Part IB — Statistics

Based on lectures by Prof. S. Bacallado and notes by thirdsgames.co.uk

Lent 2022

Contents

§1 Introduction and review of IA Probability

§1.1 Introduction

Statistics can be defined as the science of making informed decisions. The field comprises, for example:

- the design of experiments and studies;
- visualisation of data;
- formal statistical inference (which is the focus of this course);
- communication of uncertainty and risk; and
- formal decision theory.

This course concerns itself with *parametric inference*. Let $X_1, \dots, X_n$ be i.i.d. (independent and identically distributed) random variables, where we assume that the distribution of X_1 belongs to some family with parameter $\theta \in \Theta$. For instance, let $X_1 \sim \text{Poisson}(\mu)$, where $\theta = \mu$ and $\Theta = (0, \infty)$. Another example is $X_1 \sim N(\mu, \sigma^2)$, and $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (0, \infty)$. We use the observed $X = (X_1, \dots, X_n)$ to make inferences about the parameter θ:

1. we can estimate the value of θ using a *point estimate* written $\hat{\theta}(X)$;
2. we can make an *interval estimate* of θ, written $(\hat{\theta}_1(X), \hat{\theta}_2(X))$;
3. hypotheses about θ can be tested, for instance the hypothesis $H_0 \colon \theta = 1$, by checking whether there is evidence in the data X against the hypothesis H_0.

Remark 1. In general, we will assume that the family of distributions of the observations X_i is known *a priori*, and the parameter θ is the only unknown. There will, however, be some remarks later in the course where we can make weaker assumptions about the family.

§1.2 Review of IA Probability

This subsection reviews material covered in the IA Probability course. Some keywords are measure-theoretic, and are not defined.

Let Ω be the *sample space* of outcomes in an experiment. A *measurable* subset of Ω is called an *event*, and we denote the set of events by $\mathcal{F}$. A *probability measure* $\mathbb{P} \colon \mathcal{F} \to [0, 1]$ satisfies the following properties.

1. $\mathbb{P}(\varnothing) = 0$;
2. $\mathbb{P}(\Omega) = 1$;

3. $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ if (A_i) is a sequence of disjoint events.

A *random variable* is a *measurable function* $X \colon \Omega \to \mathbb{R}$. The *distribution function* of a random variable X is the function $F_X(x) = \mathbb{P}(X \leq x)$. We say that a random variable is *discrete* when it takes values in a countable set $\mathcal{X} \subset \mathbb{R}$. The *probability mass function* of a discrete random variable is the function $p_X(x) = \mathbb{P}(X = x)$. We say that X has a *continuous distribution* if it has a *probability density function* $f_X(x)$ such that $\mathbb{P}(x \in A) = \int_A f_X(x)\,\mathrm{d}x$ for 'nice' sets A.

The *expectation* of a random variable X is defined as

$$\mathbb{E}[X] = \begin{cases} \sum_{x \in X} x p_X(x) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x & \text{if } X \text{ continuous} \end{cases}$$

If $g \colon \mathbb{R} \to \mathbb{R}$, we define $\mathbb{E}[g(X)]$ by considering the fact that $g(X)$ is also a random variable. For instance, in the continuous case,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$$

The *variance* of a random variable X is defined as $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$.

We say that a set of random variables $X_1, \ldots, X_n$ are *independent* if, for all $x_1, \ldots, x_n$, we have

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

If and only if $X_1, \ldots, X_n$ have probability density (or mass) functions $f_1, \ldots, f_n$, then the *joint probability density (respectively mass) function* is

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

If $Y = \max\{X_1, \ldots, X_n\}$ where the X_i are independent, then the distribution function of Y is given by

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X_1 \leq y) \cdots \mathbb{P}(X_n \leq y)$$

The probability density function of Y (if it exists) is obtained by the differentiating the above.

Under a linear transformation, the expectation and variance have certain properties. Let $a = (a_1, \ldots, a_n)^\intercal \in \mathbb{R}^n$ be a constant in $\mathbb{R}^n$.

$$\mathbb{E}[a_1 X_1 + \cdots + a_n X_n] = \mathbb{E}[a^\intercal X] = a^\intercal \mathbb{E}[X]$$

where $\mathbb{E}[X]$ is defined componentwise. Note that independence of X_i is not required for linearity of the expectation to hold. Similarly,

$$\operatorname{Var}(a^\intercal X) = \sum_{i,j} a_i a_j \operatorname{Cov} X_i, X_j = a^\intercal \operatorname{Var}(X)\, a$$

where we define $\operatorname{Cov} X, Y \equiv \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$, and $\operatorname{Var}(X)$ is the *variance-covariance matrix* with entries $(\operatorname{Var}(X))_{ij} = \operatorname{Cov} X_i, X_j$. We can say that the variance is bilinear.

§1.3 Standardised statistics

Suppose that $X_1, \dots, X_n$ are i.i.d. and $\mathbb{E}[X_1] = \mu$, $\operatorname{Var}(X_1) = \sigma^2$. We define

$$S_n = \sum_i X_i; \quad \overline{X_n} = \frac{S_n}{n}$$

where $\overline{X_n}$ is called the *sample mean*. By linearity of expectation and bilinearity of variance,

$$\mathbb{E}\left[\overline{X_n}\right] = \mu; \quad \operatorname{Var}\left(\overline{X_n}\right) = \frac{\sigma^2}{n}$$

We further define

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

which has the properties that

$$\mathbb{E}\left[\overline{Z}_n\right] = 0; \quad \operatorname{Var}(Z_n) = 1$$

§1.4 Moment generating functions

The *moment generating function* of a random variable X is the function $M_X(t) = \mathbb{E}\left[e^{tX}\right]$, provided that this function exists for t in some neighbourhood of zero, This can be thought of as the Laplace transform of the probability density function. Note that

$$\mathbb{E}[X^n] = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\, M_X(t)\bigg|_{t=0}$$

Under broad conditions, moment generating functions uniquely define a distribution function of a random variable. In other words, the Laplace transform is invertible. They are also useful for finding the distribution of sums of independent random variables. For instance, let $X_1, \dots, X_n$ be i.i.d. Poisson random variables with parameter μ. Then, the moment generating function of X_i is

$$M_{X_1}(t) = \mathbb{E}\left[e^{tX_i}\right] = \sum_{x=0}^{\infty} e^{tx} e^{-\mu} \frac{\mu^x}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{(e^t \mu)^x}{x!} = e^{-\mu} e^{\mu e^t} = e^{-\mu(1-e^t)}$$

Now,

$$M_{S_n}(t) = \mathbb{E}\left[e^{tS_n}\right] = \prod_{i=1}^{n} \mathbb{E}\left[e^{tX_i}\right] = e^{-n\mu(1-e^t)}$$

This defines a Poisson distribution with parameter $n\mu$ by inspection.

§1.5 Limit theorems

The *weak law of large numbers* states that for all $\varepsilon > 0$, $\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \varepsilon\right) \to 0$ as $n \to \infty$. Note that the event $\left|\overline{X}_n - \mu\right| > \varepsilon$ depends only on $X_1, \dots, X_n$.

The *strong law of large numbers* states that $\mathbb{P}\left(\overline{X}_n \to \mu\right) = 1$. In this formulation, the event depends on the whole sequence of random variables X_i, since the limit is inside the probability calculation.

The *central limit theorem* states that $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ is approximately a $\mathrm{N}(0,1)$ random variable when n is large. More precisely, $\mathbb{P}\left(Z_n \leq z\right) \to \Phi(z)$ for all $z \in \mathbb{R}$.

§1.6 Conditional probability

If X, Y are discrete random variables, we can define the conditional probability mass function to be

$$p_{X|Y}(x \mid y) = \frac{\mathbb{P}\left(X = x, Y = y\right)}{\mathbb{P}\left(Y = y\right)}$$

when $\mathbb{P}\left(Y = y\right) \neq 0$. If X, Y are continuous, we define the joint probability density function to be $f_{X,Y}(x, y)$ such that

$$\mathbb{P}\left(X \leq x, Y \leq y\right) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y') \,\mathrm{d}y' \,\mathrm{d}x'$$

The conditional probability density function is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, y) \,\mathrm{d}x}$$

The denominator is sometimes referred to as the *marginal probability density function* of Y, written $f_Y(y)$. Now, we can define the conditional expectation by

$$\mathbb{E}\left[X \mid Y\right] = \begin{cases} \sum_x x p_{X|Y}(x \mid Y) & \text{if } X \text{ discrete} \\ \int_x x f_{X|Y}(x \mid Y) \,\mathrm{d}x & \text{if } X \text{ continuous} \end{cases}$$

The conditional expectation is itself a random variable, as it is a function of the random variable Y. The conditional variance is defined similarly, and is a random variable. The *tower property* is that

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

The *law of total variance* is that

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

§1.7 Change of variables in two dimensions

Suppose that $(x, y) \mapsto (u, v)$ is a differentiable bijection from $\mathbb{R}^2$ to itself. Then, the joint probability density function of U, V can be written as

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

where J is the Jacobian matrix,

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}$$

§1.8 Common distributions

X has the binomial distribution with parameters n, p if X represents the number of successes in n independent Bernoulli trials with parameter p.

X has the multinomial distribution with parameters $n; p_1, \dots, p_k$ if there are n independent trials with k types, where p_j is the probability of type j in a single trial. Here, X takes values in $\mathbb{N}^k$, and X_j is the amount of trials with type j. Each X_j is marginally binomially distributed.

X has the negative binomial distribution with parameters k, p if, in i.i.d. Bernoulli trials with parameter p, the variable X is the time at which the kth success occurs. The negative binomial with parameter $k = 1$ is the geometric distribution.

The Poisson distribution with parameter λ is the limit of the distribution $\mathrm{Bin}(n, \lambda/n)$ as $n \to \infty$.

If $X_i \sim \Gamma(\alpha_i, \lambda)$ for $i = 1, \ldots, n$ with $X_1, \ldots, X_n$ independent, then the distribution of S_n is given by the product of the moment generating functions. By inspection,

$$M_{S_n}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\sum_i \alpha_i}$$

or ∞ if $t \geq \lambda$. Hence the sum of these random variables is $S_n \sim \Gamma(\sum_i \alpha_i, \lambda)$, where the shape parameter α is constructed from the sum of the shape parameters of the original functions. We call λ the rate parameter, and λ^{-1} is called the scale parameter. If $X \sim \Gamma(\alpha, \lambda)$, then for all $b > 0$ we have $bX \sim \Gamma(x, \lambda/b)$. Special cases of the Γ distribution include:

- $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$;
- $\Gamma(k/2, 1/2) = \chi_k^2$ with k degrees of freedom, which is the distribution of a sum of k i.i.d. squared standard normal random variables.

§2 Estimation

§2.1 Estimators

Suppose $X_1, \dots, X_n$ are i.i.d. observations with a p.d.f. (or p.m.f.) $f_X(x \mid \theta)$, where θ is an unknown parameter in some parameter space Θ. Let $X = (X_1, \dots, X_n)$.

Definition 2.1 (Estimator)
An **estimator** is a statistic, or a function of the data, written $T(X) = \hat{\theta}$, which is used to approximate the true value of θ. This does not depend (explicitly) on θ. The distribution of $T(X)$ is called its **sampling distribution**.

Example 2.1
Let $X_1, \dots, X_n \sim N(0, 1)$ be i.i.d. Let $\hat{\mu} = T(X) = \overline{X}_n$. The sampling distribution is $T(X) \sim N\left(\mu, \frac{1}{n}\right)$. Note that this sampling distribution in general depends on the true parameter μ.

Definition 2.2 (Bias)
The **bias** of $\hat{\theta}$ is

$$\operatorname{bias}\left(\hat{\theta}\right) = \mathbb{E}_\theta\left[\hat{\theta}\right] - \theta$$

Note that $\hat{\theta}$ is a function only of $X_1, \dots, X_n$, and the expectation operator $\mathbb{E}_\theta$ assumes that the true value of the parameter is θ.

Remark 2. In general, the bias is a function of the true parameter θ, even though it is not explicit in the notation.

Definition 2.3 (Unbiased Estimator)
An estimator with zero bias for all θ is called an **unbiased estimator**.

Example 2.2
The estimator $\hat{\mu}$ in the above example is unbiased, since

$$\mathbb{E}_\mu\left[\hat{\mu}\right] = \mathbb{E}_\mu\left[\overline{X}_n\right] = \mu$$

for all $\mu \in \mathbb{R}$.

Definition 2.4 (Mean Squared Error)

The **mean squared error** of θ is defined as

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

Remark 3. Like the bias, the mean squared error is, in general, a function of the true parameter θ.

§2.2 Bias-variance decomposition

The mean squared error can be written as

$$\operatorname{mse}\left(\hat{\theta}\right) = \mathbb{E}_\theta\left[\left(\hat{\theta} - \mathbb{E}_\theta\left[\hat{\theta}\right] + \mathbb{E}_\theta\left[\hat{\theta}\right] - \theta\right)^2\right] = \operatorname{Var}_\theta\left(\hat{\theta}\right) + \operatorname{bias}^2\left(\hat{\theta}\right)$$

Note that both the variance and bias squared terms are positive. This implies a tradeoff between bias and variance when minimising error.

Example 2.3

Let $X \sim \operatorname{Bin}(n, \theta)$ where n is known and θ is an unknown probability. Let $T_U = X/n$. This is the proportion of successes observed. This is an unbiased estimator, since $\mathbb{E}_\theta\left[T_U\right] = \mathbb{E}_\theta\left[X\right]/n = \theta$. The mean squared error for the estimator is then

$$\operatorname{Var}_\theta\left(T_n\right) = \operatorname{Var}_\theta\left(\frac{X}{n}\right) = \frac{\operatorname{Var}_\theta\left(X\right)}{n^2} = \frac{\theta(1-\theta)}{n}$$

Now, consider an alternative estimator which has some bias:

$$T_B = \frac{X+1}{n+2} = w\underbrace{\frac{X}{n}}_{T_U} + (1-w)\frac{1}{2}; \quad w = \frac{n}{n+2}$$

This interpolates between the estimator T_U and the fixed estimator $\frac{1}{2}$. Here,

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

The bias is nonzero for all but one value of θ. Further,

$$\operatorname{Var}_\theta\left(T_B\right) = \frac{\operatorname{Var}_\theta\left(X+1\right)}{(n+2)^2} = \frac{n\theta(1-\theta)}{(n+2)^2}$$

We can calculate

$$\operatorname{mse}(T_B) = (1-w)^2\left(\frac{1}{2}-\theta\right)^2 + w^2\underbrace{\frac{\theta(1-\theta)}{n}}_{\operatorname{mse}(T_U)}$$

There exists a range of θ such that T_B has a lower mean squared error, and similarly there exists a range such that T_U has a lower error. This indicates that prior judgement of the true value of θ can be used to determine which estimator is better.

It is not necessarily desirable that an estimator is unbiased.

Example 2.4

Suppose $X \sim \text{Poisson}(\lambda)$ and we wish to estimate $\theta = \mathbb{P}(X=0)^2 = e^{-2\lambda}$. For some estimator $T(X)$ of θ to be unbiased, we need that

$$\mathbb{E}_\lambda[T(X)] = \sum_{x=0}^{\infty} T(x)\frac{\lambda^x e^{-\lambda}}{x!} = e^{-2\lambda}$$

Hence,

$$\sum_{x=0}^{\infty} T(x)\frac{\lambda^x}{x!} = e^{-\lambda}$$

But $e^{-\lambda}$ has a known power series expansion, giving $T(X) \equiv (-1)^X$ for all X. This is not a good estimator, for example because it often predicts negative numbers for a positive quantity.

§2.3 Sufficiency

Definition 2.5 (Sufficiency)

A statistic $T(X)$ is **sufficient** for θ if the conditional distribution of X given $T(X)$ does not depend on θ. Note that θ and $T(X)$ may be vector-valued, and need not have the same dimension.

Example 2.5

Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter θ where $\theta \in [0,1]$. The mass function is

$$f_X(x \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Note that this dependent only on x via the statistic $T(X) = \sum_{n=1}^{n} x_i$. Here,

$$f_{X|T=t}(x \mid \theta) = \frac{\mathbb{P}_\theta\left(X = x, T(X) = t\right)}{\mathbb{P}_\theta\left(T(x) = t\right)}$$

If $\sum x_i = t$, we have

$$f_{X|T=t}(x \mid \theta) = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-\sum x_i}} = \frac{1}{\binom{n}{t}}$$

Hence $T(X)$ is sufficient for θ.

§2.4 Factorisation criterion

Theorem 2.1

T is sufficient for θ if and only if

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

for suitable functions g, h.

Proof. This will be proven in the discrete case; the continuous case can be handled analogously. Suppose that the factorisation criterion holds. Then, if $T(x) = t$,

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

which does not depend on θ. By definition, $T(X)$ is sufficient.

Conversely, suppose that $T(X)$ is sufficient.

$$\begin{aligned} f_X(x \mid \theta) &= \mathbb{P}_\theta\left(X = x\right) \\ &= \mathbb{P}_\theta\left(X = x, T(X) = T(x)\right) \\ &= \underbrace{\mathbb{P}_\theta\left(X = x \mid T(X) = T(x)\right)}_{h(x)} \underbrace{\mathbb{P}_\theta\left(T(X) = T(x)\right)}_{g(T(X), \theta)} \end{aligned}$$

□

Example 2.6

Consider the above example with n Bernoulli random variables with mass function

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

Let $T(X) = \sum x_i$, and then the above mass function is in the form of $g(T(X), \theta)$ and we can set $h(x) \equiv 1$. Hence $T(X)$ is sufficient.

Example 2.7

Let $X_1, \dots, X_n$ be i.i.d. from a uniform distribution on the interval $[0, \theta]$ for some $\theta > 0$. The mass function is

$$f_X(x \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}\{x_i \in [0,\theta]\} = \left(\frac{1}{\theta}\right)^n \mathbb{1}\left\{\min_i x_i \geq 0\right\} \mathbb{1}\left\{\max_i x_i \leq \theta\right\}$$

Let $T(X) = \max_i X_i$. Then

$$g(T(X), \theta) = \left(\frac{1}{\theta}\right)^n \mathbb{1}\left\{\max_i x_i \leq \theta\right\}; \quad h(x) \equiv \mathbb{1}\left\{\min_i x_i \geq 0\right\}$$

We can then conclude that $T(X)$ is sufficient for θ.

§2.5 Minimal sufficiency

Sufficient statistics are not unique. For instance, any bijection applied to a sufficient statistic is also sufficient. Further, $T(X) = X$ is always sufficient. We instead seek statistics that maximally compress and summarise the relevant data in X and that discard extraneous data.

Definition 2.6 (Minimal Sufficiency)

A sufficient statistic $T(X)$ for θ is **minimal** if it is a function of every other sufficient statistic for θ. More precisely, if $T'(X)$ is sufficient, $T'(x) = T'(y) \implies T(x) = T(y)$.

Remark 4. Any two minimal statistics S, T for the same θ are bijections of each other. That is, $T(x) = T(y)$ if and only if $S(x) = S(y)$.

Theorem 2.2

Suppose that $f_X(x \mid \theta)/f_X(y \mid \theta)$ is constant in θ if and only if $T(x) = T(y)$. Then T is minimal sufficient.

Remark 5. This theorem essentially states the following. Let $x \overset{1}{\sim} y$ if the above ratio of probability density or mass functions is constant in θ. This is an equivalence relation. Similarly, we can define $x \overset{2}{\sim} y$ if $T(x) = T(y)$. This is also an equivalence relation. The hypothesis in the theorem is that the equivalence classes of $\overset{1}{\sim}$ and $\overset{2}{\sim}$ are equal. Further, we may always construct a minimal sufficient statistic for any parameter since we can use the construction $\overset{1}{\sim}$ to create equivalence classes, and set T to be constant for all such equivalence classes.

Proof. Let $t \in \operatorname{Im} T$. Then let z_t be a representative of the equivalence class $\{x \colon T(x) = t\}$. Then

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

By the hypothesis, the ratio on the right hand side does not depend on θ, so let this ratio be $h(x)$. Further, the other term depends only on $T(x)$, so it may be $g(T(x), \theta)$. Hence T is sufficient by the factorisation criterion.

To prove minimality, let S be any other sufficient statistic, and then by the factorisation criterion there exist g_S and h_S such that $f_X(x \mid \theta) = g_S(S(x), \theta) h_S(x)$. Now, suppose $S(x) = S(y)$ for some x, y. Then,

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

which is constant in θ. Hence, $x \overset{1}{\sim} y$. By the hypothesis, we have $x \overset{2}{\sim} y$, so $T(x) = T(y)$, which is the requirement for minimality. $\square$

Remark 6. Sometimes the range of X depends on θ (e.g. $X_1, \dots, X_n \overset{iid}{\sim} \mathrm{Unif}([0, \theta])$). In this case we can interpret "$\frac{f_X(x|\theta)}{f_X(y|\theta)}$ constant in θ" to mean that $f_X(x \mid \theta) = c(x, y) f_X(y \mid \theta)$ for some function c which does not depend on θ.

Example 2.8

Let $X_1, \dots, X_n$ be normal with unknown μ, σ^2.

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

Hence, for minimality, this is constant in the parameters μ, σ^2 if and only if $\sum_i x_i^2 =$

$\sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$. Thus, a minimal sufficient statistic is $(\sum_i x_i^2, \sum_i x_i)$ is a minimal sufficient statistic. A more common way of expressing the minimal sufficient statistic is

$$S(x) = \left(\overline{X}_n, S_{xx}\right); \quad \overline{X}_n = \frac{1}{n}\sum_i x_i; \quad S_{xx} = \sum_i \left(X_i - \overline{X}_n\right)^2$$

which is a bijection of the above minimal sufficient statistic so is also minimal sufficient.

Remark 7. θ and a minimal statistic T need not have the same dimension.

Example 2.9

Consider $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \mu^2)$. Here, there is a single parameter μ but the minimal sufficient statistic is still $S(x)$ as defined above.

§2.6 Rao-Blackwell theorem

Previously, the notation $\mathbb{E}_\theta$ and $\mathbb{P}_\theta$ have been used to denote expectations and probabilities under the model where the observations are i.i.d. with p.d.f. or p.m.f. f_X. From now, we omit this subscript, as it will be implied for much of the remainder of the course.

Theorem 2.3

Let T be a sufficient statistic for θ, and define an estimator $\widetilde{\theta}$ with $\mathbb{E}\left[\widetilde{\theta}^2\right] < \infty$ for all θ. Now we define another estimator

$$\hat{\theta} = \mathbb{E}\left[\widetilde{\theta} \mid T(x)\right]$$

Then, for all values of θ, we have

$$\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \leq \mathbb{E}\left[\left(\widetilde{\theta} - \theta\right)^2\right]$$

In other words, the mean squared error of $\hat{\theta}$ is not greater than the mean squared error of $\widetilde{\theta}$. Further, the inequality is strict unless $\widetilde{\theta}$ is a function of T.

Remark 8. Starting from any estimator $\widetilde{\theta}$, if we condition on the sufficient statistic T we obtain a 'better' statistic $\hat{\theta}$. Note that T must be sufficient, otherwise $\hat{\theta}$ may be a function of θ and thus not an estimator:

$$\hat{\theta}(X) = \hat{\theta}(T) = \int \hat{\theta}(x) \underbrace{f_{X|T}(x \mid T)}_{\text{does not depend on } \theta \text{ as } T \text{ is sufficient}} \mathrm{d}x$$

The message to take away from this theorem is that we can improve the mse of any estimate $\tilde{\theta}$ by taking a conditional expectation given $T(x)$.

Proof. By the tower property of the expectation, we can find

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

Hence, $\text{bias}\left(\hat{\theta}\right) = \text{bias}\left(\tilde{\theta}\right)$. By the conditional variance formula,

$$\text{Var}\left(\tilde{\theta}\right) = \mathbb{E}\left[\underbrace{\text{Var}\left(\tilde{\theta} \mid T\right)}_{\geq 0}\right] + \underbrace{\text{Var}\left(\mathbb{E}\left[\tilde{\theta} \mid T\right]\right)}_{\text{Var}(\hat{\theta})} \geq \text{Var}\left(\hat{\theta}\right) \quad \forall\, \theta.$$

By the bias-variance decomposition, we know that $\text{mse}\left(\tilde{\theta}\right) \geq \text{mse}\left(\hat{\theta}\right)$. The inequality is strict unless $\text{Var}\left(\tilde{\theta} \mid T\right) = 0$ almost surely. This requires that $\tilde{\theta}$ is a function of T. □

Example 2.10

Let $X_1, \dots, X_n$ be i.i.d. Poisson random variables with parameter λ. Then let $\theta = \mathbb{P}\left(X_1 = 0\right) = e^{-\lambda}$. Here,

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

Using the factorisation criterion, we find

$$g(T(x), \theta) = g\left(\sum x_i, \theta\right) = \theta^n(-\log\theta)^{\sum x_i}; \quad h(x) = \frac{1}{\prod x_i!}$$

so $T(x) = \sum x_i$ is sufficient.

Note that $\sum X_i$ has a Poisson distribution with parameter $n\lambda$.

Consider the estimator $\tilde{\theta} = \mathbb{1}\{X_1 = 0\}$. This depends only on X_1, hence it is a weak estimator. However, it is unbiased, so when we apply the Rao-Blackwell theorem we will construct an unbiased $\hat{\theta}$, which is precisely

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

$$
\vdots
$$
$$
= \left(\frac{n-1}{n}\right)^t
$$

This may also be written

$$
\hat{\theta} = \left(1 - \frac{1}{n}\right)^{\sum x_i}
$$

which is an estimator with strictly lower mean squared error than $\widetilde{\theta}$ for all θ by Rao-Blackwell and as $\widetilde{\theta}$ doesn't depend solely upon T.

Note that $\hat{\theta} = \left(1 - \frac{1}{n}\right)^{n\overline{X}_n}$ converges in the limit to $e^{-\overline{X}_n}$. By the strong law of large numbers, $\overline{X}_n \to \mathbb{E}[X_1] = \lambda$ almost surely, so we arrive at $\hat{\theta} \to e^{-\lambda} = \theta$ almost surely.

[a] We know the distribution of $\sum X_i$, so simply sub this pmf in.

Example 2.11

Let $X_1, \dots, X_n$ be i.i.d. uniform random variables in an interval $[0, \theta]$. We wish to estimate $\theta \geq 0$. We observed that $T = \max X_i$ is sufficient for θ.
Let $\widetilde{\theta} = 2X_1$. This is an unbiased estimator of θ. Then the Rao-Blackwellised estimator $\hat{\theta}$ is

$$
\begin{aligned}
\hat{\theta} &= \mathbb{E}\left[\widetilde{\theta} \mid T = t\right] \\
&= 2\mathbb{E}\left[X_1 \mid \max X_i = t\right] \\
&= 2\mathbb{E}\left[X_1 \mid \max X_i = t, X_1 = \max X_i\right] \mathbb{P}\left(X_1 = \max X_i \mid \max X_i = t\right) \\
&+ 2\mathbb{E}\left[X_1 \mid \max X_i = t, X_1 \neq \max X_i\right] \mathbb{P}\left(X_1 \neq \max X_i \mid \max X_i = t\right)
\end{aligned}
$$

Since $X_1, \dots, X_n$ are i.i.d., the conditional probability $\mathbb{P}\left(X_1 = \max X_i \mid \max X_i = t\right)$ can be reduced to $\mathbb{P}\left(X_1 = \max X_i\right) = \frac{1}{n}$. The complementary event may be reduced in an analogous way. The expectation $\mathbb{E}\left[X_1 \mid \max X_i = t, X_1 = \max X_i\right]$ can be reduced to t.

$$
\begin{aligned}
\hat{\theta} &= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}\left[X_1 \mid X_1 < t, \max_{i=2}^{n} X_i = t\right] \\
&= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}\left[X_1 \mid X_1 < t\right]^{a} \\
&= \frac{2t}{n} + \frac{2(n-1)}{n}\frac{t}{2} \\
&= \frac{2t}{n} + \frac{t(n-1)}{n} = \frac{n+1}{n}\max_i X_i
\end{aligned}
$$

By the Rao-Blackwell theorem, the mean squared error of $\hat{\theta}$ is strictly better than the mean squared error of $\widetilde{\theta}$. This is also an unbiased estimator.

[a]By independence

§2.7 Maximum likelihood estimation

Let $X_1, \dots, X_n$ be i.i.d. random variables with mass or density function $f_X(x \mid \theta)$.

Definition 2.7 (Likelihood Function)

For fixed observations x, the **likelihood function** $L\colon \Theta \to \mathbb{R}$ is given by

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

Definition 2.8 (Log-Likelihood Function)

We will denote the **log-likelihood** by

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

Definition 2.9

A **maximum likelihood estimator** is an estimator that maximises the likelihood function L over Θ. Equivalently, the estimator maximises ℓ.

Example 2.12

Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter p. The log-likelihood function is

$$\ell(p) = \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1-p)] = \log p \sum X_i + \log(1-p)\Big(n - \sum X_i\Big)$$

The derivative is

$$\ell'(p) = \frac{\sum X_i}{p} + \frac{n - \sum X_i}{1-p}$$

which has a single stationary point at $p = \frac{1}{n} \sum X_i = \overline{X}$. We have $\mathbb{E}\left[\hat{p}\right] = p$, so the maximum likelihood estimator in this case is unbiased.

Example 2.13

Let $X_1, \dots, X_n$ be i.i.d. normal random variables with unknown mean μ and variance σ^2.

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

This function is concave in μ and σ^2, so there exists a unique maximiser. In particular, ℓ is maximised when $\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial \sigma^2} = 0$.

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

This is zero if $\mu = \overline{X}$.
Further,

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(X_i - \mu)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(X_i - \overline{X})^2$$

This is zero iff

$$\sigma^2 = \frac{1}{n}\sum(X_i - \overline{X})^2 = \frac{S_{xx}}{n}$$

Hence, the maximum likelihood estimator is $(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{X}_n, \frac{1}{n}S_{xx}\right)$.

We can show that $\hat{\mu} = \overline{X}$ is unbiased.

We will later prove that

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

Hence

$$\mathbb{E}\left[\hat{\sigma}^2\right] = \frac{\sigma^2}{n}\mathbb{E}\left[\chi^2_{n-1}\right] = \sigma^2\frac{n-1}{n} \neq \sigma^2$$

This is therefore a biased estimator, but the bias converges to zero as $n \to \infty \; \forall \, \sigma^2$: $\hat{\sigma}^2$ is **asymptotically unbiased**.

Example 2.14

Let $X_1, \dots, X_n$ be i.i.d. uniform random variables on $[0, \theta]$. Here, we derived the unbiased estimator $\hat{\theta} = \frac{n+1}{n}\max X_i$.

The likelihood is given by

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

This function is maximised at $\hat{\theta}_{\text{mle}} = \max X_i$.

By comparison to the $\hat{\theta}$ derived from the Rao-Blackwell process, $\hat{\theta}_{\text{mle}}$ is biased but asymptotically unbiased. In particular,

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

Remark 9.

1. If T is a sufficient statistic for θ, then the maximum likelihood estimator is a function of $T(X)$. Indeed, since X and $T(X)$ are fixed, the maximiser of $L(\theta) = g(T(X), \theta)h(X)$ depends on X only through T. This is good as otherwise we could use Rao-Blackwell to get a better estimator in terms of the mse.
2. If $\varphi = H(\theta)$ for a bijection H, then if $\hat{\theta}$ is the maximum likelihood estimator for θ, we have that $H(\hat{\theta})$ is the maximum likelihood estimator for φ.
3. Asymptotic Normality: Under some regularity conditions, as $n \to \infty$ the statistic $\sqrt{n}(\hat{\theta} - \theta)$ is approximately normal with mean zero and covariance matrix Σ. More precisely, for 'nice' sets A and 'regular' values of θ, we have

   $$\mathbb{P}\left(\sqrt{n}\left(\hat{\theta}(n) - \theta\right) \in A\right) \overset{n\to\infty}{\to} \mathbb{P}\left(Z \in A\right); \quad Z \sim N(0, \Sigma)$$

   We say that the maximum likelihood estimator is *asymptotically normal*. The limiting covariance matrix Σ is a known function of ℓ, which will not be defined in this course. There is a theorem (Cramer-Rao) which says in some sense, Σ is the smallest variance that any estimator can achieve asymptotically.
4. For practical purposes, this estimator can often be found numerically by maximising ℓ or L.

§3 Inference

§3.1 Confidence Intervals

Question

A vaccine has 76% efficacy in a 3-month period, with a 95% confidence interval $(59\%, 86\%)$. What does this mean?

Definition 3.1

A $100\gamma\%$ **confidence interval** for a parameter θ is a random interval $(A(X), B(X))$ such that $\mathbb{P}\left(A(X) \le \theta \le B(X)\right) = \gamma$ for all $\theta \in \Theta$. Note that the parameter θ is assumed to be fixed for the event $\{A(X) \le \theta \le B(X)\}$, and the confidence interval holds uniformly over θ.

Answer

There exist some fixed true parameter θ. Suppose that an experiment is repeated many times. On average, $100\gamma\%$ of the time, the random interval $(A(X), B(X))$ will contain the true parameter θ. This is the *frequentist* interpretation of the confidence interval.

A misleading interpretation is as follows. Given that a single value of $X = x$ is observed, there is a probability γ that $\theta \in (A(x), B(x))$. This is wrong, as will be demonstrated later.

Example 3.1

Let $X_1, \dots, X_n \sim \mathcal{N}(\theta, 1)$ be iid. We will find the 95% confidence interval for θ. We have

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\theta, \frac{1}{n}\right); \quad Z = \sqrt{n}\left(\overline{X} - \theta\right) \sim \mathcal{N}(0,1)$$

Z has this distribution $\forall\, \theta$.

Let a, b be numbers such that $\Phi(b) - \Phi(a) = 0.95$. Then

$$\mathbb{P}\left(a \le \sqrt{n}\left(\overline{X} - \theta\right) \le b\right) = 0.95 \implies \mathbb{P}\left(\overline{X} - \frac{b}{\sqrt{n}} \le \theta \le \overline{X} - \frac{a}{\sqrt{n}}\right) = 0.95$$

Hence, $\left(\overline{X} - \frac{b}{\sqrt{n}}, \overline{X} - \frac{a}{\sqrt{n}}\right)$ is a 95% confidence interval for θ.

Typically, we wish to centre the interval around some estimator $\hat{\theta}$ such that its range is minimised for a given γ. In this case, we want to set $-a = b = z_{0.025} \approx 1.96$, where $z_\alpha = \Phi^{-1}(1-\alpha)$. Hence, the confidence interval is $\left(\overline{X} \pm \frac{1.96}{\sqrt{n}}\right)$.

Remark 10. In general, to find a confidence interval:

1. Find a quantity $R(X, \theta)$ where the distribution $\mathbb{P}_\theta$ does not depend on θ. This is known as a *pivot*. In the example above, $R(X, \theta) = \sqrt{n}\left(\overline{X} - \theta\right)$.
2. Consider $\mathbb{P}\left(c_1 \leq R(X, \theta) \leq c_2\right) = \gamma$. Given some desired level of confidence γ, find c_1 and c_2 using the distribution function of the pivot.
3. Rearrange such that $\mathbb{P}\left(A(X) \leq \theta \leq B(X)\right) = \gamma$, then $(A(X), B(X))$ is the confidence interval as required.

Proposition 3.1

Let T be a monotonically increasing function, and let $(A(X), B(X))$ be a $100\gamma\%$ confidence interval for θ. Then $(T(A(X)), T(B(X)))$ is a $100\gamma\%$ confidence interval for $T(\theta)$.

Remark 11. If θ is a vector, we can consider confidence sets instead of confidence intervals. A confidence set is a set $A(X)$ such that $\mathbb{P}\left(\theta \in A(X)\right) = \gamma$.

Example 3.2

Let $X_1, \ldots, X_n$ be i.i.d. normal random variables with zero mean and unknown variance σ^2. We will find a 95% confidence interval for σ^2. Note that $\frac{X_1}{\sigma} \sim N(0,1)$ is a valid pivot, but it considers only one data point. We will instead consider

$$R(X, \sigma^2) = \sum_i \frac{X_i^2}{\sigma^2} \sim \chi_n^2$$

Now, we can define $c_1 = F_{\chi_n^2}^{-1}(0.025)$ and $c_2 = F_{\chi_n^2}^{-1}(0.975)$, giving

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

Rearranging, we have

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

Hence, the interval $\sum_{i=1}^n X_i^2\left(\frac{1}{c_2}, \frac{1}{c_1}\right)$ is a 95% confidence interval for σ^2.

Example 3.3

Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with parameter p. Suppose n is large. We will find an approximate 95% confidence interval for p. The maximum likelihood estimator is

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

By the central limit theorem, $\hat{p}$ is asymptotically distributed according to $N\left(p, \frac{p(1-p)}{n}\right)$. Hence,

$$\sqrt{n}\frac{\hat{p} - p}{\sqrt{p(1-p)}}$$

has approximately a standard normal distribution. We have

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

Instead of directly rearranging the inequalities, we will make an approximation for the denominator of the central term, letting $\sqrt{p(1-p)} \mapsto \sqrt{\hat{p}(1-\hat{p})}$. When n is large, this approximation becomes more accurate.

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

This is much easier to rearrange, leading to

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

This gives the approximate 95% confidence interval as required.

Remark 12. Note that the size of the confidence interval is maximised at $p = \frac{1}{2}$, with a length of $2z_{0.025}\frac{1}{2\sqrt{n}} \approx \frac{1}{\sqrt{n}}$. This is a *conservative* 95% confidence interval; it may be wider than necessary but holds for all values of θ.

§3.2 Interpreting the confidence interval

Example 3.4

Let X_1, X_2 be i.i.d. uniform random variables in $\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. We wish to estimate the value of θ with a 50% confidence interval. Observe that

$$\mathbb{P}\left(\theta \in (\min X_i, \max X_i)\right) = \mathbb{P}\left(X_1 \leq \theta \leq X_2\right) + \mathbb{P}\left(X_2 \leq \theta \leq X_1\right) = \frac{1}{2}$$

Hence, $(\min X_1, \max X_i)$ is a 50% confidence interval for θ. The frequentist interpretation is exactly correct; 50% of the time, θ will lie between X_1 and X_2. However, suppose that $|X_1 - X_2| > \frac{1}{2}$. Then we know that $\theta \in (\min X_i, \max X_i)$. Suppose $X_1 = 0.1, X_2 = 0.9$, then it is not sensible to say that there is a 50% chance that $\theta \in [0.1, 0.9]$.