Statistics

Notes by Finley Cooper

27th January 2026

Contents

1 Parametric Estimation

1.1 Review of IA Probability

1.1.1 Starting axioms

We observe some data $X_1, \ldots, X_n$ iid random variables taking values in a sample space $\mathcal{X}$. Let $X = (X_1, \ldots, X_n)$. We assume that X_1 belongs to a *statistical model* $\{p(x;\theta) : \theta \in \Theta\}$ with θ 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 $\hat{\theta} : \mathcal{X}^n \to \Theta$ of the true value of θ.

(ii) We also want to give an interval estimator $(\hat{\theta}_1(X), \hat{\theta}_2(X))$ of θ.

(iii) Further we want to test of hypothesis about θ. For example we might make the hypothesis that $H_0 : \theta = 0$.

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

The cumulative distribution function (cdf) of X is $F_X(s) = \mathbb{P}(X \le x)$. 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 $P(X \in A) = \int_A f_X(x)\mathrm{d}x$ (for measurable sets A). We say that $X_1, \ldots, X_n$ are independent if $\mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \prod_{i=1}^n \mathbb{P}(X_i \le x_i)$ for all choices $x_1, \ldots, x_n$. If $X_1, \ldots, X_n$ have pdfs (or pmfs) $f_{X_1}, \ldots, f_{X_n}$, 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 varaible 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

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

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

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

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

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

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

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

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

and the marginal pdf of Y is

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

The *conditional pdf* of X given Y is $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$.

The conditional expectation of X given Y is

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

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

We also have the law of total expectation,

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

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

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

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

Theorem. (Law of total variance)

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

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

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

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

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

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

Theorem. (Strong Law of Large Numbers)

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

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

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 $\mathbb{P}(Z_n \le z) \to \Phi(z)$ 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 $f_X(x \mid \theta)$ and parameter θ unknown.

Definition. (Estimator) A function of the data $T(X) \to \hat{\theta}$ which is used to approximate the true parameter θ 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 $X_1 \sim f_X(\cdot \mid \theta)$.

Remark. In general the bias might be a function of θ which is not explicit in the notation.

Definition. (Unbiased estimator) We say that an estimator is *unbiased* if $\operatorname{bias}(\hat{\theta}) = 0$ 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

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

Remark. Note that the MSE is generally a function of θ like the bias. Again this is not clear from the notation.

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

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

Proof.

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