0.025)$ ,  $c_2 = F_{\chi_n^2}^{-1}(0.975)$ . Then,

$$\mathbb{P}(c_1 \leq R(X, \sigma^2) \leq c_2) = 0.95.$$

Rearranging,

$$\mathbb{P}\left(\frac{\sum x_i^2}{c_2} \leq \sigma^2 \leq \frac{\sum x_i^2}{c_1}\right) = 0.95.$$

Hence  $[\sum x_i^2/c_2, \sum x_i^2/c_1]$  is a 95% confidence interval  $\sigma^2$ .

Applying the proposition,  $[\sqrt{\sum x_i^2/c_2}, \sqrt{\sum x_i^2/c_1}]$  is a 95% confidence interval for  $\sigma$ .

### Example 2.16.

Let  $X_1, \dots, X_n$  be  $\text{Ber}(p)$ , for  $n$ . To find an approximate 95% for confidence interval  $p$ .

Recall that the maximum likelihood estimator for  $p$  is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i.$$

By the central limit theorem, when  $n$  is large,  $\hat{p}$  is approximately  $N(p, \frac{p(1-p)}{n})$ . Hence,

$$\sqrt{n} \frac{(\hat{p} - p)}{\sqrt{p(1-p)}} \sim N(0, 1),$$

approximately. If  $z = \Phi^{-1}(0.975)$ , then

$$\mathbb{P}\left(-z \leq \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} \leq z\right) \approx 0.95.$$

Rearranging this is tricky. Instead, we argue that  $n \rightarrow \infty$ ,  $\hat{p}(1-\hat{p}) \rightarrow p(1-p)$ . So replacing this in the denominator,

$$\mathbb{P}\left(-z \leq \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \leq z\right) \approx 0.95.$$

Rearranging this, we get

$$\mathbb{P}\left(\hat{p} - z \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + z \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) \approx 0.95.$$

Hence this is an approximate 95% confidence interval for  $p$ .

Note that  $z \approx 1.96$  and  $\sqrt{\hat{p}(1-\hat{p})} \leq \frac{1}{2}$  for all  $\hat{p} \in (0, 1)$ . So a conservative confidence interval is  $[\hat{p} \pm 1.96 \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{n}}]$ .

## 2.7 Interpreting Confidence Intervals

Suppose  $X_1, X_2$  are iid  $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ . We find a sensible 50% confidence interval for  $\theta$ . Consider

$$\begin{aligned} \mathbb{P}(\theta \text{ between } X_1, X_2) &= \mathbb{P}(\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)) \\ &= \mathbb{P}(X_1 \leq \theta \leq X_2) + \mathbb{P}(X_2 \leq \theta \leq X_1) \\ &= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}. \end{aligned}$$

Hence we can immediately conclude that  $(\min(X_1, X_2), \max(X_1, X_2))$  is a 50% confidence interval for  $\theta$ .

This does not mean for specific  $X_1 = x_1, X_2 = x_2$  the value  $\theta$  lies in  $(\min(x_1, x_2), \max(x_1, x_2))$  with probability  $\frac{1}{2}$ : consider when  $|x_1 - x_2| > \frac{1}{2}$ .

In this case, we can be sure that  $\theta$  is in  $(\min(x_1, x_2), \max(x_1, x_2))$ , as the value  $\theta$  can be at most distance  $\frac{1}{2}$  from  $x_1$  and  $x_2$ .

However the frequentist interpretation makes sense: if we repeat the experiment many times, we see  $\theta \in (\min(X_1, X_2), \max(X_1, X_2))$  exactly 50% of the time. We cannot say, given a specific observation that we are 50% certain that  $\theta$  is in the confidence interval.

### 3 Bayesian Inference

So far, we have assumed there is some true parameter  $\theta$ . That data  $X$  has pdf (or pmf)  $f_X(\cdot | \theta)$ .

Bayesian analysis is a different framework, where we treat  $\theta$  as a random variable, taking values in  $\Theta$ .

We begin by assigning to  $\theta$  a *prior distribution*  $\pi(\theta)$ , which represents the opinions or information about  $\theta$  before seeing on any data.

Conditional on  $\theta$ , the data  $X$  has pdf (or pmf)  $f_X(x | \theta)$ . Having observed a specific value of  $X = x$ , this information is combined with the prior to form the *posterior distribution*  $\pi(\theta | x)$ , which is the conditional distribution of  $\theta$  given  $X = x$ . By Bayes' rule,

$$\pi(\theta | x) = \frac{\pi(\theta) \cdot f_X(x | \theta)}{f_X(x)},$$

where  $f_X(x)$  is the marginal probability of  $X$ , and

$$f_X(x) = \begin{cases} \int_{\Theta} f_X(x | \theta) \pi(\theta) d\theta & \theta \text{ continuous,} \\ \sum_{\Theta} f_X(x | \theta) \pi(\theta) & \theta \text{ discrete.} \end{cases}$$

# Index

- bias, 7
- central limit theorem, 5
- change of variables, 6
- conditional expectation, 6
- conditional probability density function, 5
- conditional probability mass function, 5
- confidence interval, 18
- continuous random variable, 2
- covariance, 3
- discrete random variable, 2
- distribution function, 2
- estimator, 7
- event, 2
- expectation, 3
- factorization criterion, 10
- independence, 3
- joint probability density function, 5
- joint probability mass function, 5
- marginal probability density function, 5
- marginal probability mass function, 5
- maximum likelihood estimator, 15
- mean squared error, 7
- minimal sufficiency, 11
- pivot, 19
- posterior distribution, 22
- prior distribution, 22
- probability density function, 2
- probability mass function, 2
- probability measure, 2
- random variable, 2
- Rao-Blackwell theorem, 13
- sample space, 2
- sampling distribution, 7
- strong law of large numbers, 4
- sufficiency, 9
- tower property, 6
- unbiased estimator, 7
- variance, 3
- weak law of large numbers, 4