

# Statistics

Notes by Finley Cooper

10th February 2026

# Contents

<b>1</b>	<b>Parametric Estimation</b>	<b>3</b>
1.1	Review of IA Probability . . . . .	3
1.1.1	Starting axioms . . . . .	3
1.1.2	Joint random variables . . . . .	4
1.1.3	Limit theorems . . . . .	5
1.2	Estimators . . . . .	5
1.2.1	Bias-variance decomposition . . . . .	6
1.3	Sufficient statistics . . . . .	7
1.4	Minimal sufficiency . . . . .	8
1.5	Likelihood . . . . .	10
1.6	Confidence intervals . . . . .	11
1.7	Bayesian estimation . . . . .	13

# 1 Parametric Estimation

## 1.1 Review of IA Probability

### 1.1.1 Starting axioms

We observe some data  $X_1, \dots, X_n$  iid random variables taking values in a sample space  $\mathcal{X}$ . Let  $X = (X_1, \dots, X_n)$ . We assume that  $X_1$  belongs to a *statistical model*  $\{p(x; \theta) : \theta \in \Theta\}$  with  $\theta$  unknown. For example  $p(x; \theta)$  could be a pdf.

Let's see some examples

- (i) Suppose that  $X_1 \sim \text{Poisson}(\lambda)$  where  $\theta = \lambda \in \Theta = (0, \infty)$ .
- (ii) Suppose that  $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ , where  $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ .

We have some common questions about these statistical models.

- (i) We want to give an estimate  $\hat{\theta} : \mathcal{X}^n \rightarrow \Theta$  of the true value of  $\theta$ .
- (ii) We also want to give an interval estimator  $(\hat{\theta}_1(X), \hat{\theta}_2(X))$  of  $\theta$ .
- (iii) Further we want to test of hypothesis about  $\theta$ . For example we might make the hypothesis that  $H_0 : \theta = 0$ .

Let's do a quick reievw of IA Probability. Let  $X : \Omega \rightarrow \mathbb{R}$  be a random variable defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . So  $\Omega$  is the sample space,  $\mathcal{F}$  is the set of events, and  $\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$  is the probability measure.

The cumulative distribution function (cdf) of  $X$  is  $F_X(s) = \mathbb{P}(X \leq x)$ . A discrete random variable takes values in a countable set  $\mathcal{X}$  and has probability mass function (pmf) given by  $p_X(x) = \mathbb{P}(X = x)$ . A continuous random variable has probability density function (pdf)  $f_X$  satisfying  $P(X \in A) = \int_A f_X(x) dx$  (for measurable sets  $A$ ). We say that  $X_1, \dots, X_n$  are independent if  $\mathbb{P}(X_1 \leq x_1, \dots, X_n \leq x_n) = \prod_{i=1}^n \mathbb{P}(X_i \leq x_i)$  for all choices  $x_1, \dots, x_n$ . If  $X_1, \dots, X_n$  have pdfs (or pmfs)  $f_{X_1}, \dots, f_{X_n}$ , then this is equivalent to  $f_X(x) = \prod_{i=1}^n f_{X_i}(x_i)$  for all  $x_i$ . The expectation of  $X$  is,

$$\mathbb{E}(x) = \begin{cases} \sum_{x \in \mathcal{X}} x p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_X(x) & \text{if } X \text{ is continuous} \end{cases}.$$

The variance of  $X$  is  $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$ . The moment generating function of  $X$  is  $M(t) = \mathbb{E}[e^{tX}]$  and can be used to generate the momentum of a random variable by taking derivatives. If two random variables have the same moment generating functions, then they have the same distribution.

The expectation operator is linear and

$$\text{Var}(a_1 X_1 + \dots + a_n X_n) = \sum_{i,j=1}^n a_i a_j \text{Cov}(X_i, X_j),$$

where  $\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]$ . In vector notation writing  $X$  as the column vector of  $X_i$  and  $a$  as the column vector for  $a_i$  we get that

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

Similar for the variance we get that

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

where  $\text{Var}(X)$  is the covariance matrix for  $X$  with entries  $\text{Cov}(X_i, X_j)$ .

### 1.1.2 Joint random variables

If  $X$  is a discrete random variable with pmf  $P_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$  and marginal pmf  $P_Y(y) = \sum_{x \in X} P_{X,Y}(x, y)$ , then the conditional pmf is

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

If  $X, Y$  are continuous then the join pdf  $f_{X,Y}$  satisfies

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

and 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 of  $X$  given  $Y$  is

$$E(X | Y) = \begin{cases} \sum_{x \in X} x \mathbb{P}_{X|Y}(x | Y) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f_{X|Y}(x | Y) dy & \text{if } Y \text{ is continuous} \end{cases}.$$

*Remark.*  $\mathbb{E}(X | Y)$  is a function of  $Y$  so  $\mathbb{E}(X | Y)$  is a random variable.

We also have the law of total expectation,

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

This is a consequence of the law of total probability which is

$$p_X(x) = \sum_y p_{X|Y}(x | y) p_Y(y).$$

Now we have a new (but less useful) theorem similar to the tower property of expectation.

**Theorem.** (Law of total variance)

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | Y)] + \text{Var}(\mathbb{E}[X | Y]).$$

*Proof.* Write  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ , so

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

We also have the change of variables formula. If we have a mapping  $(x, y) \rightarrow (u, v)$ , a bijection from  $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ , then

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v)) |\det J|,$$

where  $J$  is the Jacobian matrix.

### 1.1.3 Limit theorems

Suppose  $X_1, \dots, X_n$  are iid random variables with mean  $\mu$  and variance  $\sigma^2$ . Define the sum  $S = \sum_{i=1}^n X_i$  and the sample mean  $\bar{X}_n = \frac{S_n}{n}$ . We have the following theorems.

**Theorem.** (Weak Law of Large Numbers)

$$\bar{X}_n \rightarrow \mu$$

where  $\rightarrow$  means that  $\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \rightarrow 0$  as  $n \rightarrow \infty$  for all  $\varepsilon > 0$ .

**Theorem.** (Strong Law of Large Numbers)

$$\bar{X}_n \rightarrow \mu$$

almost surely. So  $\mathbb{P}(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$ .

**Theorem.** (Central Limit Theorem) The random variables

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

is approximately  $\mathcal{N}(0, 1)$  for large  $n$ . Or we can write this as

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2).$$

Formally this means that  $\mathbb{P}(Z_n \leq z) \rightarrow \Phi(z)$  for all  $z \in \mathbb{R}$  where  $\Phi(z)$  is the cdf of  $\mathcal{N}(0, 1)$ .

## 1.2 Estimators

Suppose that  $X_1, \dots, X_n$  are iid with pdf  $f_X(x | \theta)$  and parameter  $\theta$  unknown.

**Definition.** (Estimator) A function of the data  $T(X) \rightarrow \hat{\theta}$  which is used to approximate the true parameter  $\theta$  is called an *estimator* (or sometimes a *statistic*). The distribution of  $T(X)$  is the *sampling distribution*.

For an example suppose that  $X_1, \dots, X_n \sim \mathcal{N}(\mu, 1)$  and let  $\hat{\mu} = T(x) = \frac{1}{n} \sum_{i=1}^n X_i$ . The sampling distribution of  $\hat{\mu}$  is  $T(X) \sim \mathcal{N}(\mu, \frac{1}{n})$ .

**Definition.** (Bias) The *bias* of a random variable  $\hat{\theta} = T(X)$  is

$$\text{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta,$$

where the expectation is taken over the model  $X_1 \sim f_X(\cdot | \theta)$ .

*Remark.* In general the bias might be a function of  $\theta$  which is not explicit in the notation.

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

So for our estimator from before,  $\hat{\mu}$ , is unbiased since

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

### 1.2.1 Bias-variance decomposition

**Definition.** (Mean squared error) The *mean squared error* of an estimator  $\hat{\theta}$  is

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

*Remark.* Note that the MSE is generally a function of  $\theta$  like the bias. Again this is not clear from the notation.

**Proposition.** (Bias-variance decomposition) For an estimator  $\hat{\theta}$  of a parameter  $\theta$ , we have that

$$\text{mse}(\hat{\theta}) = \left(\text{bias}(\hat{\theta})\right)^2 + \text{Var}_{\theta}(\hat{\theta}).$$

*Proof.*

$$\begin{aligned} \text{mse}(\hat{\theta}) &= \mathbb{E}_{\theta}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}_{\theta} \left[ \left( \hat{\theta} - \mathbb{E}_{\theta}(\hat{\theta}) + \mathbb{E}_{\theta}(\hat{\theta}) - \theta \right)^2 \right] \\ &= \mathbb{E}_{\theta}[(\hat{\theta} - \mathbb{E}_{\theta}(\hat{\theta}))^2] + (\mathbb{E}_{\theta}(\hat{\theta}) - \theta)^2 + 2(\mathbb{E}_{\theta}(\hat{\theta}) - \theta) \cdot \mathbb{E}_{\theta}[\hat{\theta} - \mathbb{E}_{\theta}(\hat{\theta})] \\ &= \left(\text{bias}(\hat{\theta})\right)^2 + \text{Var}_{\theta}(\hat{\theta}). \quad \square \end{aligned}$$

Let's see an example. Suppose that  $X \sim \text{Binomial}(n, \theta)$  where  $n$  is known and we want to estimate  $\theta \in [0, 1]$ . Let  $T_u = \frac{X}{n}$  be an estimator, so  $\mathbb{E}_{\theta}(T_u) = \frac{\mathbb{E}(X)}{n} = \frac{n\theta}{n} = \theta$ , hence this estimator is unbiased. And  $\text{mse}(T_u) = \text{Var}(T_u) + \text{bias}(T_u) = \frac{\theta(1-\theta)}{n}$ .

Instead if we used the estimator  $T_b = \frac{X+1}{n+2} = \omega \frac{X}{n} + (1-\omega) \frac{1}{2}$  where  $\omega = \frac{n}{n+2}$ . We get that

$$\begin{aligned} \text{bias}(T_b) &= (1-\omega)\left(\frac{1}{2} - \theta\right) \\ \text{Var}(T_b) &= \omega^2 \frac{\theta(1-\theta)}{n}. \end{aligned}$$

Giving that

$$\text{mse}(T_b) = \omega^2 \theta(1-\theta) + (1-\omega)^2 \left(\frac{1}{2} - \theta\right)^2$$

### 1.3 Sufficient statistics

Suppose  $X_1, \dots, X_n$  are iid random variables taking values in  $\mathcal{X}$  with pdf  $f_{X_1}(\cdot | \theta)$ . Consider  $\theta$  as fixed. Denote  $X = (X_1, \dots, X_n)$ .

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

*Remark.* The parameter  $\theta$  may be a vector, and  $T(X)$  may be a vector.

Suppose  $X_1, \dots, X_n \sim \text{Binomial}(1, \theta)$  iid for some  $\theta \in [0, 1]$ . Then

$$\begin{aligned} f_X(x | \theta) &= \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} \\ &= \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \end{aligned}$$

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

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

**Theorem.** (Factorisation criterion) The statistics  $T$  is sufficient for  $\theta$  if and only if  $f_X(x | \theta) = g(T(x), \theta)h(x)$  for some suitable  $g$  and  $h$ .

*Proof.* Suppose that  $f_X(x | \theta) = g(T(x), \theta)h(x)$ . We can compute

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

which doesn't depend on  $\theta$ , so  $T(X)$  is sufficient.

Conversely, suppose  $T(X)$  is sufficient. We can write

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

So we're done. □

*Remark.* For our example before we can define  $T(x) = \sum x_i$  and  $g(t, \theta) = \theta^t (1 - \theta)^{n-t}$  and  $h(x) = 1$ .

Let's see another example. Let  $X_1, \dots, X_n$  be iid uniform on  $[0, \theta]$  for some  $\theta \in (0, \infty)$ . So

$$\begin{aligned} f_X(x | \theta) &= \prod_{i=1}^n \frac{1}{\theta} \mathbf{1}\{x_i \in [0, \theta]\} \\ &= \frac{1}{\theta^n} \mathbf{1}\{\max x_i \leq \theta\} \mathbf{1}\{\min x_i \geq 0\} \\ &= g(T(x), \theta) h(x). \end{aligned}$$

## 1.4 Minimal sufficiency

**Definition.** (Minimal sufficient) A sufficient statistics  $T(X)$  is *minimal sufficient* if it is a function of every other sufficient statistic. So if  $T'(X)$  is also sufficient, then  $T'(x) = T'(y) \implies T(x) = T(y)$  for all  $x, y \in \chi$ .

*Remark.* Minimal sufficient statistics are unique up to bijection.

**Theorem.** Suppose  $T(X)$  is a statistics such that  $\frac{f_X(x|\theta)}{f_X(y|\theta)}$  is constant a function of  $\theta$  if and only if  $T(x) = T(y)$ . Then  $T$  is minimal sufficient.

Let's see an example before we prove this. Suppose that  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ . Then

$$\begin{aligned} \frac{f_X(x | \mu, \sigma^2)}{f_X(y | \mu, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right)}{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum (y_i - \mu)^2\right)} \\ &= \exp\left(-\frac{1}{2\sigma^2} \left(\sum_i x_i^2 - \sum_i y_i^2\right) + \frac{\mu}{\sigma^2} \left(\sum_i x_i - \sum_i y_i\right)\right) \end{aligned}$$

This is constant in  $(\mu, \sigma^2)$  if and only if  $\sum_i x_i = \sum_i y_i$  and  $\sum_i x_i^2 = \sum_i y_i^2$  therefore  $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$  is minimal sufficient.

*Proof.* Need to show that such a statistics is sufficient and minimal. First we'll show sufficiency. For each  $t$  pick a  $x_t$  such that  $T(x_t) = t$ . Now let  $x \in \chi_N$  and let  $T(x) = t$ . So  $T(x) = T(x_t)$ , so by the hypothesis  $\frac{f_X(x, \theta)}{f_X(x_t, \theta)}$  does not depend on  $\theta$ . Let this be  $h(x)$  and let  $g(t, \theta) = f_X(x_t, \theta)$  then we have that  $f_X(x, \theta) = g(t, \theta)h(x)$  so sufficient.

Now let  $S$  be any other sufficient statistic. By the factorisation criterion, there exists  $g_S, h_S$  such that  $f_X(x | \theta) = g_S(S(x), \theta)h_S(x)$ . Suppose  $S(x) = S(y)$ . Then

$$\frac{f_X(x | \theta)}{f_X(y | \theta)} = \frac{g_S(S(x), \theta)h_S(x)}{g_S(S(y), \theta)h_S(y)} = \frac{h_S(x)}{h_S(y)}$$

which does not depend on  $\theta$  so  $T(x) = T(y)$  so  $T$  is minimal sufficient.  $\square$

We know that bijections of minimal sufficient statistics are still minimal sufficient statistics, so we can write our minimal sufficient statistic for  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  as

$$S(X) = (\bar{X}, S_{XX})$$

where  $\bar{X} = \frac{1}{n} \sum_i X_i$  and  $S_{XX} = \sum_i (X_i - \bar{X})^2$ , since there is a bijection between them.

Until now we used  $\mathbb{E}_\theta$  and  $\mathbb{P}_\theta$  to denote expectation and probability when  $X_1, \dots, X_n$  are iid from a distribution with pdf  $f_X(x | \theta)$ . From now on we drop the subscript  $\theta$  to simplify notation.



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

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

This inequality is strict unless  $\tilde{\theta}$  is a function of  $T$ .

*Remark.* We have that  $\hat{\theta}(T) = \int \tilde{\theta}(x) f_{X|T}(x \mid T) dx$ . By sufficiency of  $T$ , the conditional pdf does not depend on  $\theta$  so  $\hat{\theta}$  does not depend on  $\theta$ , and is valid estimator.

*Proof.* By the tower property of expectation,

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

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

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

So

$$\text{mse}(\tilde{\theta}) \geq \text{mse}(\hat{\theta}).$$

Equality is achieved only when  $\text{Var}(\tilde{\theta} \mid T) = 0$  with probability 1 which requires  $\tilde{\theta}$  to be a function of  $T$ .  $\square$

Let's see an example of this. Suppose that  $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$  iid. Let  $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$ . Then

$$f_X(x \mid \theta) = \frac{e^{-n\lambda} \lambda^{\sum x_i}}{\prod_i x_i!} = \frac{\theta^n (-\log \theta)^{\sum x_i}}{\prod_i x_i!}.$$

By the factorisation criterion,  $T(X) = \sum_i x_i$  is sufficient. Recall that  $\sum x_i \sim \text{Poisson}(n\lambda)$ . Let  $\tilde{\theta} = \mathbf{1}\{X_1 = 0\}$ . Then

$$\begin{aligned} \hat{\theta} &= \mathbb{E}[\tilde{\theta} \mid T = t] = \mathbb{P}\left(X_1 = 0 \mid \sum_{i=1}^n X_i = t\right) \\ &= \frac{\mathbb{P}(X_1 = 0, \sum_{i=2}^n X_i = t)}{\mathbb{P}(\sum_{i=1}^n X_i = t)} \\ &= \frac{\mathbb{P}(X_1 = 0) \mathbb{P}(\sum_{i=2}^n X_i = t)}{\mathbb{P}(\sum_{i=1}^n X_i = t)} \\ &= \frac{e^{-\lambda} e^{-(n-1)\lambda} \frac{((n-1)\lambda)^t}{t!}}{e^{-n\lambda} \frac{(n\lambda)^t}{t!}} = \left(\frac{n-1}{n}\right)^t \end{aligned}$$

Hence  $\hat{\theta} = (1 - \frac{1}{n})^{\sum x_i}$  has  $\text{mse}(\hat{\theta}) < \text{mse}(\tilde{\theta})$  for all  $\theta$ . We can see that as  $n \rightarrow \infty$ ,  $\hat{\theta} \rightarrow e^{-\bar{X}} = e^{-\lambda} = \theta$ .

Let  $X_1, \dots, X_n \sim \text{Uniform}([0, \theta])$  and suppose we want to estimate  $\theta \geq 0$ . Last time we saw that  $T = \max X_i$  is sufficient for  $\theta$ . Let  $\tilde{\theta} = 2X_1$  be an estimator (unbias). Then

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

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

## 1.5 Likelihood

**Definition.** (Likelihood) Let  $X = (X_1, \dots, X_n)$  have a joint pdf  $f_X(x \mid \theta)$ . The *likelihood* of  $\theta$  is the function

$$L : \theta \rightarrow f_X(x \mid \theta).$$

The max likelihood estimator (MLE) is the value of  $\theta$  maximizing  $L$ .

If  $X_1, \dots, X_n \sim f_X(\cdot \mid \theta)$  iid, then  $L(\theta) = \prod_{i=1}^n f_X(x_i \mid \theta)$ .

It's usually easier to work with the log-likelihood, since this reduces to a sum. So in the iid case,

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

For example let  $X_1, \dots, X_n \sim \text{Bernoulli}(p)$  iid. Then we get that

$$\ell(p) = \left(\sum_{i=1}^n X_i\right) \log p + \left(n - \sum_{i=1}^n X_i\right) \log(1-p).$$

Taking the derivative with respect to  $p$ ,

$$\frac{\partial \ell}{\partial p} = \frac{\sum_i X_i}{p} - \frac{n - \sum_i X_i}{1-p}.$$

So setting the derivative to zero we get that

$$p = \frac{\sum X_i}{n}.$$

Hence the MLE is

$$\hat{p} = \frac{\sum_i X_i}{n},$$

and since  $\mathbb{E}[\hat{p}] = p$ , this is unbiased.

Now suppose  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ .

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

So

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

which is zero when  $\mu = \frac{\sum_i X_i}{n}$  regardless of  $\sigma$ . Also

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

If we set  $\mu = \frac{\sum_i X_i}{n}$  then we get  $\frac{\partial \ell}{\partial \sigma^2} = 0$  if  $\sigma^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{S_{xx}}{n}$ . Hence the MLE is

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \frac{S_{xx}}{n}).$$

Note that  $\mu$  is unbiased, but we will see later that

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

So  $\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left(\frac{S_{xx}}{n}\right) = \frac{n-1}{n}\sigma^2$ . So  $\hat{\sigma}^2$  is not unbiased, but is asymptotically unbiased as  $n \rightarrow \infty$ .

Suppose now that  $X_1, \dots, X_n \sim \text{Uniform}([0, \theta])$  iid. Then

$$\ell(\theta) = \frac{1}{\theta^n} \mathbf{1}_{\{\max_i X_i \leq \theta\}}.$$

Hence the MLE is  $\hat{\theta}_{\text{MLE}} = \max_i X_i$ . Recall that last time, we had an unbiased estimator  $\tilde{\theta}$  and by the Rao-Blackwell Theorem we found the estimator  $\hat{\theta} = \mathbb{E}[\tilde{\theta} | T] = \frac{n+1}{n} \max_i X_i$ . Note that  $\hat{\theta}_{\text{MLE}} = \frac{n}{n+1} \hat{\theta}$ , so  $\mathbb{E}[\hat{\theta}_{\text{MLE}}] = \frac{n}{n+1} \mathbb{E}[\hat{\theta}] = \frac{n}{n+1} \theta$ . Again this not unbiased, but is asymptotically unbiased.

Let's see some properties of the MLE.

- (i) If  $T$  is a sufficient statistic, the MLE is a function of  $T$ . We can factorise  $L(\theta) = g(T(x), \theta)h(x)$ .
- (ii) If  $\phi = h(\theta)$  where  $h$  is a bijection, the MLE of  $\phi$  is  $\hat{\phi} = h(\hat{\theta})$  where  $\hat{\theta}$  is the MLE of  $\theta$ .
- (iii) Asymptotic normality:  $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta)$  is approximately normal with mean 0 for large  $n$ .  
The covariance matrix is the "smallest attainable" (see II Principles of Statistics).

## 1.6 Confidence intervals

**Definition.** (Confidence intervals) A  $(100\gamma)\%$  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 some  $\gamma \in (0, 1)$  and all values of the true parameter  $\theta$ .

*Remark.* The incorrect interpretation: Having observed  $X = x$ , there is a  $1 - \gamma$  probability that  $\theta$  is in  $(A(X), B(X))$ . This is wrong.

Suppose that  $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$  iid. We want to find a 95% confidence interval for  $\theta$ . We know that  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim \mathcal{N}(\theta, \frac{1}{n})$ . If we define  $Z = \sqrt{n}(\bar{X} - \theta)$  Then  $Z \sim \mathcal{N}(0, 1)$  no matter

the value of  $\theta$ . Let  $z_1, z_2$  be numbers with  $\Phi(z_1) - \Phi(z_2) = 0.95$  where  $\Phi$  is the cdf of the standard normal.  $\mathbb{P}(z_1 \leq \sqrt{n}(\bar{X} - \theta) \leq z_2) = 0.95$  rearranging we get that

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

hence

$$\left(\bar{X} - \frac{z_2}{\sqrt{n}}, \bar{X} - \frac{z_1}{\sqrt{n}}\right)$$

is a 95% confidence interval.

This is the recipe for confidence intervals.

- (i) Find a quantity  $R(X, \theta)$  such that this  $\mathbb{P}_\theta$  distribution of  $R(X, \theta)$  does not depend on  $\theta$ . This is called a *pivot* for example  $R(X, \theta) = \sqrt{n}(\bar{X} - \theta)$ .
- (ii) Write down the statement

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

where  $(c_1, c_2)$  are quantiles of the distribution of  $R(X, \theta)$ .

- (iii) Rearranging the above to leave  $\theta$  in the middle of the inequality, so we get something in the form

$$\mathbb{P}(A(X) \leq \theta \leq B(X)).$$

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

Suppose that  $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$  iid. We want a 95% confidence interval for  $\sigma^2$ . Note that

$$\frac{X_i}{\sigma} \sim \mathcal{N}(0, 1),$$

so  $\sum_{i=1}^n \frac{X_i^2}{\sigma^2} \sim \chi_n^2$ . Hence

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

is a *pivot*. Let  $F_{\chi_n^2}^{-1}(0.025)$  and  $F_{\chi_n^2}^{-1}(0.975)$ . Then

$$\mathbb{P}\left(c_1 \leq \sum_{i=1}^n \frac{X_i^2}{\sigma^2} \leq c_2\right) = 0.95,$$

so rearranging we get that

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

gives our confidence interval.

Now suppose that  $X_1, \dots, X_n \sim \text{Bernoulli}(p)$  for large  $n$ . We will find an approximate 95% confidence interval for  $p$ . The maximum likelihood estimator of  $p$  is  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$ . By the central limit theorem  $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$  approximately. Thus

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

for large  $n$ . So we have our pivot, which gives

$$\mathbb{P} \left( -z_{0.025} \leq \frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} \leq z_{0.025} \right) \approx 0.95$$

Instead of inverting directly, if  $n$  is large  $\hat{p} \approx p$ , so switching  $p$  with  $\hat{p}$  on the denominator we get that

$$\mathbb{P} \left( \hat{p} - z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \approx 0.95$$

which gives our confidence interval. Since for all  $\hat{p} \in [0, 1]$  we have  $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$  we would also report a conservative confidence interval of  $\hat{p} \pm z_{0.025} \sqrt{\frac{1}{4n}}$ .

## 1.7 Bayesian estimation

So far we've been using frequentist methods treating  $\theta \in \Theta$  as fixed. For Bayesian methods, we treat  $\theta$  as random with a prior distribution  $\pi(\theta)$ . Conditional on  $\theta$  the data  $X$  has pdf  $f_X(\cdot | \theta)$ . Having observed that  $X = x$ , we combine with the prior to form a posterior distribution  $\pi(\theta | X)$ . By Bayes' rule,

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

where  $f_X(x)$  is the marginal distribution of  $X$ , so

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

More simply,

$$\pi(\theta | x) \propto \pi(\theta) f_X(x | \theta).$$

Often it is easier to recognise the RHS as proportional to a known distribution.

*Remark.* By the factorisation criterion, the posterior only depends on  $X$  through a sufficient statistic.

$$\begin{aligned} \pi(\theta | x) &\propto \pi(\theta) \cdot f_X(x | \theta) = \pi(\theta) \cdot g(T(x), \theta) h(x) \\ &\propto \pi(\theta) g(T(x), \theta). \end{aligned}$$

Suppose  $\theta \in [0, 1]$  is the mortality rate for some procedure at Addenbrookes. In the first 10 operations there are no deaths. In other hospitals across the country the mortality rate is between 3 – 20%, with average of 10%. Consider the prior distribution  $\pi(\theta) \sim \text{Beta}(a, b)$  we can choose  $(a, b) = (3, 27)$  so  $\pi(\theta)$  has mean 0.1 and  $\pi(0.03 < \theta < 0.2) = 0.9$ .

Let  $X_i \sim \text{Bernoulli}(\theta)$  be indicator for whether  $i$ th patient at Addenbrookes dies.

$$f_X(x | \theta) = \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i}.$$

The posterior is

$$\begin{aligned} \pi(\theta | X) &\propto \pi(\theta) f_X(x | \theta) \\ &\propto \theta^{a-1} (1 - \theta)^{b-1} \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} \\ &= \theta^{\sum x_i + a - 1} (1 - \theta)^{b + n - \sum x_i - 1}. \end{aligned}$$

Hence

$$\pi(\theta \mid X) \sim \text{Beta}(a + \sum x_i, b + n - \sum x_i)$$

so pluggin in  $a = 3, b = 27, n = 10, \sum X_i = 0$ , so the posterior is  $\text{Beta}(3, 37)$ .

*Remark.* In the example the prior and posterior were from the same family of distributions known as conjugacy.

Suppose we put a  $\text{Beta}(a, b)$  prior on the parameter  $\theta$  of kidney cancer death rates in each county. We can estimate  $(a, b) = (27, 58000)$  with  $\frac{a}{a+b} \approx 4.65 \times 10^{-9}$  being the kidney cancer death rate in the United States. The previous example shows that if we observe  $\sum_{i=1}^n X_i$  deaths in a county, the posterior mean estimate is  $\frac{a + \sum X_i}{a + b + n}$ . This is equal to

$$\frac{n}{a + b + n} \cdot \frac{\sum X_i}{n} + \frac{a + b}{a + b + n} \cdot \frac{a}{a + b}.$$

For large  $n$ , we use  $\approx \frac{\sum X_i}{n}$  as our estimate, for small  $n$  we use  $\frac{a}{a+b}$  and in between we shrink our estimate between them.