

# Statistics

Notes by Finley Cooper

29th January 2026

# Contents

<b>1 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 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 review 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  $F_X(s) = \mathbb{P}(X \leq s)$ . 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

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

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 joint 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

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

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

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

*Proof.*

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

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

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

### 1.3 Sufficient statistics

Suppose  $X_1, \dots, X_n$  are iid random variables taking values in  $\chi$  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.  $\square$

*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

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

## 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 statistics.

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  $g(t, \theta) = f_X(x, \theta)$  then we have that  $f_X(x, \theta) = g(t, \theta)h(x)$  so sufficient.