IB Statistics

Martin von Hodenberg (mjv43@cam.ac.uk)

January 27, 2022

These are my notes for the IB course Statistics, which was lectured in Lent 2022 at Cambridge by Dr S.Bacallado. These notes are written in LaTeX for my own revision purposes. Any suggestions or feedback is welcome.

Contents

§0 Introduction

Statistics can be defined as the science of *making informed decisions*. It can include:

1. Formal statistical inference
2. Design of experiments and studies
3. Visualisation of data
4. Communication of uncertainty and risk
5. Formal decision theory

In this course we will only focus on formal statistical inference.

Definition 0.1 (Parametric inference)

Let $X_1, \dots, X_n$ be iid. random variables. We will assume the distribution of X_1 belongs to some family with parameter $\theta \in \Theta$.

Example 0.2

We will give some examples of such families:

1. $X_1 \sim \text{Po}(\mu), \theta = \mu \in \Theta = (0, \infty)$.
2. $X_1 \sim N(\mu, \sigma^2) \quad N(\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

We will use the observed $X = (X_1, \dots X_n)$ to make inferences about θ such as:

1. Point estimate $\theta(X)$ of θ.
2. Interval estimate of θ: $(\theta_1(x), \theta_2(x))$
3. Testing hypotheses about θ: for example checking if there is evidence in X against the hypothesis $H_0 : \ \theta = 1$.

Remark. In general, we'll assume the distribution of the family $X_1, \dots, X_n$ is known but the parameter is unknown. Some results (on mean square error, bias, Gauss-Markov theorem) will make weaker assumptions.

§1 Probability

First we will briefly recap IA Probability.

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

Definition 1.1 (Probability measure)

A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies:

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

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

Definition 1.2 (Random variable)
A random variable is a (measurable) function $X : \Omega \to \mathbb{R}$.

Example 1.3
Tossing two coins has $\Omega = \{HH, HT, TH, TT\}$. Since Ω is countable, $\mathcal{F}$ is the power set of Ω. We can define X to be the random variable that counts the number of heads. Then

$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$

Definition 1.4 (Distribution function)
The distribution function of X is $F_X(x) = \mathbb{P}(X \leq x)$.

A discrete random variable takes values in a countable set $S \subset \mathbb{R}$. Its probability mass function is

$$p_X(x) = \mathbb{P}(X = x).$$

A random variable X has a continuous distribution if it has a probability density function $f_X(x)$ which satisfies

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

for measurable sets A.

The expectation of X is

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

If $g : \mathbb{R} \to \mathbb{R}$, then for a continuous r.v

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

The variance of X is

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

We say $X_1, \ldots, X_n$ are independent if for all $x_1, \ldots, x_n$ we have

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

If $X_1, \ldots, X_n$ have pdfs or pmfs $f_{X_1}, \ldots, f_{X_n}$ then their joint pdf or pmf is

$$f_X(x) = \prod_i f_{X_i}(x_i).$$

If $Y = \max(X_1, \ldots, X_n)$ independent, then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X_1 \leq y, \ldots, X_n \leq y) = \prod_i F_{X_i}(y).$$

The pdf of Y (if it exists) is obtained by differentiating F_Y.

§1.1 Linear transformations

Let $(a_1, \dots a_n)^T = a \in \mathbb{R}^n$ be a constant.

$$\mathbb{E}(a_1 X_1 + \dots + a_n X_n) = \mathbb{E}(a^T X) = a^T \mathbb{E}(X).$$

This gives linearity of expectation (does not require independence).

$$\mathrm{Var}(a^T X) = \sum_{i,j} a_i a_j \underbrace{\mathrm{Cov}(X_i, X_j)}_{=\mathbb{E}((X_i - \mathbb{E}(X_i)(X_j - \mathbb{E}(X_j))))} = a^T \, \mathrm{Var}(X) a.$$

where the matrix $[\mathrm{Var}(X)]_{ij} = \mathrm{Cov}(X_i, X_j)$. This gives the "bilinearity of variance".

§1.2 Standardised statistics

Let $X_1, \dots, X_n$ be iid. with $\mathbb{E}(X_1) = \mu$, $\mathrm{Var}(X_1) = \sigma^2$. We define $S_n = \sum_i X_i$ and $\overline{X_n} \frac{S_n}{n}$ (the sample mean). By linearity

$$\mathbb{E}(\overline{X_n}) = \mu, \quad \mathrm{Var}(\overline{X_n}) = \frac{\sigma^2}{n}.$$

Define $Z_n = \frac{S_n - n\mu}{n}$. Then $\mathbb{E}(Z_n) = 0$ and $\mathrm{Var}(Z_n) = 1$.

§1.3 Moment generating functions

The mgf of a random variable X is the function

$$M_x(t) = \mathbb{E}(e^{tx}).$$

provided that it exists for t in some neighbourhood of 0. This is the Laplace transform of the pdf. It relates to moments of the pdf, for example $M_x^{(n)}(0) = \mathbb{E}(X^n)$.

Under broad conditions $M_x = M_y \iff F_X = F_Y$. (The Laplace transform is invertible.) The mgf is also useful for finding distributions of sums of independent random variables:

Example 1.5

Let $X_1, \dots, X_n \sim \mathrm{Po}(\mu)$. Then

$$M_{X_i}(t) = \mathbb{E}(e^{tX_i}) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\mu}\mu^x}{x!} = e^{-\mu} \sum_{x=0}^{\infty} \frac{(e^t \mu^x)}{x!} = e^{-\mu(1-e^t)}.$$

What is M_{S_n}? We have

$$M_{S_n}(t) = \mathbb{E}(e^{t(X_1 + \dots + X_n)}) = \prod_{i=1}^{n} e^{tX_i} = e^{-n\mu(1-e^t)}.$$

So we conclude $S_n \sim \mathrm{Po}(n\mu)$

§1.4 Limits of r.v's

The weak law of large numbers states that $\forall \varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}\left(|\overline{X_n} - \mu > \epsilon|\right) \to 0.$$

The strong law of large numbers states that as $n \to \infty$,

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

The central limit theorem states that if we have the variable $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$, then as $n \to \infty$ we have

$$\mathbb{P}(Z_n \leq z) \to \Phi(z) \quad \forall z \in \mathbb{R}.$$

where Φ is the distribution function of a $N(0,1)$ random variable.

§1.5 Conditional probability

If X, Y are discrete r.v's then

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

If X, Y are continuous then the joint pdf of X, Y satisfies:

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

The conditional pdf of X given Y is

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

The conditional expectation of X given Y is

$$\mathbb{E}(X|Y) = \begin{cases} \sum_x x p_{X|Y}(x|Y) & \text{discrete} \\ \int x f_{X|Y}(x|Y)\mathrm{d}x & \text{continuous} \end{cases}$$

Note this is itself a random variable, as it is a function of Y. We define $\mathrm{Var}(X|Y)$ similarly.

Tower property: $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$

Law of total variance: $\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\mathbb{E}(X|Y))$.

Change of variables (in 2D):

Let $(x, y) \mapsto (u, v)$ be a differentiable bijection $\mathbb{R}^2 \to \mathbb{R}^2$. Then

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

where $J = \frac{\partial(x,y)}{\partial(u,v)}$ is the Jacobian matrix we have seen before.

If $X_i \sim \Gamma(\alpha_i, \lambda)$ for $i = 1, \ldots, n$ with $X_1, \ldots X_n$ independent, then what is the distribution of $S_n = \sum_{i=1}^{n} X_i$?

$$M_{S_n}(t) = \prod_i M_{X_i}(t) = \begin{cases} \left(\frac{\lambda}{\lambda t}\right)^{\sum_i \alpha_i} & t < \lambda \\ \infty & t > \lambda \end{cases}.$$

So S_n is $\Gamma(\sum_i a_i, \lambda)$. We call the first parameter the "shape parameter", and the second one the "rate parameter". A consequence of what we have just done is that if $X \sim \Gamma(\alpha, \lambda)$, then for all $b > 0$ we have $bX \sim \Gamma(\alpha, \frac{\lambda}{b})$.

Special cases:

- $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$
- $\Gamma\left(\frac{k}{2}, \frac{1}{2}\right) = \chi_k^2$ (the chi-squared distribution with k degrees of freedom, i.e the distribution of a sum of k independent squared $N(0,1)$ r.v's.)

§1.6 Estimation

Suppose $X_1, \ldots X_n$ are iid observations with pdf or pdf (or pmf) $f_X(x|\theta)$ where θ is an unknown parameter in Θ. Let $X = (X_1, \ldots, X_n)$.

Definition 1.6 (Estimator)
An estimator is a statistic or function of the data $T(X) = \hat{\theta}$ which does not depend on θ, and is used to approximate the true parameter θ. The distribution of $T(X)$ is called its "sampling distribution".

Example 1.7
Let $X_1, \ldots, X_n \sim N(\mu, 1)$ iid. Here $\hat{\mu} = \frac{1}{n}\sum_i X_i = \overline{X_n}$. The sampling distribution of $\hat{\mu}$ is $T(X) = N(\mu, \frac{1}{n})$.

Definition 1.8 (Bias)
The bias of $\hat{\theta} = T(X)$ is

$$\mathrm{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta.$$

Here $\mathbb{E}_\theta$ is the expectation in the model where $X_1, X_2, \ldots, X_n \sim f_X(x|\theta)$.

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

Definition 1.9 (Unbiased estimator)
We say $\hat{\theta}$ is unbiased if $\mathrm{bias}(\hat{\theta}) = 0$ for all values of the true parameter θ.

In our example, $\hat{\mu}$ is unbiased because

$$\mathbb{E}_\mu(\hat{\mu}) = \mathbb{E}_\mu(\overline{X_n}) = \mu \quad \forall \mu \in \mathbb{R}.$$

Definition 1.10 (Mean squared error)
The mean squared error (mse) of θ is

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

It tells us "how far" $\hat{\theta}$ is from θ "on average".

§1.7 Bias-variance decomposition

We expand the square in the definition of mse to get

$$\begin{aligned}\operatorname{mse}(\hat{\theta}) &= \mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right] \\ &= \mathbb{E}_\theta\left((\hat{\theta}-\mathbb{E}_\theta\hat{\theta}-\theta)^2\right) \qquad\qquad = \operatorname{Var}_\theta(\hat{\theta}) + \operatorname{bias}^2(\hat{\theta}) \\ &\geq 0\end{aligned}$$

There is a tradeoff between bias and variance. For example, let $X \sim \operatorname{Bin}(n,\theta)$. Suppose n is known, and $\theta \in [0,1]$ is our unknown parameter. We define $T_u = \frac{X}{n}$, i.e the proportion of successes observed. Clearly T_u is unbiased since

$$\mathbb{E}_\theta(T_u) = \frac{E_\theta(X)}{n} = n\theta/n = \theta.$$

We can caculate

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

Consider another estimator $T_B = \frac{X+1}{n+2} = w\frac{X}{n} + (1-w)\frac{1}{2}$ for $w = \frac{n}{n+2}$. This is called a "fixed estimator". In this case we have

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

This is $\neq 0$ for all but one value of θ. Note that

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

Remark. In this example, there are regions where either estimator is better. Prior judgement on the true value of θ determines which estimator is better.

Unbiasedness is not necessarily desirable. Let's look at a pathological example:

Example 1.11

Suppose $X \sim \operatorname{Po}(\lambda)$. We want to estimate $\theta = \mathbb{P}(X=0)^2 = e^{-2\lambda}$. For some estimator $T(X)$ to be unbiased, we need

$$\mathbb{E}_\lambda(T(x)) = \sum_{x=0}^\infty T(x)\frac{\lambda^x e^{-\lambda}}{x!} = e^{-2\lambda} = \theta \iff \sum_{x=0}^\infty T(x)\frac{\lambda^x}{x!} = e^{-\lambda} = \sum_{x=0}^\infty (-1)^x\frac{\lambda^x}{x!}.$$

The only function $T : N \to \mathbb{R}$ satisfying this equality is $T(x) = (-1)^x$. This is clearly an absurd estimator.

§1.8 Sufficiency

Definition 1.12 (Sufficiency)

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

Remark. θ can be a vector and $T(X)$ can also be vector-valued.

Example 1.13

Let $X_1, \ldots, X_n$ be iid. Bernoulli(θ) variables for some θ. Then

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

This only depends on x through $T(X) = \sum x_i$.