IB Statistics

Ishan Nath, Lent 2023

Based on Lectures by Dr. Sergio Bacallado

January 26, 2023

Contents

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

- Estimating θ,
- Quantifying uncertainty in estimates,
- Testing a hypothesis about θ.

1.1 Probability Review

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

A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ 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 $X : \Omega \to \mathbb{R}$.

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

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

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

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

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

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

The *variance* of X is

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

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

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

If 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 $(X_1, \ldots, X_n)$ and x is the vector $(x_1, \ldots, x_n)$.

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

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

Moreover,

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

Here the *covariance* of X_i and X_j is

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

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

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

and moreover

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

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

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

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

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(μ). 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 $S_n = X_1 + \cdots + X_n$, then

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

This is the same as a Poisson(μn) 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 $\varepsilon > 0$,

$$\mathbb{P}(|\overline{X_n} - \mu| > \varepsilon) \to 0,$$

as $n \to \infty$.

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

$$\mathbb{P}(\overline{X_n} \to \mu) = 1,$$

i.e. $\overline{X_n}$ converges to μ 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 $\sigma^2 = \operatorname{Var}(X_i)$, then Z_n is "approximately" $N(0,1)$ as $n \to \infty$.

What this means is that $\mathbb{P}(Z_n \leq z) \to \Phi(z)$ as $n \to \infty$ for all $z \in \mathbb{R}$, where Φ 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

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

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

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

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

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* is given by

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

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

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

Hence we can write the variance of X as follows:

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

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

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

$X \sim \mathrm{Poisson}(\lambda)$ is the limit of $\mathrm{Bin}(n, \lambda/n)$ random variables, 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 if $S_n = X_1 + \cdots + X_n$,

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

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 $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$, 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 $X_1, X_2, \ldots, X_n$, which are iid from some pdf (or pmf) $f_X(x \mid \theta)$, with θ unknown. We let $X = (X_1, \ldots, X_n)$.

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 θ. The distribution of $T(X)$ is called the *sampling distribution*.

Example 2.1.

If $X_1, \ldots, X_n$ 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 θ, 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.

Out 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 θ.

2.1 Bias-Variance Decomposition

We can write the mean squared error as

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

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 $X \sim \mathrm{Bin}(n,\theta)$, where n is known, and we wish to estimate θ. 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

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

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

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

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

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

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

The message here is that our prior judgements about θ 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 λ,

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

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) $f_X(\cdot \mid \theta)$. 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 θ?

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

Note θ 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,

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

This only depends on X through

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

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

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

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

Theorem 2.1 (Factorization criterion)**.** *T is sufficient for θ if and only if*

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

for suitable functions g, h.

Proof: We only do the discrete case.

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

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

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

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

Therefore the pmf of X factorizes.

Example 2.6.

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

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

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

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

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

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 θ, if and only if $T(x) = T(y)$. Then T is minimal sufficient.

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

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

is constant in θ. It is easy to check that $\stackrel{1}{\sim}$ is an equivalence relation.

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

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

Remark. We can always construct a statistic T which is constant on the equivalence classes of $\stackrel{1}{\sim}$, 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,

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

This is exactly in the form $g(T(x), t)h(x)$, 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 θ. Hence $x \stackrel{1}{\sim} y$. By the hypothesis, we get that $T(x) = T(y)$.

Remark. Sometimes the range of X depends on θ. In this case we can interpret

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

to mean that

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

for some function c which does not depend on θ.

Index