IB Statistics

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

Last updated: March 5, 2022

These are my notes[1] for the IB course Statistics, which was lectured in Lent 2022 at Cambridge by Dr S.Bacallado.

[1] Notes are posted online on my website.

Contents

0 Introduction **3**
- 0.1 Probability . . . 4

1 Estimation **8**
- 1.1 Bias-variance decomposition . . . 8
- 1.2 Sufficiency . . . 9
- 1.3 Minimal sufficiency . . . 11
- 1.4 Rao-Blackwell theorem . . . 12
- 1.5 Maximum likelihood estimation . . . 14
- 1.6 Confidence intervals . . . 16
- 1.7 Bayesian analysis . . . 18

2 Hypothesis testing **22**
- 2.1 Simple hypotheses . . . 22
- 2.2 Composite hypotheses . . . 24

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

Example
We will give some examples of such families:

1. $X_1 \sim \mathrm{Po}(\mu), \theta = \mu \in \Theta = (0, \infty)$.
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 $X = (X_1, \ldots X_n)$ 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 $X_1, \ldots, X_n$ is known but the parameter is unknown. Some results (on mean square error, bias, Gauss-Markov theorem) will make weaker assumptions.

0.1 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 $\mathcal{F}$.

Random variables

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

1. $\mathbb{P}(\emptyset) = 0$
2. $\mathbb{P}(\Omega) = 1$
3. $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i = \sum_i \mathbb{P}(A_i)\right)$ if (A_i) is a sequence of disjoint events.

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

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

Definition (Discrete/continuous random variable)
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.

Definition (Expectation/variance)

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

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

Definition (Independence)
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.

Linear transformations

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

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

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".

Standardised statistics

Let $X_1, \ldots, X_n$ 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 $\mathrm{Var}(Z_n) = 1$.

Moment generating functions

Definition (Moment generating function)
The **moment generating function** (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
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

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

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

Example (mgf of the gamma distribution)
If $X_i \sim \Gamma(\alpha_i, \lambda)$ for $i = 1, \dots, n$ with $X_1, \dots X_n$ 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.)

Limits of random variables

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.

Conditional probability

Definition (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.

There are several notable properties of conditional random variables:

- 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.

1 Estimation

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

Definition (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
Let $X_1, \dots, X_n \sim N(\mu, 1)$ 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 (Bias)
The bias of $\hat{\theta} = T(X)$ is

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

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

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

Definition (Unbiased estimator)
We say $\hat{\theta}$ is unbiased if $\text{bias}(\hat{\theta}) = 0$ 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 (Mean squared error)
The mean squared error (mse) of θ is

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

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

1.1 Bias-variance decomposition

We expand the square in the definition of mse to get

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

There is a tradeoff between bias and variance. For example, let $X \sim \mathrm{Bin}(n, \theta)$. 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

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

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

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

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

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

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

Suppose $X \sim \mathrm{Po}(\lambda)$. 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.

1.2 Sufficiency

Notation. From now on in the course we drop the θ subscript on expectations etc. in order to simplify notation.

Definition (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

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 $T(X) = \sum x_i$. To check it's sufficient:

$$\begin{aligned}
f_{X|T=t}(x|T=t) &= \frac{\mathbb{P}(X=x, T(x)=t)}{\mathbb{P}(T(x)=t)} \\
\text{If } \sum x_i = t, \quad &= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\mathbb{P}(T(x)=t)} \\
&= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \text{ since } \sum X_i \sim Bin(n,\theta) \\
&= \binom{n}{t}^{-1}
\end{aligned}$$

Therefore T is sufficient.

Theorem 1.1 (Factorisation criterion)

T is sufficient for θ iff $f_x(x|\theta) = g(T(x), \theta)h(x)$ for suitable functions g, h.

Proof. We will only prove this for the discrete case; the continuous case is similar.

Reverse implication: Suppose $f_x(x|\theta) = g(T(x),\theta)h(x)$. Then if $T(x) = t$, we have

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

This doesn't depend on θ so T is sufficient.

Forward implication: Suppose $T(X)$ is sufficient. Then we have

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

By noting that $\mathbb{P}(X = x|T(X) = x)$ only depends on x by assumption and $\mathbb{P}(T(X) = T(x))$ only depends on x through $T(x)$, we are done. □

Remark. This criterion makes our previous example much easier.

Let's look at another example.

Example

Let $X_1, \dots, X_n \sim U([0,\theta])$ be iid with $\theta > 0$. Then

$$f_X(x|\theta) = \prod_{i=1}^n \frac{1}{\theta} 1_{x_i \in [0,\theta]} = \frac{1}{\theta^n} 1_{\min_i x_i \geq 0} 1_{\max_i x_i \leq \theta}.$$

Define $T(x) = \max_i x_i$. Then we can write

$$g(T(x), \theta) = \frac{1}{\theta^n} 1_{\max_i x_i \leq \theta}, \quad h(x) = 1_{\min_i x_i \geq 0}.$$

So $T(x)$ is sufficient.

1.3 Minimal sufficiency

Sufficient statistics are **not** unique.

Remark. Any bijection applied to a sufficient statistic yields another sufficient statistic.

It's not hard to find sufficient statistics, for example $T(X) = X$ is a trivial sufficient statistic (that is useless!). Instead, we want statistics which give us 'maximal' compression of the data in X. This motivates our next definition.

Definition (Minimal sufficient statistic)

$T(X)$ is **minimal sufficient** if for every other sufficient statistic T',

$$T'(x) = T'(y) \implies T(x) = T(y) \quad \forall x, y \in \mathcal{X}^n.$$

Note that it follows from this definition that minimal sufficient statistics are unique up to bijection.

Theorem 1.2

Suppose that $f_X(x|\theta)/f_X(y|\theta)$ is constant in θ iff $T(x) = T(y)$. Then T is minimal sufficient.

Proof. Let $x \overset{1}{\sim} y$ if $f_X(x|\theta)/f_X(y|\theta)$ is constant in θ. It's easily checked that this defines an equivalence relation. Similarly, let $x \overset{2}{\sim} y$ if $T(x) = T(y)$; this is also an equivalence relation. The hypothesis in the theorem states that the equivalence classes of $\overset{1}{\sim}$ and $\overset{2}{\sim}$ are the same.

We will construct a statistic T which is constant on the equivalence classes of $\overset{1}{\sim}$. For any value t of T let z_t be a representative from $\{x; T(x) = t\}$. Then

$$\begin{aligned} f_X(x|\theta) = f_X(z_{T(x)}|\theta) &= \frac{f_X(x|\theta)}{f_X(z_{T(x)}|\theta)} \\ &= g(T(x), \theta) h(x). \end{aligned}$$

Hence T is sufficient by the factorisation criterion. To prove T is minimal sufficient, let S be any other sufficient statistic. By the factorisation criterion, there exist functions g_S, h_S such that

$$f_X(x|\theta) = g_S(S(x), \theta) h_S(x).$$

Now suppose $S(x) = S(y)$ for some x, y. Then

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

which is constant in θ, so $x \overset{1}{\sim} y$. By the hypothesis, $x \overset{2}{\sim} y$ and $T(x) = T(y)$. $\square$

Example

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then

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

This is constant in $\left(\mu, \sigma^2\right)$ iff $\sum x_i^2 = \sum y_i^2$ and $\Sigma x_i = \sum y_i$. Hence $\left(\sum x_i^2, \sum x_i\right)$ is a minimal sufficient statistic.A more common minimal sufficient statistic is obtained by taking a bijection of $\left(\Sigma x_i^2, \Sigma x_i\right)$:

$$\begin{aligned} &S(x) = \left(\bar{x}_n, S_{xx}\right) \\ &\bar{x}_n = \frac{1}{n}\sum x_i \quad S_{xx} = \sum_i \left(x_i - \bar{x}_n\right)^2 \end{aligned}$$

In this example $\theta = \left(\mu, \sigma^2\right)$ has same dimension as $S(x)$. In general, they can be different.

1.4 Rao-Blackwell theorem

We will now look at the Rao-Blackwell theorem. This theorem allows us to start from any estimator $\widetilde{\theta}$, and then by conditioning on a sufficient statistic we get a better one.

Theorem 1.3 (Rao-Blackwell theorem)
Let T be a sufficient statistic for θ and define an estimator $\widetilde{\theta}$ with $\mathbb{E}(\widetilde{\theta}^2) < \infty$ for all θ. Define a new estimator

$$\hat{\theta} = \mathbb{E}(\widetilde{\theta}\,|T(x)).$$

Then for all $\theta \in \Theta$,

$$\mathbb{E}((\hat{\theta} - \theta)^2) \leq \mathbb{E}((\widetilde{\theta} - \theta)^2).$$

Furthermore, the inequality is strict unless $\widetilde{\theta}$ is a function of $T(x)$.

Proof. By tower property of $\mathbb{E}$, we have

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}(\mathbb{E}(\widetilde{\theta}|T)) = \mathbb{E}(\widetilde{\theta}).$$

By the conditional variance formula,

$$\begin{aligned}\mathrm{Var}(\widetilde{\theta}) &= \mathbb{E}(\mathrm{Var}(\widetilde{\theta}|T)) + \mathrm{Var}(\mathbb{E}(\widetilde{\theta}|T))\\ &= \mathbb{E}(\mathrm{Var}(\widetilde{\theta}|T)) + \mathrm{Var}(\hat{\theta})\\ &\geq \mathrm{Var}(\hat{\theta}).\end{aligned}$$

So by the bias-variance decomposition, $\operatorname{mse}\widetilde{\theta} \geq \operatorname{mse}\hat{\theta}$. The inequality is strict unless $\mathrm{Var}(\widetilde{\theta}|T) = 0$ with probability 1, which requires $\widetilde{\theta}$ is a function of T. □

Remark. T must be sufficient, since otherwise $\hat{\theta}$ would be a function of θ, so it wouldn't be an estimator.

We will now look at a few examples to show how powerful this theorem can be.

Example

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Poi}(\lambda)$. Let $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$. Then

$$\begin{aligned} f_X(x|\lambda) &= \frac{e^{-n\lambda}\lambda^{\Sigma x_i}}{\prod_i x_i!}\\ \implies f_X(x|\theta) &= \frac{\theta^n(-\log\theta)^{\sum x_i}}{\prod_{i=1} x_i!} = g\left(\sum x_i, \theta\right)h(x)\end{aligned}$$

So $\sum x_i = T(x)$ is sufficient by factorisation.

Recall $\sum X_i \sim \mathrm{Poi}(n\lambda)$. Let $\widetilde{\theta} = 1_{X_1=0}$ (which only depends on X_1, so it is a bad estimator). However, it is unbiased, which is desirable as the Rao-Blackwell process will then also yield an unbiased estimator. So let's calculate

$$\begin{aligned}\hat{\theta} = \mathbb{E}(\widetilde{\theta}|T = t) &= \mathbb{P}\left(X_1 = 0 \Big| \sum_{i=1}^n X_i = t\right)\\ &= \frac{\mathbb{P}(X_1 = 0, \sum_{i=1}^n X_i = t)}{\mathbb{P}(\sum_{i=1}^n X_i = t)}\end{aligned}$$

Example

Let $X_1, \ldots, X_n$ be iid. $U([0, \theta])$; want to estimate θ.

We previously saw that $T = \max_i X_i$ is sufficient. Let $\widetilde{\theta} = 2X_1$, an unbiased estimator of θ. Then

$$\begin{aligned}\hat{\theta} &= \mathbb{E}(\widetilde{\theta}|T = t) = 2\mathbb{E}(X_1 | \max_i X_i = t)\\ &= 2\mathbb{E}(X_1 | \max_i X_i = t, X_1 = \max_i X_i)\mathbb{P}(\max_i X_i = X_1 | \max_i X_i = t)\\ &\quad + 2\mathbb{E}(X_1 | \max_i X_i = t, X_1 \neq \max_i X_i)\mathbb{P}(\max_i X_i \neq X_1 | \max_i X_i = t)\end{aligned}$$

$$= \frac{2t}{n} + \frac{2(n-1)}{n} \underbrace{\mathbb{E}(X_1 | X_1 < t, \max_{2\le i\le n} X_i = t)}_{=t/2 \text{ as } X_1|X_1<t\sim U([0,t])}$$

$$= \frac{n+1}{n} \max_i X_i.$$

By Rao-Blackwell $\operatorname{mse}(\hat\theta) \le \operatorname{mse}(\widetilde{\theta})$. Also, $\hat\theta$ is unbiased.

1.5 Maximum likelihood estimation

Definition (Likelihood function/Maximum likelihood estimator)
Let $X_1, \dots, X_n$ be iid. with pdf (or pmf) $f_X(\cdot|\theta)$. The **likelihood function** $L : \theta \to \mathbb{R}$ is given by

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

(We take X to be fixed observations.) We further define the **log-likelihood**

$$l(\theta) = \log L(\theta) = \sum_{i=1}^n \log f_{X_i}(x_i|\theta).$$

A **maximum likelihood estimator** (mle) is an estimator that maximises L over Θ.

Example
Let $X_1, \dots X_n \sim^{iid} \text{Ber}(p)$. Then we have

$$\begin{aligned} l(p) &= \sum_{i=1}^n X_i \log p + (1-X_i)\log(1-p) \\ &= \log p(\sum X_i) + \log(1-p)(n - \sum X_i) \end{aligned}$$

$$\implies \frac{\mathrm{d}l}{\mathrm{d}p} = \frac{\sum X_i}{p} + \frac{n - \sum X_i}{1-p}$$

This is > 0 iff $p = \frac{1}{n}\sum X_i = \overline{X_i}$. We have $\mathbb{E}(\overline{X_i}) = \frac{n}{n}\mathbb{E}(X_1) = p$. So the mle $\hat p = \overline{X_i}$ is unbiased.

Now let's try a more involved example.

Example
Let $X_1, \dots X_n \sim^{iid} N(\mu, \sigma^2)$. Then we have

$$l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i - \mu)^2.$$

This is maximised when $\frac{\partial l}{\partial \mu} = \frac{\partial l}{\partial \sigma^2} = 0$. But $\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)$ so is equal to 0 iff

$$\mu = \hat{X_n} = \frac{1}{n}\sum X_i.$$

for all $\sigma^2 > 0$. We also have that $\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \mu)^2$. If we set $\mu = \overline{X_n}$, $\frac{\partial l}{\partial \sigma^2} = 0$ iff

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{X_n})^2 = \frac{S_{xx}}{n}.$$

Hence the mle is $(\hat{\mu}, \hat{\sigma^2}) = (\overline{X_n}, \frac{S_{xx}}{n})$. We can check $\hat{\mu}$ is unbiased. Later in the course we will see that

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

Therefore $\mathbb{E}(\sigma^2) = \frac{\sigma^2}{n}\mathbb{E}(\chi^2_{n-1}) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. Hence $\hat{\sigma}^2$ is biased. But as $n \to \infty$ the bias converges to 0, so we say $\hat{\sigma}^2$ is **asymptotically unbiased**.

The next example will focus on an example where the mle is discontinuous, and doesn't behave as nicely.

Example

Let $X_1, \ldots, X_n$ be iid. $U([0, \theta])$. Recall the estimator we derived, $\hat{\theta} = \frac{n+1}{n}\max_i X_i$. The likelihood function is

$$L(\theta) = \frac{1}{\theta^n} 1_{\max_i X_i \leq \theta}.$$

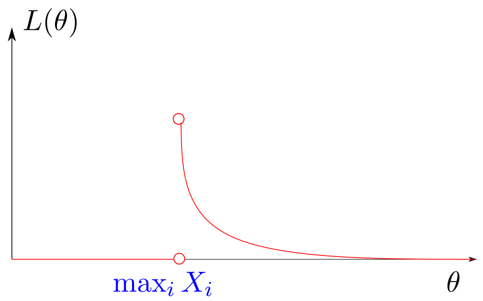

Figure 1: The plot of $L(\theta)$; note the discontinuity.

Hence the mle is $\hat{\theta}^{\text{mle}} = \max_i X_i$. As $\hat{\theta}$ is unbiased, $\hat{\theta}^{\text{mle}}$ is **not** unbiased.

Properties of the mean likelihood estimator

1. If T is sufficient for θ, then the mle is a function of T. Recall

$$L(\theta) = g(T, \theta)h(X).$$

   So the maximiser of L only depends on X through T.

2. If we parameterise θ in some way, say $\phi = H(\theta)$ where H is a bijection, and $\hat{\theta}$ is the mle for θ, then $H(\hat{\theta})$ is the mle for ϕ.

3. Asymptotic normality: Under regularity conditions, as $n \to \infty$ the statistic $\sqrt{n}(\hat{\theta} -$

$\theta)$ is approx $N(0, \Sigma)$, i.e for some 'nice' set A we have

$$\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \in A) \xrightarrow{n\to\infty} \mathbb{P}(Z \in A), \qquad \text{where } Z \sim N(0, \Sigma).$$

The limiting covariance matrix Σ is a known function of L. In some sense it is the 'best' or 'smallest' variance that any estimator can achieve asymptotically. See Part II Principles of Statistics for more on this.

4. When the mle is not available analytically in closed form, in real-world applications it is often found numerically (see Part IB Numerical Analysis).

1.6 Confidence intervals

The idea of confidence intervals is omnipresent; it is used in the real world to describe a measure of certainty, and you may well have used the term in conversation or seen it in media before. We will give a rigorous mathematical definition of confidence.

Definition (Confidence interval)
A $100 \cdot \gamma\%$ **confidence interval** with $\gamma \in (0, 1)$ and for a parameter θ is a random interval $(A(x), B(x))$ such that

$$\mathbb{P}(A(x) \leq \theta \leq B(x)) = \gamma \qquad \text{for all } \theta \in \Theta.$$

Note that we consider θ to be a fixed parameter, but the endpoints of the interval are randomly changing.

Remark. When θ is a vector, we talk about confidence sets instead of confidence intervals.

A **frequentist interpretation** is that if we repeat the experiment many times, on average $100 \cdot \gamma\%$ of the time the interval will contain θ.

A **misleading interpretation** is: "having oberved $X = x$, there is now a probability γ that $\theta \in [A(x), B(x)]$". This is actually **incorrect**, and we will later see an example that shows this.

Example
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. We want to find the 95% confidence interval for θ. We know $\overline{X} \sim N(\theta, \frac{1}{n})$ and $Z = \sqrt{n}(\overline{X} - \theta) \sim N(0, 1)$ for all $\theta \in \mathbb{R}$.

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

$$\mathbb{P}(a \leq \sqrt{n}(\overline{X} - \theta) \leq b) = 0.95.$$

Rearrange:

$$\mathbb{P}(\overline{X} - \frac{b}{\sqrt{n}} \leq \theta \leq \overline{X} - \frac{a}{\sqrt{n}}) = 0.95$$

Hence $(\overline{X} - \frac{b}{\sqrt{n}}, \overline{X} - \frac{a}{\sqrt{n}})$ is a 95% C.I for θ.

Note a, b are not unique. Typically we centre the interval around some estimator $\hat{\theta}$ and aim to minimise its length. In this case, we would choose $a = -b$, which would give $b = Z_{0.025} \approx 1.96$ where Z_α is equal to $\Phi^{-1}(1-\alpha)$. We call this the "upper α-point" of the $N(0,1)$ distribution.

Therefore our final C.I is $(\overline{X} \pm \frac{1.96}{\sqrt{n}})$. A quick sanity check is to note that our interval decreases as n gets larger (with more observations).

We can generalise the method we used in this example.

Remark. Recipe for finding a confidence interval:

1. Find a quantity $R(X, \theta)$ whose $\mathbb{P}_\theta$-distribution *doesn't* depend on θ. This is called a **pivot**, for example in the above example our pivot was $R(X, \theta) = \sqrt{n}(\overline{X} - \theta)$.
2. Write down a statement
$$\mathbb{P}(c_1 \leq R(X, \theta) \leq c_2) = \gamma.$$
Given some γ, we find c_1, c_2 using the distribution of R.
3. Rearrange to leave θ in the middle of two inequalities.

Proposition 1.4
If T is a monotone increasing function and $(A(X), B(X))$ is a $100 \cdot \gamma\%$ C.I for θ, then $T(A(X), T(B(X)))$ is a $100 \cdot \gamma\%$ C.I for $T(\theta)$.

Proof. Immediate from definitions. (Exercise) □

Example
Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We want to find a 95% C.I for σ^2. Let's follow our recipe:

1. Note that $\frac{X_1}{\sigma} \sim N(0,1)$. This is a pivot, but ideally we would want one that depends on all the observations. So let our pivot be
$$\sum_{i=1}^{n} \frac{X_i}{\sigma^2} \sim \chi_n^2.$$
2. Let $c_1 = F_{\chi_n^2}{}^{-1}(0.025)$ and $c_2 = F_{\chi_n^2}{}^{-1}(0.975)$.
3. Now rearrange to get σ^2 in the middle:
$$\mathbb{P}(\frac{\sum X_i^2}{c_2} \leq \sigma^2 \leq \frac{\sum X_i^2}{c_1}) = 0.95$$
Hence this is our 95% confidence interval for σ^2.

Example
Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Ber}(p)$ with n "large". We want to find an approximate 95% C.I for p.

1. The mle of p is $\hat{p} = \overline{X} = \frac{1}{n} \sum X_i$. By the central limit theorem, $\hat{p}$ is approx

$N(p, p(1-p)/n)$. Therefore

$$\sqrt{n}\frac{(\hat{p}-p)}{\sqrt{p(1-p)}} \quad \text{is approx } N(0,1).$$

2. $\mathbb{P}(-Z_{0.025} \leq \sqrt{n}\frac{(\hat{p}-p)}{\sqrt{p(1-p)}} \leq Z_{0.025}) \approx 0.95$.

3. Note that if we wanted to rearrange for p here, we would have to solve a quadratic inequality. So instead of this, we'll approximate $\sqrt{p(1-p)} \approx \sqrt{\hat{p}(1-\hat{p})}$. We argue when n is large

$$\mathbb{P}(-Z_{0.025} \leq \sqrt{n}\frac{(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \leq Z_{0.025}) \approx 0.95.$$

This is easier to rearrange, which gives

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

So we have found an approximate 95% confidence interval for p.

Remark. In the above example, $p(1-p) \leq \frac{1}{4}$ on $p \in (0,1)$ hence $(\hat{p} \pm \frac{Z_{0.025}}{2\sqrt{n}})$ is a "conservative" 95% C.I for p.

Let's go back to the issue of how to interpret a confidence interval, and the two interpretations that were mentioned. This can be seen in the following example:

Example

Suppose X_1, X_2 are iid. $\mathrm{Unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. What is a sensible 50% confidence interval for θ? Note

$$\begin{aligned}
\mathbb{P}(\theta \text{ between } X_1, X_2) &= \mathbb{P}(\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)) \\
&= \mathbb{P}(X_1 \leq \theta \leq X_2) + \mathbb{P}(X_2 \leq \theta \leq X_1) \\
&= \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{2}.
\end{aligned}$$

Hence the frequentist interpretation is *exactly* correct.

But suppose $|X_1 - X_2| > 0.5$. Then we *know* that θ is in $(\min(X_1, X_2), \max(X_1, X_2))$

1.7 Bayesian analysis

So far we've talked about *frequentist* inference; we think of θ as being fixed. Inferential statements are interpreted in terms of repetitions of the experiment.

Bayesian analysis is a different framework. In this view, we treat θ as a random variable taking values in Θ. The **prior distribution** $\pi(\theta)$ represents the investigator's beliefs or information about θ before observing the data. Conditional on θ, the data X has pdf (or pmf) $f_X(\cdot|\theta)$. (This is why we've been writing f_X in this way when we use the frequentest interpretation.) Having observed X, the information in X is combined with the prior to form the **posterior distribution** $\pi(\theta|X)$, which is the conditional distribution of θ given X. By Bayes' Rule:

$$\pi(\theta|X) = \frac{\pi(\theta) f_X(X|\theta)}{f_X(X)}.$$

where $f_X(x)$ is the marginal distribution of X:

$$f_X(X) = \begin{cases} \int_\Theta f_X(X|\theta)\pi(\theta)\ \mathrm{d}\theta & \text{if } \theta \text{ continuous;} \\ \sum_{\theta\in\Theta} f_X(X|\theta)\pi(\theta) & \text{if } \theta \text{ discrete.} \end{cases}$$

More simply,

$$\pi(\theta|X) \propto \pi(\theta) \cdot f_X(X|\theta).$$

Often, it is easy to recognize that the RHS is in some family of distributions up to a normalising constant.

Remark. By the factorisation criterion, if T is sufficient then

$$\pi(\theta|X) \propto \pi(\theta) \cdot g(T(X), \theta).$$

Therefore the posterior only depends on X through $T(X)$, since we can absorb the $h(x)$ in the decomposition of T into the constant.

Example
Suppose a patient walks into a COVID-19 testing clinic (we have no prior info about the patient). Our random variable is $\theta = 1_{\text{patient infected}}$. We know the sensitivity of the test $f_X(X=1|\theta=1)$ and the specificity $f_X(X=0|\theta=0)$.

How do we choose the prior? Let $\pi(\theta=1)$ to be the proportion of people in the UK with COVID on that day. What is the probability of an infection given a positive test?

$$\pi(\theta=1|X=1) = \frac{\pi(\theta=1) f_X(X=1|\theta=1)}{\pi(\theta=1) f_X(X=1|\theta=1) + \pi(\theta=0) f_X(X=1|\theta=0)}.$$

Sometimes $\pi(\theta=1) \ll \pi(\theta=0)$ which can make $\pi(\theta=1|X=1)$ small, which can be surprising.

In the second example, choosing the prior will be less clear-cut.

Example
Let $\theta \in [0,1]$ be the mortality rate for a new surgery in Addenbrooke's hospital. We observe the first 10 operations and see no deaths. We model $X \sim \mathrm{Bin}(10, \theta)$.

How do we choose the prior? Suppose that in other hospitals mortality ranges between 3% and 20% with an average of 10%. For example, take $\pi(\theta) \sim \beta(a,b)$. We can choose $a=3$ and $b=27$ so that $\pi(\theta)$ has mean 0.1 and $\pi(0.03 < \theta < 0.2) \approx 0.9$. The posterior is

$$\begin{aligned} \pi(\theta|X) &\propto \pi(\theta) \cdot f_X(X=10|\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} \\ &\propto \theta^{x+a-1}(1-\theta)^{b+n-x-1} \qquad \text{for } \theta \in [0,1]. \end{aligned}$$

We recognise this as a $\beta(X+a, n+X-b)$ distribution. In our example this is $\beta(3, 10+27)$. Plotting our prior and posterior, we can see that the posterior has a distribution shifted slightly to the left (since we observed no deaths in 10 trials), and that its variance is smaller (since we have already seen some trials). [picture]

Remark. In this example the prior and posterior are in the same family of distributions. This is called **conjugacy**.

What do we do with the posterior?

$\pi(\theta|X)$ represents information about θ *after* seeing X. This can be used to make decisions under uncertainty.

1. We must pick some decision $\delta \in D$. For example, $D = \{$ ask patient to isolate (or not) $\}$.
2. Define the **loss function** $L(\theta, \delta)$. For example, $L(\theta = 1, \delta = 1)$ would be the loss incurred by asking a patient to isolate if they test positive.
3. Pick δ that minimises
$$\int_\Theta L(\theta, \delta)\,\pi(\theta|X)\ \mathrm{d}\theta.$$
This is called the "**posterior expectation of loss**".[2]

Point estimation

An example of a decision is a "best guess" for θ. The **Bayes estimator** minimises

$$h(\delta) = \int_\Theta L(\theta, \delta)\,\pi(\theta|X)\ \mathrm{d}\theta.$$

Example
If we choose a quadratic loss $L(\theta, \delta) = (\theta - \delta)^2$. Then $h(\delta) = \int_\Theta L(\theta - \delta)^2\,\pi(\theta|X)\ \mathrm{d}\theta$. Differentiate: $h'(\delta) = 0$ if

$$\begin{aligned}\int_\Theta (\theta - \delta)\,\pi(\theta|X)\ \mathrm{d}\theta = 0 \\ \iff \delta = \int_\Theta \theta\pi(\theta|X)\ \mathrm{d}\theta.\end{aligned}$$

This is $\hat{\theta}$ under quadratic loss.

Example
If we use the absolute error loss $L(\theta, \delta) = |\theta - \delta|$, then

$$\begin{aligned}h(\delta) &= \int_\Theta |\theta - \delta|\,\pi(\theta|X)\ \mathrm{d}\theta \\ &= \int_{-\infty}^{\delta} -(\theta - \delta)\,\pi(\theta|X)\ \mathrm{d}\theta + \int_\delta^\infty (\theta - \delta)\,\pi(\theta|X)\ \mathrm{d}\theta.\end{aligned}$$

[2] See Von-Neumann/Morgenstein.

Take the derivative wrt. θ using the FTC:

$$h'(\theta) = \int_{-\infty}^{\delta} \pi(\theta|X)\ \mathrm{d}\theta - \int_{\delta}^{\infty} \pi(\theta|X)\ \mathrm{d}\theta.$$

So $h'(\delta) = 0$ iff $\delta = \hat{\theta}$ is the median of the posterior $\pi(\theta|X)$.

We would also want some kind of Bayesian interpretation of a confidence interval. The next definition makes this more concrete.

Definition (Credible interval)
A $100 \cdot \gamma\%$ credible interval satisfies

$$\pi(A(x) \leq \theta \leq B(x)|x) = \gamma.$$

Remark. Unlike confidence intervals, credible intervals **can** be interpreted conditionally; for example, "given a specific observation x, we are 95% certain that θ is in (A, B)". The caveat here is that the credible interval depends on the choice of prior.

We'll quickly look at some examples of Bayesian computations in order to get a better feel for them. [todo]

Remark. Asymptotically, we will often see credible intervals approach confidence intervals (as in the previous example).

[next example] This is where our discussion of Bayesian analysis ends - we now return to a frequentist viewpoint.

2 Hypothesis testing

2.1 Simple hypotheses

Definition (Hypothesis)
A **hypothesis** is an assumption about the distribution of data X.

Scientific questions are often phrased as a decision between a **null hypothesis** H_0 and an **alternative hypothesis** H_1.

Example
1. $X = (X_1, \dots, X_n)$ are iid. $\text{Ber}(\theta)$. Suppose $H_0: \theta = 1/2$, $H_1: \theta = 3/4$.
2. In 1. we could also have $X_1: \theta \neq 1/2$.
3. Let $X = (X_1, \dots, X_n)$ be iid. where X_i takes values in $\mathbb{N}$. Suppose

$$H_0: X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda) \text{ for some } \lambda > 0, \quad H_1: X_i \overset{\text{iid}}{\sim} f_1 \text{ for some other dist. } f_1.$$

This is known as a **goodness-of-fit test**.

Definition (Simple hypothesis)
A **simple hypothesis** is one which fully specifies the pdf (resp. pmf) of X. Otherwise, we say the hypothesis is **composite**.

In the last example, the test in 1. was composite, and the test in 2. was simple.

Definition (Test)
A **test** of the null H_0 is defined by a **critical region** $C \subseteq \mathcal{X}$. When $X \in C$, we *reject the null*. When $X \notin C$, we *fail to reject the null*.

Remark. When $X \notin C$, we simply don't find evidence to reject the null; it doesn't mean the null is false. We will see examples of this later.

Definition (Error)
There are two types of error:

- **Type I error**: rejecting H_0 when H_0 is true.
- **Type II error**: fail to reject H_0 when H_0 isn't true.

Write

$$\begin{aligned}
\alpha &= \mathbb{P}_{H_0}(H_0 \text{ is rejected}) = \mathbb{P}_{H_0}(X \in C), \\
\beta &= \mathbb{P}_{H_1}(H_0 \text{ is not rejected}) = \mathbb{P}_{H_1}(X \in C).
\end{aligned}$$

The **size** of the test is α, the *power* is $1-\beta$. There is a tradeoff between α and β. What we typically do is choose an acceptable probability of type I errors (say 1%); set α to that, and then pick the test which minimises β (maximises power).

Definition (Likelihood ratio statistic/test)
Let H_0 and H_1 be simple, with X having pdf (or pmf) f_i under H_i for $i \in \{0, 1\}$. The **likelihood ratio statistic** is

$$\Lambda_x(H_0, H_1) = \frac{f_1(x)}{f_0(x)}.$$

A **likelihood ratio test** (LRT) rejects when $\Lambda_x(H_0, H_1)$ is large, i.e

$$C = \{x : \Lambda_x(H_0, H_1) > k\} \text{ for some } k.$$

Lemma 2.1 (Neyman-Pearson lemma)
Suppose that f_0, f_1 are nonzero on some sets. Suppose there is $k > 0$ such that the LRT with critical region $C = \{x : \Lambda_x(H_0, H_1) > k\}$ has size α. Then out of all tests with size $\leq \alpha$, this test has the smallest β.

Remark. An LRT of size α does not always exist. (Exercise.) But in general, we can find a "randomised" test of size α.

Proof. Let $\overline{C}$ be the complement of C. We know that the LRT has

$$\alpha = \mathbb{P}_{H_0}(X \in C) = \int_C f_0(x)\ \mathrm{d}x, \quad \beta = \mathbb{P}_{H_1}(X \notin C) = \int_{\overline{C}} f_1(x)\ \mathrm{d}x.$$

Let C^* be some other critical region with type I/II error probabilities α^*, β^*. Suppose $\alpha^* \leq \alpha, \beta \leq \beta^*$. We want to show $\beta \leq \beta^*$. Note that

$$\begin{aligned}
\beta - \beta^* &= \int_{\overline{C}} f_1(x)\ \mathrm{d}x - \int_{\overline{C^*}} f_1(x)\ \mathrm{d}x \\
&= \int_{\overline{C} \cap C^*} f_1(x)\ \mathrm{d}x - \int_{\overline{C^*} \cap C} f_1(x)\ \mathrm{d}x \text{ as the integrals over } \overline{C} \cap \overline{C^*} \text{ cancel} \\
&= \int_{\overline{C} \cap C^*} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\leq k \text{ on } \overline{C}} f_0(x)\ \mathrm{d}x - \int_{\overline{C^*} \cap C} \underbrace{\frac{f_1(x)}{f_0(x)}}_{> k \text{ on } C} f_0(x)\ \mathrm{d}x \\
&\leq k \left(\int_{\overline{C} \cap C^*} f_0(x)\ \mathrm{d}x - \int_{\overline{C^*} \cap C} f_0(x)\ \mathrm{d}x \right) \\
&\text{Now add and subtract } k \int_{C \cap C^*} f_0(x)\ \mathrm{d}x : \\
&= k \left(\int_{C^*} f_0(x)\ \mathrm{d}x - \int_C f_0(x)\ \mathrm{d}x \right) \\
&= k(\alpha^* - \alpha) \leq 0.
\end{aligned}$$

$\square$

Let's illustrate this with an example.

Example
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_i^2)$ where σ_0 is known. We want the best size α test for

H_0: $\mu = \mu_0$ vs H_1: $\mu = \mu_1$ for some fixed $\mu_1 > \mu_0$. Skipping the algebra, we get

$$\Lambda_X(H_0; H_1) = \exp\left\{\frac{\mu_1 - \mu_0}{\sigma_0^2} n\overline{X} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma_0^2}\right\}.$$

Λ_x is monotone increasing in $\overline{X}$; it is also monotone increasing in $Z = \sqrt{n}(\overline{X} - \mu_0)/\sigma_0$. Thus $\Lambda_x > k \iff z > k'$ for some $k' > 0$. Hence the LRT has critical region of the form

$$C = \left\{x : Z(x) > k'\right\}.$$

To find the most powerful test, by the Neyman-Pearson lemma, we need only find k such that C has size α. Under $H_0 : \ \mu = \mu_0$, $Z \sim N(0,1)$. Thus if we choose $k' = \Phi^{-1}(1-\alpha)$ we have

$$\mathbb{P}_{H_0}(z > k') = \alpha,$$

i.e the test $C = \{x : Z(x) > k'\}$ has size α. This is called a **z-test**.

The p-value

If we have a critical region $\{x : T(x) > k\}$ for some **test statistic** $T(X)$, we usually report a **p-value** in addition to the test's conclusion, which is defined by

$$p = \mathbb{P}_{H_0}(T(X) > T(x^*)),$$

where x^* is the observed data.

In the example above, suppose $\mu_0 = 5, \mu_1 = 6, \alpha = 0.05$. Suppose we are given the data $x^* = (5.1, 5.5, 4.9, 5.3)$. We have $\overline{x^*} - 5.2, z^* = 0.4$. The LRT is

$$\left\{x : Z(x) > \Phi^{-1}(0.95) = 1.645\right\}.$$

. The conclusion of the LRT is that we do not reject H_0. [drawing todo] Here $p = 1 - \Phi(z^*) = 0.35$.

Proposition 2.2
Under H_0, the p-value is $\sim U[0,1]$.

Proof. Let F be the distribution of T. Assume that F is continuous. Then

$$\begin{aligned}
\mathbb{P}_{H_0}(p < u) &= \mathbb{P}_{H_0}(1 - F(T) < u) \\
&= \mathbb{P}_{H_0}(F(T) > 1 - u) \\
&= \mathbb{P}_{H_0}(T > F^{-1}(1-u)) \\
&= 1 - F(F^{-1}(1-u)) = u.
\end{aligned}$$

□

2.2 Composite hypotheses

Let $X \sim f_X(\cdot|\theta)$, where $\theta \in \Theta$. We define a **composite hypothesis**

$$H_0 : \theta \in \Theta_0 \subset \Theta$$

$$H_1 : \theta \in \Theta_1 \subset \Theta$$

Now the probabilities of Type I or II error may depend on the value of θ within Θ_0 or Θ_1. They are not constants.

Definition (Power function/size)
The **power function** for a test C is

$$W(\theta) = \mathbb{P}_\theta(X \in C).$$

The **size** of a test C is

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta).$$

We say that a test is **uniformly most powerful** (UMP) if for any other test C^* with power function W^* and size α,

$$W(\theta) \geq W^*(\theta) \text{ for all } \theta \in \Theta_1.$$

Remark. UMP tests need not exist. However, in simple models many LRTs are UMP.

Example
One-sided test for normal location: let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ with σ_0^2 known. We define our composite hypotheses:

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0.$$

The LRT for the simple hypotheses

$$H_0' : \mu = \mu_0, H_1' : \mu = \mu_1 > \mu_0$$

is

$$C = \left\{ x : \ z = \sqrt{n}\frac{(\overline{x} - \mu_0)}{\sigma_0} > \Phi^{-1}(1 - \alpha) \right\}.$$

Claim. This test is UMP for H_0 against H_1.

Proof. Last time we found $W(\mu) = 1 - \Phi(z_0 + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0})$. Indeed we have $\sup_{\mu \leq \mu_0} W(\mu) = \alpha$. Now we need to check that for any test C^* of size α with power function W^*, $W(\mu) \geq W^*(\mu)$ for all $\mu > \mu_0$.

First note that the critical region C depends on μ_0, not on μ_1. Take any $\mu_1 > \mu_0$, then C is the LRT for H_0' vs H_1'. We can alse see C^* as a test of H_0' vs H_1'. And for these simple hypotheses C^* has size

$$W^*(\mu_0) \leq \sup_{\mu < \mu_0} W^*(\mu) \leq \alpha.$$

By the Neyman-Pearson Lemma (2.1), C has power no smaller than C^* for H_0' vs H_1', i.e

$$W(\mu_1) \geq W^*(\mu_2).$$

Hence C is a UMP. $\square$

Generalised likelihood ratio test

Let $H_0: \ \theta \in \Theta_0$, $H_1: \ \theta \in \Theta_1$ with $\Theta_0 \subseteq \Theta_1$. We say the hypotheses are nested. The **generalised likelihood ratio test** (GRT) is

$$\Lambda_X(H_0; H_1) = \frac{\sup_{\theta \in \Theta_1} f_X(X|\theta)}{\sup_{\theta \in \Theta_0} f_X(X|\theta)}.$$

Large values indicate a better fit under the alternative hypothesis. A GLR test rejects H_0 when $\Lambda_X(H_0; H_1)$ is large.

Example

Again take our two-sided test for normal location. Let our nested hypotheses be

$$H_0: \ \mu = \mu_0, \quad H_1: \ \mu \in \mathbb{R}.$$

In this model we have

$$\Lambda_X(H_0; H_1) = \frac{(2\pi\sigma_0^2)^{-n/2} \exp(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \overline{X})^2)}{(2\pi\sigma_0^2)^{-n/2} \exp(-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (X_i - \mu_0)^2)}.$$

To make it easier, we consider

$$2 \log \Lambda_x = \frac{n}{\sigma_0^2}(\overline{X} - \mu_0^2).$$

Recall that under H_0, $\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma_0} \sim N(0,1)$. So $2 \log \Lambda_x \sim \chi_1^2$. So the critical region of GLR test is

$$C = \left\{ x : \frac{n}{\sigma_0^2}(\overline{x} - \mu_0^2) > \underbrace{\chi_1^2(\alpha)}_{\text{upper } \alpha\text{-point of } \chi_1^2} \right\}.$$

Definition (Dimension of a hypothesis)
The **dimension** of a hypothesis $H_0 : \theta \in \Theta_0$ is the number of "free parameters" in Θ_0.

Example

- If $\Theta_0 = \{\theta \in \mathbb{R}^k : \ \theta_1 = \ldots = \theta_p = 0\}$, then $\dim \Theta_0 = k - p$.
- If $A \in \mathbb{R}^{p\times k}$ has linearly independent rows, and $b \in \mathbb{R}^p$ where $p < k$, and $\Theta_0 = \{\theta \in R^k : \ A\theta = b\}$, then $\dim \Theta_0 = k - p$.
- We could generalise this to Θ_0 being a differential manifold, which is overkill for this course, and we would need a notion of differential geometry which we have not yet encountered.

Theorem 2.3 (Wilks' theorem)
Suppose $\Theta_0 \subset \Theta_1$ and $\dim \Theta_1 - \dim \Theta_0 = p$. Then if $X = (X_1, \ldots, X_n)$ are iid. under $f_X(\cdot|\theta)$ with $\theta \in \inf(\Theta_0)$, then [under some topological conditions] the limiting distribution of $2 \log \Lambda_x$ is χ^2_p.

Remark. This is *very* useful because it allows us to implement a generalised ratio test even if we can't find the exact distribution of $2 \log \Lambda_x$ (assuming that n is large; any frequentist guarantee will be approximate).

Proof. Omitted; this is proved in Part II Principles of Statistics. □

Example
In the two-sided normal location example, $\dim \Theta_0 = 0$ and $\dim \Theta_1 = 1$. So Wilks' theorem tells us $2 \log \Lambda_X$ is exactly 1 (in this example, this happens to be exact).

Goodness-of-fit test

Let $X_1, \ldots, X_n$ be iid. samples taking values in $\{1, \ldots, k\}$. Let $p_i = \mathbb{P}(X_1 = i)$, and let N_i be the number of samples equal to i. Hence $\sum_i N_i = n$, and $\sum_i p_i = 1$.

We can view this as a model with parameters $p := (p_1, \ldots, p_k)$. The parameter space has dimension $k - 1$ (since we have one equality constraint). A **goodness-of-fit** (GoF) test has a null of the form

$$H_0 : p_i = \widetilde{p}_i, \quad i = 1, \ldots, k$$

for some fixed distribution $\widetilde{p}$. The alternative puts no constraints on p.

The model is $(N_1, \ldots, N_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$. So $L(p) \propto p_1^{N_1} \ldots p_k^{N_k}$. Hence

$$l(p) = \log L(p) = c + \sum_i N_i \log p_i$$

for some constant c. The GLR Λ_x has

$$2 \log \Lambda_x = 2(\sup_{p\in\Theta_1} l(p) - \underbrace{\sup_{p\in\Theta_0} l(p)}_{=l(\widetilde{p})}).$$