Statistics

Cambridge University Mathematical Tripos: Part IB

4th May 2024

Contents

1 Introdution 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 $X_1 \sim \text{Poi}(\mu)$, 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:

(i) we can estimate the value of θ using a *point estimate* written $\hat{\theta}(X)$;

(ii) we can make an *interval estimate* of θ, written $(\hat{\theta}_1(X), \hat{\theta}_2(X))$;

(iii) hypotheses about θ can be tested, for instance the hypothesis $H_0 : \theta = 1$, by checking whether there is evidence in the data X against the hypothesis H_0.

Remark. 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* $\mathbb{P} : \mathcal{F} \to [0, 1]$ satisfies the following properties.

(i) $\mathbb{P}(\varnothing) = 0$;

(ii) $\mathbb{P}(\mathcal{F}) = 1$;

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

A *random variable* is a *measurable function* $X : \Omega \to \mathbb{R}$. 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 $g : \mathbb{R} \to \mathbb{R}$, 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 $X_1, \dots, X_n$ are *independent* if, for all $x_1, \dots, x_n$, we have

$$\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n)$$

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

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

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

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

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 $a = (a_1, \dots, a_n)^\intercal \in \mathbb{R}^n$ 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

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

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 $X_1, \ldots, X_n$ 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 $x \to \infty$. Note that the event $\left|\overline{X}_n - \mu\right| > \varepsilon$ depends only on $X_1, \ldots, X_n$.

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, $\mathbb{P}(Z_n \le z) \to \Phi(z)$ 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

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

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

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

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

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

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

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

The *law of total variance* is that

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

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 $n; p_1, \ldots, p_k$ 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 $S_n \sim \Gamma\left(\sum_i \alpha_i, \lambda\right)$, 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:

- $\Gamma(1, \lambda) = \text{Exp}(\lambda)$;
- $\Gamma(k/2, 1/2) = \chi^2_k$ 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. 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. 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. The *bias* of $\hat{\theta}$ is

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

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

Definition. An estimator with zero bias for all θ is called an *unbiased estimator*.

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

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

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

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

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

Remark. 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

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

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

Example. Let $X \sim \text{Bin}(n, \theta)$ 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 $\mathbb{E}_\theta[T_U] = \mathbb{E}_\theta[X]/n = \theta$. The mean squared error for the estimator is then

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

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,

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

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

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

We can calculate

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

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. Suppose $X \sim \text{Poi}(\lambda)$ 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. 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. 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,

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

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. 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 $T(x) = t$,

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

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

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

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

□

Example. 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. 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. 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. 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. 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. 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 $\{x : T(x) = t\}$. 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. □

Example. 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 $\left(\sum_i x_i^2, \sum_i x_i\right)$ 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.

Example. θ and a minimal statistic T need not have the same dimension. Consider $X_1, \dots, X_n \sim N(\mu, \mu^2)$. 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. Let T be a sufficient statistic for θ, and define an estimator $\tilde{\theta}$ with $\mathbb{E}\left[\tilde{\theta}^2\right] < \infty$ for all θ. Now we define another estimator

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

Then, for all values of θ, we have

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

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

Remark. Starting from any estimator $\tilde{\theta}$, 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:

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

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, subtracting $\tilde{\theta}$ from both sides, we find $\operatorname{bias}(\hat{\theta}) = \operatorname{bias}(\tilde{\theta})$. By the conditional variance formula,

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

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

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

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

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

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

This may also be written

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

which is an estimator with lower mean squared error than $\tilde{1}$ for all θ. 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, $\overline{X}_n \to \mathbb{E}[X_1] = \lambda$, so we arrive at $\hat{\theta} \to e^{-\lambda} = \theta$ almost surely.

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

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

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.

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

By the Rao–Blackwell theorem, the mean squared error of $\hat{\theta}$ is not greater than the mean squared error of $\tilde{\theta}$. This is also an unbiased estimator.

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

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

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

Example. 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 $p = \frac{1}{n}\sum X_i = \overline{X}_n$. We have $\mathbb{E}[\hat{p}] = p$, so the maximum likelihood estimator in this case is unbiased.

Example. 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 $\mu = \overline{X}_n$. Further,

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

This is zero if and only if

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

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 $\hat{\mu}$ 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

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

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

Example. 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

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

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. 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. If T is a sufficient statistic for θ, then the maximum likelihood estimator is a function of T. Indeed, since X and T are fixed, the maximiser of $L(\theta) = g(T, \theta)h(X)$ depends on X only through T. 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.

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, we have

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

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. In some sense, Σ is the smallest variance that any estimator can achieve asymptotically.

For practical purposes, this estimator can often be found numerically by maximising ℓ or L.

3 Inference

3.1 Confidence intervals

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

Remark. 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 is observed, there is a probability γ that $\theta \in (A(x), B(x))$. This is wrong, as will be demonstrated later.

Example. Let $X_1, \ldots, X_n$ be i.i.d. normal random variables with unit variance. We will find the 95% confidence interval for $\mu = \theta$. We have

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

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

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

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. In general, to find a confidence interval:

(i) 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, $R(X, \theta) = \sqrt{n}\left(\overline{X} - \theta\right)$.

(ii) Consider $\mathbb{P}(c_1 \leq R(X, \theta) \leq c_2) = \gamma$. Given some desired level of confidence γ, find c_1 and c_2 using the distribution function of the pivot.

(iii) Rearrange such that $\mathbb{P}(A(X) \leq \theta \leq B(X)) = \gamma$, then $(A(X), B(X))$ is the confidence interval as required.

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

Example. Let $X_1, \dots, X_n$ 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. 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. 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. 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 $\theta \in [0.1, 0.9]$.

4 Bayesian analysis

4.1 Introduction

Frequentist analysis considers the value θ to be fixed, and then we can make inferential statements about θ in the context of repeated experiments on a random variable X. Bayesian analysis is an alternative to frequentist analysis, where θ is itself treated as a random variable taking values in the parameter space Θ. We say that the *prior* distribution $\pi(\theta)$ is a distribution representing the beliefs of the investigator about θ before observing data. The data X has a p.d.f. or p.m.f. conditional on θ given by $f_X(\,\cdot \mid \theta)$. Having observed X, we can combine this information with the prior distribution to form the *posterior* distribution $\pi(\theta \mid X)$, which is the conditional distribution of θ given X. This contains updated information about the value of θ. By Bayes' rule,

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

where $f_X(x)$ is the marginal distribution of X, defined by

$$f_X(x) = \begin{cases} \int_\Theta f_X(x \mid \theta)\pi(\theta)\,\mathrm{d}\theta & \theta \text{ continuous} \\ \sum_\Theta f_X(x \mid \theta)\pi(\theta) & \theta \text{ discrete} \end{cases}$$

More simply,

$$\pi(\theta \mid X) \propto \pi(\theta) \cdot f_X(X \mid \theta)$$

The proportionality here is with respect to θ. So the posterior is proportional to the prior multiplied by the likelihood. It is often easy to recognise that the right hand side of this expression is in some family of distributions, such as N or Γ, up to some normalising constant.

Remark. By the factorisation criterion, if T is a sufficient statistic for θ, the posterior $\pi(\theta \mid x)$ depends on X only through T. More precisely,

$$\pi(\theta \mid X) \propto \pi(\theta) g(T(X), \theta) h(X) \propto \pi(\theta) g(T(C), \theta)$$

Example. Consider a patient who we will test for the presence of a disease, where we have no information about the health or lifestyle of the patient. Let θ take the value 1 if the patient is infected and 0 otherwise. We have a random variable X which takes the value 1 if a given test returns a positive result and 0 if the test is negative. We know the *sensitivity* of the test $f_X(X = 1 \mid \theta = 1)$, and the *specificity* of the test $f_X(X = 0 \mid \theta = 0)$. This fully specifies the likelihood function.

We now must choose a prior distribution. For example, let $\pi(\theta = 1)$ be the estimated proportion of the general population that have the given disease. The posterior is the probability of an infection given the test result.

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

Even with a positive test result, the posterior distribution may still yield a low probability for θ, which may happen if $\pi(\theta = 1) \ll \pi(\theta = 0)$.

Example. Let θ be the mortality rate of a particular surgery, which will take values in $[0, 1]$. In the first ten operations, we observed that none of the patients died. We will model $X \sim B(10, \theta)$ and observe $X = 0$.

We must choose a prior. Suppose that we have data from other hospitals that suggests that the mortality for the surgery ranges from 3% to 20%, with an average of 10%. We can choose the prior to be the beta distribution, $\pi(\theta) \sim \text{Beta}(a, b)$, since the value of θ should range between zero and one. Let $a = 3$ and $b = 27$, which will give $\mathbb{E}[\theta] = 0.1$ and $\mathbb{P}(0.03 < \theta < 0.2) \approx 0.9$. In this case, the posterior is

$$\pi(\theta \mid X) \propto \pi(\theta) f_X(x = 0 \mid \theta) \propto \theta^{a-1}(1-\theta)^{b-1}\theta^x(1-\theta)^{n-x} = \theta^{x+a-1}(1-\theta)^{b-n-x-1}$$

This is again a beta distribution with parameters $x + a$ and $n - x + b$. The normalising constant does not need to be explicitly calculated since the form of the distribution can be recognised.

With the above data, we obtain $\pi(\theta \mid x = 0) \sim \text{Beta}(3, 37)$. This posterior has a smaller variance than the prior, and a smaller expectation due to observing no deaths. In this case, the prior and posterior have the same distribution. This is known as *conjugacy*.

4.2 Inference from the posterior

The posterior distribution $\pi(\theta \mid x)$ represents information about θ after having observed some data X. This can be used to make decisions under uncertainty.

(i) We first choose some decision $\delta \in \Delta$. For instance, in the first example, a decision could be to ask the patient to isolate from others to reduce transmission.

(ii) We define a *loss function* $L(\theta, \delta)$, which defines what loss is incurred by making decision δ given the true value of θ. In the above example, $L(\theta = 1, \delta = 1)$ is the loss incurred by asking the patient to isolate given that they have the disease.

(iii) We can now choose the decision δ that minimises

$$\int_\Theta L(\theta, \delta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

which is the posterior expectation of the loss.

4.3 Point estimation

We can use Bayesian analysis to represent an estimate for the value of θ as a decision.

Definition. The *Bayes estimator* $\hat{\theta}^{(B)}$ minimises

$$h(\delta) = \int_\Theta L(\theta, \delta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

Example. Suppose the loss function is quadratic, given by $L(\theta, \delta) = (\theta - \delta)^2$. Here,

$$h(\delta) = \int_\Theta (\theta - \delta)^2 \pi(\theta \mid x)\,\mathrm{d}\theta$$

Thus, $h(\delta) = 0$ if

$$\int_\Theta (\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta = 0 \iff \delta = \int_\Theta \theta\pi(\theta \mid x)\,\mathrm{d}x$$

Under the quadratic loss function, $\hat{\theta}^{(B)}$ can be described as the expectation of θ under the posterior distribution.

Example. Consider the absolute error loss, given by $L(\theta, \delta) = |\theta - \delta|$. In this case we have

$$h(\delta) = \int_\Theta |\theta - \delta|\pi(\theta \mid x)\,\mathrm{d}\theta = \int_{-\infty}^{\delta} -(\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta + \int_\delta^\infty (\theta - \delta)\pi(\theta \mid x)\,\mathrm{d}\theta$$

We can differentiate, using the fundamental theorem of calculus, to find

$$h'(\delta) = \int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta - \int_\delta^\infty \pi(\theta \mid x)\,\mathrm{d}\theta$$

This is zero if and only if

$$\int_{-\infty}^{\delta} \pi(\theta \mid x)\,\mathrm{d}\theta = \int_\delta^\infty \pi(\theta \mid x)\,\mathrm{d}\theta$$

This yields the median of the posterior distribution.

4.4 Credible intervals

Definition. A $100\gamma\%$ *credible interval* $(A(x), B(x))$ satisfies

$$\pi(A(x) \le \theta \le B(x) \mid x) = \gamma$$

Remark. Unlike confidence intervals, credible intervals can be interpreted conditionally on the data. For example, we could say that given a specific observation x, we are $100\gamma\%$ certain that θ lies within $(A(x), B(x))$. This credible interval is also dependent on the choice of prior distribution.

5 Hypothesis testing

5.1 Hypotheses

Definition. A *hypothesis* is an assumption about the distribution of the data X. Scientific questions are often phrased as a decision between two hypotheses. The *null hypothesis* H_0 is usually a basic hypothesis, often representing the simplest possible distribution of the data. The *alternative hypothesis* H_1 is the alternative, if H_0 were found to be false.

Example. Let $X = (X_1, \dots, X_n)$ be i.i.d. Bernoulli random variables with parameter θ. We could take, for example, $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta = \frac{3}{4}$. Alternatively, we could take $H_0 : \theta = \frac{1}{2}$ and $H_1 : \theta \neq \frac{1}{2}$.

Example. Suppose X_i takes values $0, 1, \dots$. We can take $H_0 : X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$ for some λ, and $H_1 : X_i \overset{\text{iid}}{\sim} f_1$ for some other distribution f_1. This is known as a *goodness of fit* test, which checks how well the model used for the data fits.

Definition. A *simple hypothesis* is a hypothesis which fully specifies the p.d.f. or p.m.f. of the data. A hypothesis that is not simple is called *composite*.

Example. In the first example above, $H_0 : \theta = \frac{1}{2}$ is simple, and $H_1 : \theta \neq \frac{1}{2}$ is composite. In the second example, $H_0 : X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda)$ is composite since λ was not fixed.

5.2 Testing hypotheses

Definition. A *test* of the null hypothesis H_0 is defined by a *critical region* $C \subseteq \mathcal{X}$. When $X \in C$, we *reject* the null hypothesis. This is a positive result. When $X \notin C$ we *fail to reject* the null hypothesis, or find *no sufficient evidence against* the null hypothesis. This is the negative result.
A *type I* error, or a *false positive*, is the error made by rejecting the null hypothesis when it is true. A *type II* error, or a *false negative*, is the error made by failing to reject the null hypothesis when it is not true. When H_0, H_1 are simple, we define

$$\alpha = \mathbb{P}_{H_0}(H_0 \text{ is rejected}) = \mathbb{P}_{H_0}(X \in C); \quad \beta = \mathbb{P}_{H_1}(H_0 \text{ is not rejected}) = \mathbb{P}_{H_1}(X \notin C)$$

The *size* of a test is α, which is the probability of a type I error. The *power* of a test is $1 - \beta$, which is the probability of not finding a type II error.
There is typically a tradeoff between α and β. Often, statisticians will choose an 'acceptable' value for the probability of type I errors α, and then maximise the power with respect to this fixed α. Computing the size of a test is typically simpler since it does not depend on H_1.

5.3 Neyman–Pearson lemma

Let H_0 and H_1 be simple, and let X have a p.d.f. or p.m.f. f_i under H_i. The *likelihood ratio statistic* is defined by

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

The *likelihood ratio test* is a test that rejects H_0 when Λ_x exceeds a set value k, or more formally, $C = \{x : \Lambda_x(H_0; H_1) > k\}$.

Lemma. Suppose that f_0, f_1 are nonzero on the same set, and suppose that there exists $k > 0$ such that the likelihood ratio test with critical region $C = \{x : \Lambda_x(H_0; H_1) > k\}$ has size α. Then out of all tests of size upper bounded by α, this test has the largest power.

Remark. A likelihood ratio test with size α does not always exist for any given α. However, in general we can find a *randomised test* with arbitrary size α. This is a test where, for some values of X, we reject the null hypothesis; for some values, we fail to reject the null hypothesis; and for some values we reject the null hypothesis with a random chance of rejecting the null hypothesis.

Proof. Let $\overline{C}$ be the complement of C in $\mathcal{X}$. Then, the likelihood ratio test has

$$\alpha = \int_C f_0(x)\,\mathrm{d}x\,; \quad \beta = \int_{\overline{C}} f_1(x)\,\mathrm{d}x$$

Let $C^\star$ be a critical region for a different test, with type I and II error probabilities $\alpha^\star, \beta^\star$. Here,

$$\alpha^\star = \int_{C^\star} f_0(x)\,\mathrm{d}x\,; \quad \beta^\star = \int_{\overline{C^\star}} f_1(x)\,\mathrm{d}x$$

Suppose $\alpha^\star \le \alpha$. Then, we will show $\beta \le \beta^\star$.

$$\beta - \beta^\star = \int_{\overline{C}} f_1(x)\,\mathrm{d}x - \int_{\overline{C^\star}} f_1(x)\,\mathrm{d}x$$

By cancelling the integrals on the intersection, and using the definition of C,

$$\begin{aligned}
\beta - \beta^\star &= \int_{\overline{C}\cap C^\star} f_1(x)\,\mathrm{d}x - \int_{\overline{C^\star}\cap C} f_1(x)\,\mathrm{d}x \\
&= \int_{\overline{C}\cap C^\star} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\le k} f_0(x)\,\mathrm{d}x - \int_{\overline{C^\star}\cap C} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\ge k} f_0(x)\,\mathrm{d}x \\
&\le k\left[\int_{\overline{C}\cap C^\star} f_0(x)\,\mathrm{d}x - \int_{\overline{C}^\star\cap C} f_0(x)\,\mathrm{d}x\right] \\
&= k\left[\int_{\overline{C}\cap C^\star} f_0(x)\,\mathrm{d}x + \int_{C\cap C^\star} f_0(x)\,\mathrm{d}x - \int_{C\cap C^\star} f_0(x)\,\mathrm{d}x - \int_{\overline{C}^\star\cap C} f_0(x)\,\mathrm{d}x\right] \\
&= k\left[\int_{\cap C^\star} f_0(x)\,\mathrm{d}x - \int_C f_0(x)\,\mathrm{d}x\right] \\
&= k[\alpha^\star - \alpha] \\
&\le 0
\end{aligned}$$

□

Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d., where σ_0^2 is known and μ is an unknown. We wish to find the most powerful test of fixed size α for the hypotheses $H_0 : \mu = \mu_0$ and $H_1 : \mu = \mu_1 > \mu_0$. The

likelihood ratio is

$$\begin{aligned}\Lambda_x(H_0;H_1) &= \frac{(2\pi\sigma_0^2)^{-n/2}\exp\left\{\frac{-1}{2\sigma_0^2}\sum(x_i-\mu_0)^2\right\}}{(2\pi\sigma_0^2)^{-n/2}\exp\left\{\frac{-1}{2\sigma_0^2}\sum(x_i-\mu_1)^2\right\}} \\ &= \exp\left\{\underbrace{\frac{\mu_1-\mu_0}{\sigma_0^2}}_{\geq 0} n\overline{X} + \frac{n(\mu_0-\mu_1)^2}{2\sigma_0^2}\right\}\end{aligned}$$

which depends only on $\overline{X}$, and is monotonically increasing with respect to the sample mean $\overline{X}$. Therefore, this is also monotonically increasing with respect to the statistic

$$Z = \sqrt{n}\frac{\overline{X}-\mu_0}{\sigma_0}$$

Thus, $\Lambda_x > k$ if and only if $Z > k'$ for some k'. Hence, the likelihood ratio test has critical region $\{x : Z(x) > k'\}$ for some k'. It thus suffices to find a critical region of Z with size α in order to construct the most powerful test of this size. Under H_0, $Z \sim N(0,1)$. Hence, the critical region is given by $k' = \Phi^{-1}(1-\alpha)$. This is known as a *Z-test*, since we are using the Z statistic.

5.4 p-values

Definition. Let C be a critical region of the form $\{x : T(x) > k\}$ for some test statistic T. Let $x^\star$ denote the observed data. Then, the *p-value* is

$$\mathbb{P}_{H_0}(T(X) > T(x^\star))$$

Typically, when reporting the results of a test, we describe the conclusion of the test as well as the p-value. In the example above, suppose $\mu_0 = 5$, $\mu_1 = 6$, $\alpha = 0.05$, and $x^\star = (5.1, 5.5, 4.9, 5.3)$. Here, $\overline{x^\star} = 5.2$ and $z^\star = 0.4$. The likelihood ratio test has critical region

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

The conclusion of the test here is to not reject H_0. The p-value is $1 - \Phi(z^\star) \approx 0.35$.

Proposition. Under the null hypothesis H_0, the p-value is a uniform random variable in $[0,1]$.

Proof. Let F be the distribution of the test statistic T, which we will assume for this proof is continuous. Then,

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

□

5.5 Composite hypotheses

Let $X \sim f_X(\cdot \mid \theta)$ where $\theta \in \Theta$. Let $H_0 = \theta \in \Theta_0 \subset \Theta$ and $H_1 = \theta \in \Theta_1 \subseteq \Theta$. The probabilities of type I and type II error are now dependent on the precise value of θ, rather than simply on which hypothesis is taken.

Definition. The *power function* for a test C is

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

The *size* of a test C is

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

A test is *uniformly most powerful* of size α if, for any test $C^\star$ with power function $W^\star$ and size upper bounded by α, for all $\theta \in \Theta_1$ we have $W(\theta) \geq W^\star(\theta)$. Such tests need not exist. In simple models, many likelihood ratio tests are uniformly most powerful.

Example (one-sided test for normal location)**.** Let $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. where σ_0^2 is known and μ is unknown. Let $H_0 : \mu \leq \mu_0$ and $H_1 : \mu > \mu_0$ for some fixed μ_0. We claim that the simple hypothesis test given by $H_0' : \mu = \mu_0$ and $H_1' : \mu = \mu_1 > \mu_0$ is uniformly most powerful for H_0 and H_1. The power function is

$$\begin{aligned} W(\mu) &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma_0} = Z < z_\alpha = \Phi^{-1}(1-\alpha)\right) \\ &= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma_0} > z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}\right) \\ &= 1 - \Phi\left(x_\alpha + \sqrt{n}\frac{\mu_0 - \mu}{\sigma_0}\right) \end{aligned}$$

The test has size α since $\sup_{w \in \Theta_0} W(\mu) = \alpha$. It remains to show that this power function dominates all other power functions $W^\star$ of size α in the alternative space Θ_1. First, observe that the critical region depends only on μ_0, and not on μ_1. In particular, for any $\mu_1 > \mu_0$, we have that the critical region C is the likelihood ratio test for the simple hypothesis test $H_0' : \mu = \mu_0$ and $H_1' : \mu = \mu_1$. We can also see $C^\star$ as a test of H_0' versus H_1', and for these simple hypotheses, $C^\star$ has size

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

By the Neyman–Pearson lemma, C has power no smaller than $C^\star$ for H_0' against H_1':

$$W(\mu_1) \geq W^\star(\mu_1)$$

Since this is true for all $\mu_1 > \mu_0$, the result holds, and the test C satisfies the property for being uniformly most powerful.

5.6 Generalised likelihood ratio test

Definition. Suppose we have *nested hypotheses*, i.e. $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where $\Theta_0 \subset \Theta_1$. The *generalised likelihood ratio* is given by

$$\Lambda_x(H_0; H_1) = \frac{\sup_{\theta \in \Theta_1} f_X(x \mid \theta)}{\sup_{\theta \in \Theta_0} f_X(x \mid \theta)}$$

Large values indicate a better fit under the alternative hypothesis. The *generalised likelihood ratio test* rejects the null hypothesis when Λ_x is sufficiently large.

Example (two-sided test for normal location)**.** Let $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. where σ_0^2 is known and μ is unknown. Let $H_0 : \mu = \mu_0$ and $H_1 : \mu \in \mathbb{R}$ for some fixed μ_0. In this model, the generalised likelihood ratio is

$$\Lambda_x(H_0; H_1) = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma_0^2}\Sigma_{i=1}^n (x_i - \overline{X})^2\right\}}{(2\pi\sigma_0^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma_0^2}\Sigma_{i=1}^n (x_i - \mu_0)^2\right\}}$$

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

Under H_0, $\sqrt{n}\frac{\overline{X}-\mu_0}{\sigma_0} \sim N(0,1)$. Hence, $2 \log \Lambda_x \sim \chi_1^2$. Therefore, the critical region of this generalised likelihood ratio test is

$$C = \left\{x : \frac{n}{\sigma_0^2}(\overline{X} - \mu_0)^2 > \chi_1^2(\alpha)\right\}$$

where $\chi_1^2(\alpha)$ is the upper α point of χ_1^2. This is called a *two-sided test* since there are two tails on the critical region, plotting with respect to $\sqrt{n}\frac{\overline{X}-\mu_0}{\sigma_0}$.

5.7 Wilks' theorem

Definition. The *dimension* of a hypothesis $H_0 : \theta \in \Theta_0$ is the number of 'free parameters' in this space.

Example. If $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_1 = \cdots = \theta_p = 0\}$, then the dimension of H_0 is $k - p$.

Let $A \in \mathbb{R}^{p \times k}$ be a $p \times k$ matrix with linearly independent rows. Let $b \in \mathbb{R}^p$ for $p < k$, then we define $\Theta_0 = \{\theta \in \mathbb{R}^k : A\theta = b\}$. Then the dimension of θ is $k - p$.

Let Θ_0 be a Riemannian manifold. We use differential geometry to deduce the dimensionality of such a manifold.

Theorem. Suppose $\Theta_0 \subset \Theta_1$, and $\dim \Theta_1 - \dim \Theta_0 = p$. Let $X = (X_1, \dots, X_n)$ be i.i.d. random variables under $f_x(\cdot \mid \theta)$ where $\theta \in \Theta_0^\circ$. Then, under some regularity conditions, as $n \to \infty$ we have

$$2 \log \Lambda_x \sim \chi_p^2$$

More precisely, for all $\ell \in \mathbb{R}_+$,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(2\log\Lambda_x \le \ell\right) = \mathbb{P}\left(\Xi \le \ell\right); \quad \Xi \sim \chi^2_p$$

Remark. If n is large, this theorem allows us to implement a generalised likelihood ratio test even if we cannot find the exact distribution of $2\log\Lambda_x$. Frequentist guarantees obtained from such a test will be approximate.

Example. In the two-sided test for normal location, $\dim\Theta_1 = 1$ and $\dim\Theta_0 = 0$ hence the difference in dimensions is 1. Then, Wilks' theorem implies that $2\log\Lambda_x$ is approximately distributed according to χ^2_1, although the result is exact in this particular case.

5.8 Goodness of fit

Let $X_1, \dots, X_n$ be i.i.d. samples taking values in $\{1, \dots, k\}$. Let $p_i = \mathbb{P}(X_1 = i)$, and let N_i be the number of samples equal to i, so $\sum_i p_i = 1$ and $\sum_i N_i = n$. The parameters here are $p = (p_1, \dots, p_k)$, which has $k-1$ dimensions. A *goodness of fit test* has a null hypothesis of the form $H_0 :\ p_i = \widetilde{p}_i$ for all i, for a fixed $\widetilde{p} = (\widetilde{p}_1, \dots, \widetilde{p}_k)$. The alternative hypothesis H_1 does not constrain p.

The model is $(N_1, \dots, N_k) \sim \text{Multi}(n; p_1, \dots, p_k)$. The likelihood function is

$$L(p) \propto p_1^{N_1} \cdots p_k^{N_k} \implies \ell(p) = \text{constant} + \sum_i N_i \log p_i$$

The generalised likelihood ratio is

$$2\log\Lambda_x = 2\left(\sup_{p\in\Theta_1} \ell(p) - \sup_{p\in\Theta_0} \ell(p)\right) = 2(\ell(\hat{p}) - \ell(\widetilde{p}))$$

where $\hat{p}$ is the maximum likelihood estimator under H_1. To find $\hat{p}$, we typically use the method of Lagrange multipliers.

$$\mathcal{L}(p, \lambda) = \sum_i N_i \log p_i - \lambda\Big(\sum p_i - 1\Big)$$

We can compute that

$$\hat{p}_i = \frac{N_i}{n}$$

This is simply the fraction of observed samples of type i.

5.9 Pearson statistic

Let $o_i = N_i$ be the observed number of samples of type i, and $e_i = n\widetilde{p}_i$ be the expected value under the null hypothesis of the number of samples of type i. Here, we can write

$$2\log\Lambda = 2\sum_i N_i \log\left(\frac{N_i}{n\widetilde{p}_i}\right) = 2\sum_i o_i \log\frac{o_i}{e_i}$$

Let $\delta_i = o_i - e_i$. Then

$$2\log\Lambda = 2\sum_i (e_i + \delta_i)\log\left(1 + \underbrace{\frac{\delta_i}{e_i}}_{\text{small when } n \text{ large}}\right)$$

By taking the Taylor expansion, we arrive at

$$2\sum_i \left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i}\right)$$

Note that $\sum_i \delta_i = \sum_i (o_i - e_i) = n - n = 0$, so we can simplify and find

$$\sum_i \frac{\delta_i^2}{e_i} = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

This is *Pearson's χ^2 statistic*. This is also referred to a χ^2_{k-1} when performing a hypothesis test.

Example. Mendel performed an experiment in which 556 different pea plants were created from a small set of ancestors. Each descendent was either yellow or green, and either wrinkled or smooth, giving four possibilities in total. The observed result was

$$N = \left(\underbrace{315}_{SG}, \underbrace{108}_{SY}, \underbrace{102}_{WG}, \underbrace{31}_{WY}\right)$$

Mendel's theory gives a null hypothesis $H_0 : p = \widetilde{p} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$. Here,

$$2\log\Lambda = 0.618; \quad \sum_i \frac{(o_i - e_i)^2}{e_i} = 0.604$$

These are referred to a χ^2_3 distribution. We observe that $\chi^2_3(0.05) = 7.815$, so we fail to reject the null hypothesis with a test of size 5%. We can compute that the p-value is $\mathbb{P}\left(\chi^2_3 > 0.6\right) \approx 0.96$, so there is a very high probability of observing a more extreme value than observed.

5.10 Goodness of fit for composite null

Suppose $H_0 : p_i = p_i(\theta)$ for some $\theta \in \Theta_0$, and $H_1 : p$ has any distribution on $\{1, \dots, k\}$. We can compute

$$2\log\Lambda = 2\left(\sup_p \ell(p) - \sup_{\theta\in\Theta} \ell(p(\theta))\right)$$

We can sometimes compute these quantities explicitly, and hence find a test which refers this test statistic to a χ^2_p distribution where $p = \dim\Theta_1 - \dim\Theta_0 = (k-1) - \dim\Theta_0$.

Example. Consider a population of individuals who may have one of three genotypes, which occur with probabilities $(p_1, p_2, p_3) = (\theta^2, 2\theta(1-\theta), (1-\theta)^2)$. In this case, we can find the maximum likelihood estimator under the null hypothesis to be

$$\hat{\theta} = \frac{2N_1 + N_2}{2n}$$

Hence,

$$2\log\Lambda = 2(\ell(\hat{p}) - \ell(\hat{\theta}))$$

where $\hat{p}_i = \frac{N_1}{n}$ as found previously. This can be computed explicitly and referred to a χ^2_1 distribution. We can check that, in this model,

$$2\log\Lambda = \sum_i o_i \log\frac{o_i}{e_i}$$

where $o_i = N_i$ and $e_i = np_i(\hat{\theta})$. We can approximate this using the Pearson statistic, $\sum_i \frac{(o_i - e_i)^2}{e_i}$.

5.11 Testing independence in contingency tables

Suppose we have observations $(X_1, Y_1), \dots, (X_n, Y_n)$ which are i.i.d., where the X_i take values in $1, \dots, r$ and the Y_i take values in $1, \dots, c$. We wish to test whether the X_i and Y_i are independent. We will summarise this data into a sufficient statistic known as a *contingency table* N, given by

$$N_{ij} = |\{\ell : 1 \le \ell \le n, (X_\ell, Y_\ell) = (i, j)\}|$$

So N_{ij} is the number of samples of type (i, j).

Example. Suppose we observe n samples, and each sample has probability p_{ij} of being of type (i, j). Flattening (N_{ij}) into a vector, this has a multinomial distribution with parameters (p_{ij}) (also flattened into a vector). The null hypothesis is $H_0 : p_{ij} = p_{i+}p_{+j}$ where $p_{i+} = \sum_j p_{ij}$ and $p_{+j} = \sum_i p_{ij}$. The alternative hypothesis places no restrictions on the p_{ij} apart from that it sums to 1 and has nonnegative entries. We find

$$2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}$$

where $\hat{p}_{ij}$ is the maximum likelihood estimator under H_1, and where $\hat{p}_{i+}$ and $\hat{p}_{+j}$ are the maximum likelihood estimators under H_0. These can be found using the method of Lagrange multipliers. In particular,

$$\hat{p}_{ij} = \frac{N_{ij}}{n}; \quad \hat{p}_{i+} = \frac{N_{i+}}{n} = \frac{1}{n} \sum_{j=1}^{c} N_{ij}; \quad \hat{p}_{+j} = \frac{N_{+j}}{n} = \frac{1}{n} \sum_{i=1}^{r} N_{ij}$$

Writing $o_{ij} = N_{ij}$ and $e_{ij} = n\hat{p}_{i+}\hat{p}_{+j}$,

$$2 \log \Lambda = \sum_{i,j} o_{ij} \log \frac{o_{ij}}{e_{ij}} \approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

By Wilks' theorem, these test statistics have an approximate χ^2_p distribution, where $p = \dim \Theta_1 - \dim \Theta_0 = (rc - 1) - (r - 1 + c - 1) = (r - 1)(c - 1)$.

The χ^2 test for independence has a number of weaknesses.

(i) The χ^2 approximation requires n to be large. A reasonable heuristic is to require $N_{ij} \ge 5$ for *all* i, j. If this is not possible, we can perform an *exact test* (which is non-examinable).

(ii) The χ^2 test often has a low power. Heuristically, this is because the alternative space Θ_1 is too large, and there are many possible models that lie in this space.

Note that this test also applies when n is a random variable with a Poisson distribution. This is often the case when we do not fix the number of samples. The proof is not provided in this course.

5.12 Testing homogeneity in contingency tables

Example. Suppose we perform a clinical trial on 150 patients, who are randomly assigned to one of three groups of equal size. The first two sets take a drug with different doses, and the third set takes a placebo.

	improved	no difference	worse	
placebo	18	17	15	50
half dose	20	10	20	50
full dose	25	13	12	50

In the previous section, we fixed the total number of samples. Here, we fix the total number of samples, and the total number of samples in each row. We suppose

$$N_{i1}, \dots, N_{ic} \sim \text{Multinomial}(n_{i+}; p_{i1}, \dots, p_{ic})$$

which are independent for each row i of the table. The null hypothesis for homogeneity is that $p_{1j} = p_{2j} = \cdots = p_{rj}$ for all j. The alternative hypothesis assumes that $p_{i1}, \dots, p_{ic}$ is any arbitrary probability vector for each row i. Under the alternative hypothesis,

$$L(p) = \prod_{i=1}^{r} \frac{n_{i+}!}{N_{i1}! \cdots N_{ic}!} p_{i1}^{N_{i1}} \cdots p_{ic}^{N_{ic}}$$

Hence,

$$\ell(p) = \text{constant} + \sum_{i,j} N_{ij} \log p_{ij}$$

This is the same likelihood as the independence test above. To define the maximum likelihood estimator we can again use the method of Lagrange multipliers with constraints $\sum_j p_{ij} = 1$ for each i. We find

$$\hat{p}_{ij} = \frac{N_{ij}}{n_{i+}}$$

Under the null hypothesis, we let $p_j = p_{ij}$ for any i.

$$\ell(p) = \text{constant} + \sum_{i,j} N_{ij} \log p_j = \sum_j N_{+j} \log p_j$$

We have the constraint $\sum_j p_j = 1$. Using the method of Lagrange multipliers,

$$\hat{p}_j = \frac{N_{+j}}{n_{++}}$$

Hence,

$$2 \log \Lambda = 2 \sum_{i,j} N_{ij} \log \frac{\hat{p}_{ij}}{\hat{p}_j} = 2 \sum_{i,j} N_{ij} \log \frac{N_{ij}}{n_{i+} N_{+j} / n_{++}}$$

This is precisely the same test statistic as the test for independence above. The only difference is that n_{i+} is fixed in this model. Further, if $o_{ij} = N_{ij}$ and $e_{ij} = n_{i+} \hat{p}_j = \frac{n_{i+} N_{+j}}{n_{++}}$, we have

$$2 \log \Lambda = 2 \sum_{i,j} o_{ij} \log \frac{o_{ij}}{e_{ij}} \approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

By Wilks' theorem, this is asymptotically a χ^2_p distribution. Here,

$$p = \dim \Theta_1 - \dim \Theta_0 = r(c-1) - (c-1) = (r-1)(c-1)$$

This is again exactly the same as in the χ^2 test for independence. Operationally, the tests for homogeneity and independence are therefore completely identical; we reject the null hypothesis for one test if and only if we reject the null for the other. In the example above,

$$2 \log \Lambda = 5.129; \quad \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 5.173$$

Referring this to a χ^2_4 distribution, the upper 0.05-point is 9.488. Hence, we do not reject the null hypothesis at the 5% significance level.

5.13 Tests and confidence sets

Definition. The *acceptance region* A of a test is the complement of the critical region.

Theorem. Let $X \sim f_X(\cdot \mid \theta)$ for some $\theta \in \Theta$. Suppose that for each $\theta_0 \in \Theta$, there exists a test of size α with acceptance region $A(\theta_0)$ for the null hypothesis $\theta = \theta_0$. Then

$$I(X) = \{\theta : X \in A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence set.
Now suppose there exists a set $I(X)$ which is a $100(1-\alpha)\%$ confidence set for θ. Then

$$A(\theta_0) = \{x : \theta_0 \in I(x)\}$$

is the acceptance region of a test of size α for the hypothesis $\theta = \theta_0$.

Proof. Observe that for both parts of the theorem,

$$\theta_0 \in I(X) \iff X \in A(\theta_0) \iff \text{fail to reject } H_0 \text{ with data } X$$

For the first part, we assume that $\mathbb{P}_\theta(\text{fail to reject } H_0 \text{ with data } X) = 1-\alpha$, and we want to show $\mathbb{P}_\theta(\theta_0 \in I(X)) = 1-\alpha$. The second part is the converse. □

Example. Let $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ be i.i.d. with σ_0^2 known and μ unknown. We found that a $100(1-\alpha)\%$ confidence interval for μ is

$$I(X) = \left(\overline{X} \pm \frac{Z_{\alpha/2}\sigma_0}{\sqrt{n}}\right)$$

Hence, by the second part of the theorem above, we can find a test for $H_0 : \mu = \mu_0$ with size α by

$$A(\mu_0) = \{x : \mu_0 \in I(x)\} = \left\{x : \mu_0 \in \left[\overline{x} \pm \frac{Z_{\alpha/2}\sigma_0}{\sqrt{n}}\right]\right\}$$

This is equivalent to rejecting H_0 when

$$\left|\sqrt{n}\frac{\mu_0 - \overline{X}}{\sigma_0}\right| > Z_{\alpha/2}$$

This is a two-sided test for normal location.

6 The normal linear model

6.1 Multivariate normal distribution

Let $X = (X_1, \dots, X_n)$ be a vector of random variables. Then we define

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}; \quad \operatorname{Var}(X) = \left(\mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right]\right)_{i,j}$$

The familiar linearity results are

$$\mathbb{E}[AX + b] = A\mathbb{E}[X] + b; \quad A\operatorname{Var}(X)A^\intercal$$

where $A \in \mathbb{R}^{k\times n}, b \in \mathbb{R}^k$ are constant.

Definition. We say that X has a *multivariate normal distribution* if, for any fixed $t \in \mathbb{R}^n$, we have $t^\intercal X \sim N(\mu, \sigma^2)$ for some parameters μ, σ^2.

Proposition. Let X be multivariate normal. Then $AX + b$ is multivariate normal, where $A \in \mathbb{R}^{k\times n}, b \in \mathbb{R}^k$ are constant.

Proof. Let $t \in \mathbb{R}^k$. Then,

$$t^\intercal(Ax + b) = \underbrace{(A^\intercal t)^\intercal X}_{\sim N(\mu,\sigma^2)} + t^\intercal b$$

which is the sum of a normal random variable and a constant. So this is $N(\mu + t^\intercal b, \sigma^2)$. □

Proposition. A multivariate normal distribution is fully specified by its mean and covariance matrix.

Proof. Let X_1, X_2 be multivariate normal vectors with the same mean μ and the same covariance matrix Σ. We will show that these two random variables have the same moment generating function, and hence the same distribution.

$$M_{X_1}(t) = \mathbb{E}\left[e^{1\cdot t^\intercal X_1}\right]$$

Note that $t^\intercal X_1$ is univariate normal. Hence, this is equal to

$$M_{X_1}(t) = \exp\left(1 \cdot \mathbb{E}[t^\intercal X_1] + \frac{1}{2}\operatorname{Var}(t^\intercal X_1) \cdot 1^2\right) = \exp\left(t^\intercal\mu + \frac{1}{2}t^\intercal\Sigma t\right)$$

This depends only on μ and Σ, and we obtain the same moment generating function for X_2. □

6.2 Orthogonal projections

Definition. A matrix $P \in \mathbb{R}^{n\times n}$ is an *orthogonal projection* onto its column space $\operatorname{col}(P)$ if, for all $v \in \operatorname{col}(P)$, we have $Pv = v$, and for all $w \in \operatorname{col}(P)^\perp$, we have $Pw = 0$.

Proposition. P is an orthogonal projection if and only if it is idempotent and symmetric.

Proof. If P is idempotent and symmetric, let $v \in \operatorname{col}(P)$, so $v = Pa$ for some $a \in \mathbb{R}^n$. Then, $Pv = PPa = Pa = v$. Now, let $w \in \operatorname{col}(P)^\perp$. By definition, $P^\intercal w = 0$. By symmetry, $Pw = 0$.

Now, suppose P is an orthogonal projection. Any vector $a \in \mathbb{R}^n$ can be uniquely written as $a = v + w$ where $v \in \operatorname{col}(P)$ and $w \in \operatorname{col}(P)^\perp$. Then $PPa = PPv + PPw = Pv = P(v + w) = Pa$. As this holds for all a, we have that P is idempotent. Let $u_1, u_2 \in \mathbb{R}^n$, and note $(Pu_1)\cdot((I - P)u_2) = 0$, as $Pu_1 \in \operatorname{col}(P)$ and $(I - P)u_2 \in \operatorname{col}(P)^\perp$. We have $u_1^\intercal P^\intercal(I - P)u_2 = 0$. Since this holds for all u_1, u_2, $P^\intercal(I - P) = 0$ so $P^\intercal = P^\intercal P$. Note that $P^\intercal P$ is symmetric, so $P^\intercal$ is symmetric, and hence P is symmetric. □

Corollary. Let P be an orthogonal projection matrix. Then $I - P$ is also an orthogonal projection matrix.

Proof. Clearly, if P is symmetric, so is $I-P$, so it suffices to prove idempotence. We have $(I-P)(I-P) = I - 2P + P^2 = I - 2P + P = I - P$ as required. □

Proposition. If P is an orthogonal projection, then $P = UU^\intercal$ where the columns of U are an orthonormal basis for the column space of P.

Proof. First, we show that $UU^\intercal$ is an orthogonal projection. This is clearly symmetric. It is idempotent: $UU^\intercal UU^\intercal = UU^\intercal$ since $U^\intercal U = I$, as the columns of U form an orthonormal basis for the column space of P. Further, the column space of P is exactly the column space of $UU^\intercal$. □

Proposition. The rank of an orthogonal projection matrix is equal to its trace.

Proof. The rank is the dimension of the column space, which is $\operatorname{rank} P = \operatorname{rank}(U^\intercal U) = \operatorname{tr}(U^\intercal U) = \operatorname{tr}(UU^\intercal) = \operatorname{tr} P$. □

Theorem. Let X be multivariate normal, where $X \sim N(0, \sigma^2 I)$, and let P be an orthogonal projection. Then

(i) $PX \sim N(0, \sigma^2 P)$, and $(I - P)X \sim N(0, \sigma^2(I - P))$, and these two random variables are independent;

(ii) $\frac{\|PX\|^2}{\sigma^2} \sim \chi^2_{\operatorname{rank} P}$.

Proof. The vector $(P, I - P)^\intercal X$ is multivariate normal, since it is a linear function of X. This distribution is fully specified by its mean and variance.

$$\mathbb{E}\left[\begin{pmatrix} PX \\ (I-P)X \end{pmatrix}\right] = \begin{pmatrix} P \\ I - P \end{pmatrix} \mathbb{E}[X] = 0$$

Further,

$$\operatorname{Var}\left(\begin{pmatrix} PX \\ (I-P)X \end{pmatrix}\right) = \begin{pmatrix} P \\ I - P \end{pmatrix} \sigma^2 I \begin{pmatrix} P \\ I - P \end{pmatrix}^\intercal = \sigma^2 \begin{pmatrix} P^2 & P(I-P) \\ P(I-P) & (I-P)^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} P & 0 \\ 0 & I - P \end{pmatrix}$$

Now we must show that the variables $PX, (I - P)X$ are independent. Let $Z \sim N(0, \sigma^2 P), Z' \sim N(0, \sigma^2(I - P))$ be independent. Then we can see that $(Z, Z')^\intercal$ is multivariate normal with

$$\mu = 0; \quad \Sigma = \begin{pmatrix} P & 0 \\ 0 & I - P \end{pmatrix}$$

Hence $(PX, (1 - P)X)^\intercal$ is equal in distribution to $(Z, Z')^\intercal$. So PX is independent of $(I - P)X$.

We must show that $\frac{\|PX\|^2}{\sigma^2} \sim \chi^2_{\operatorname{rank} P}$. Note that

$$\frac{\|PX\|^2}{\sigma^2} = \frac{X^\intercal P^\intercal PX}{\sigma^2} = \frac{X^\intercal (UU^\intercal)^\intercal UU^\intercal X}{\sigma^2} = \frac{\|U^\intercal X\|^2}{\sigma^2}$$

Note, $U^\intercal X \sim N(0, \sigma^2 U^\intercal U) = N(0, \sigma^2 I_{\operatorname{rank} P})$. So

$$\frac{(U^\intercal X)_i}{\sigma} \overset{\text{iid}}{\sim} N(0,1)$$

for $i = 1, \dots, \operatorname{rank} P$. Hence

$$\frac{\|PX\|^2}{\sigma^2} = \sum_{i=1}^{\operatorname{rank} P} \left(\frac{(U^\intercal X)_i}{\sigma}\right)^2 \sim \chi^2_{\operatorname{rank} P}$$

□

Theorem. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ for some unknown $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. The maximum likelihood estimators for μ and σ are

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_i X_i; \quad \hat{\sigma}^2 = \frac{S_{xx}}{n} = \frac{\sum_i \left(X_i - \overline{X}\right)^2}{n}$$

Further,

(i) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;

(ii) $\frac{S_{xx}}{\sigma^2} \sim \chi^2_{n-1}$;

(iii) $\overline{X}, S_{xx}$ are independent.

Proof. Let P be the square $n \times n$ matrix with all entries $\frac{1}{n}$. This is an orthogonal projection matrix, as it is symmetric and idempotent. Note that

$$PX = \begin{pmatrix} \overline{X} \\ \vdots \\ \overline{X} \end{pmatrix}$$

We will write the observations X as

$$X = \underbrace{\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}}_{M} + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2 I)$$

Note that $\overline{X}$ is a function of $P\varepsilon$, since $\overline{X} = (PX)_1 = (PM + P\varepsilon)_1$. Further,

$$S_{xx} = \sum_i \left(X_i - \overline{X}\right)^2 = \|X - PX\|^2 = \|(I-P)X\|^2 = \|(I-P)\varepsilon\|^2$$

Hence S_{xx} is a function of $(I-P)\varepsilon$. Since $P\varepsilon$ and $(I-P)\varepsilon$ are independent, $\overline{X}$ and S_{xx} are independent. Since $I - P$ is a projection with rank equal to its trace $n - 1$, we apply the previous theorem to obtain

$$S_{xx} = \|(I-P)\varepsilon\|^2 \chi^2_{n-1}$$

□

6.3 Linear model

Suppose we have data in pairs $(x_1, Y_1), \dots, (x_n, Y_n)$, where $Y_i \in \mathbb{R}, x_i \in \mathbb{R}^p$. The Y_i are known as the *response* variables, or the *dependent* variables. The x_{i1}, x_{ip} are the *predictors*, or *independent* variables. We will model the expectation of the response Y_i as a linear function of the predictors $(x_{i1}, \dots, x_{ip})$.

Example. Let Y_i be the number of insurance claims that driver i makes in a given year, and $x_{i1}, \dots, x_{ip}$ is a set of variables about the specific driver. Predictors include age, the number of years they have held their license, and the number of points on their license, for instance.

We assume that

$$Y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

where $\alpha \in \mathbb{R}$ is an *intercept*, β_i are the *coefficients*, and ε is a *noise vector*, which is a random variable. The intercept and coefficients are the parameters of interest. We will often eliminate the intercept by making one of the predictors $x_{i1} = 1$ for all i, so β_1 plays the role of the intercept.

Note that we can use a linear model to model nonlinear relationships. For example, suppose $Y_i = a + bz_i + cz_i^2 + \varepsilon_i$. We can rephrase this as a linear model with $x_i = (1, z_i, z_i^2)$.

The coefficient β_j can be interpreted as the effect on Y_i of increasing x_{ij} by one, while keeping all other predictors fixed. This cannot be interpreted as a causal relationship, unless this is a randomised control experiment.

6.4 Matrix formulation

Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}; \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}; \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}; \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

We call X the *design matrix.* The linear model is that

$$Y = X\beta + \varepsilon$$

$X\beta$ is considered fixed. Since ε is random, this makes Y into a random variable.

6.5 Assumptions

We make a number of *moment assumptions* on the noise vector ε. This allows us to deduce more results about the linear model.

(i) $\mathbb{E}[\varepsilon] = 0 \implies \mathbb{E}[Y_i] = x_i^\intercal \beta$;

(ii) $\operatorname{Var}(\varepsilon) = \sigma^2 I$, which is equivalent to both $\operatorname{Var}(\varepsilon_i) = \sigma^2$ and $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$. This property is known as *homoscedasticity*.

We will always assume that the design matrix X has full rank p, or equivalently, that it has linearly independent columns. Since $X \in \mathbb{R}^{n \times p}$, this requires that $n \geq p$, so we need at least as many samples as we have predictors.

6.6 Least squares estimation

Definition. The *least squares estimator* $\hat{\beta}$ minimises the *residual sum of squares*, which is

$$S(\beta) = \|Y - X\beta\|^2 = \sum_i \left(Y_i - x_i^\intercal \beta\right)^2$$

The term $Y_i - x_i^\intercal \beta$ is called the ith residual.

Since $S(\beta)$ is a positive definite quadratic in β, it is minimised at the stationary point.

$$\left.\frac{\partial S(\beta)}{\partial \beta_k}\right|_{\beta=\hat{\beta}} = 0 \iff \forall k,\ -2\sum_{i=1}^{n} x_{ik}\left(Y_i - \sum_k x_{ij}\hat{\beta}_j\right) = 0 \iff X^\intercal X\hat{\beta} = X^\intercal Y$$

As X has full column rank, $X^\intercal X$ is invertible.

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$$

This is notably a linear function of Y, given fixed X. Note that

$$\mathbb{E}\left[\hat{\beta}\right] = (X^\intercal X)^{-1} X^\intercal \mathbb{E}\left[Y\right] = (X^\intercal X)^{-1} X^\intercal X\beta = \beta$$

So $\hat{\beta}$ is an unbiased estimator. Further,

$$\begin{aligned}\operatorname{Var}\left(\hat{\beta}\right) &= (X^\intercal X)^{-1} X^\intercal \operatorname{Var}\left(Y\right)\left[(X^\intercal X)^{-1} X^\intercal\right]^\intercal \\ &= (X^\intercal X)^{-1} X^\intercal \sigma^2 I\left[(X^\intercal X)^{-1} X^\intercal\right]^\intercal \\ &= \sigma^2 (X^\intercal X)^{-1}\end{aligned}$$

Theorem (Gauss–Markov theorem)**.** Let an estimator $\beta^\star$ of β be unbiased and a linear function of Y, so $\beta^\star = CY$. Then, for any fixed $t \in \mathbb{R}^p$, we have

$$\operatorname{Var}\left(t^\intercal \hat{\beta}\right) \leq \operatorname{Var}\left(t^\intercal \beta^\star\right)$$

where $\hat{\beta}$ is the least squares estimator. We say that $\hat{\beta}$ is the best linear unbiased estimator (BLUE).

Remark. We can think of $t \in \mathbb{R}^p$ as a vector of predictors for a new sample. Then $t^\intercal \hat{\beta}$ is the prediction for $\mathbb{E}\left[Y_i\right]$ for this new sample, using the least squares estimator. $t^\intercal \beta^\star$ is the prediction with $\beta^\star$. In both cases, the prediction is unbiased.

Proof. Note that

$$\operatorname{Var}\left(t^\intercal \beta^\star\right) - \operatorname{Var}\left(t^\intercal \hat{\beta}\right) = t^\intercal\left[\operatorname{Var}\left(\beta^\star\right) - \operatorname{Var}\left(\hat{\beta}\right)\right]t$$

To prove that this quantity is always non-negative, we must show that $\operatorname{Var}\left(\beta^\star\right) - \operatorname{Var}\left(\hat{\beta}\right)$ is positive semidefinite. Let $A = C - (X^\intercal X)^{-1} X^\intercal$. Note that $\mathbb{E}\left[AY\right] = \mathbb{E}\left[\beta^\star\right] - \mathbb{E}\left[\hat{\beta}\right] = 0$. Also, $\mathbb{E}\left[AY\right] =$

$A\mathbb{E}[Y] = AX\beta$. This holds for all β, so $AX = 0$. Now, since $X^\intercal X$ is symmetric,

$$\begin{aligned}
\operatorname{Var}(\beta^\star) &= \operatorname{Var}(CY) \\
&= \operatorname{Var}\big((A + (X^\intercal X)^{-1}X^\intercal)Y\big) \\
&= \big[A + (X^\intercal X)^{-1}X^\intercal\big]\operatorname{Var}(Y)\big[A + (X^\intercal X)^{-1}X^\intercal\big]^\intercal \\
&= \big[A + (X^\intercal X)^{-1}X^\intercal\big]\sigma^2 I\big[A + (X^\intercal X)^{-1}X^\intercal\big]^\intercal \\
&= \sigma^2\big(AA^\intercal + (X^\intercal X)^{-1} + AX(X^\intercal X)^{-1} + (X^\intercal X)^{-1}X^\intercal A^\intercal\big) \\
&= \sigma^2 AA^\intercal + \operatorname{Var}\big(\hat{\beta}\big) \\
\operatorname{Var}(\beta^\star) - \operatorname{Var}\big(\hat{\beta}\big) &= \sigma^2 AA^\intercal
\end{aligned}$$

Note that the outer product $AA^\intercal$ is always positive semidefinite. □

6.7 Fitted values and residuals

Definition. The *fitted values* are $\hat{Y} = X\hat{\beta} = X(X^\intercal X)^{-1}X^\intercal Y$, where $P = X(X^\intercal X)^{-1}X^\intercal$ is the *hat matrix*. The *residuals* are $Y - \hat{Y} = (I - P)Y$.

Proposition. P is the orthogonal projection onto the column space of the design matrix.

Proof. If v is in the column space of X, then $v = Xb$ for some b. Hence

$$Pv = X(X^\intercal X)^{-1}X^\intercal Xb = Xb = v$$

If w is in the orthogonal complement, then

$$Pw = X(X^\intercal X)^{-1}\underbrace{X^\intercal w}_{0} = 0$$

□

Corollary. The fitted values are an orthogonal projection of the response variables to the column space of the design matrix. The residuals are orthogonal to the column space.

6.8 Normal linear model

The normal linear model is a linear model under the assumption that $\varepsilon \sim N(0, \sigma^2 I)$, where σ^2 is unknown. The parameters in the model are now (β, σ^2). The likelihood function in the normal linear model is

$$L(\beta, \sigma^2) = f_Y(y \mid \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - x_i^\intercal \beta)^2\right\}$$

The log-likelihood is

$$\ell(\beta, \sigma^2) = \text{constant} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

To maximise this as a function of β for any fixed σ^2, we must minimise the residual sum of squares $S(\beta) = \|Y - X\beta\|^2$. So $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$ is the maximum likelihood estimator of β. Further, $\hat{\sigma}^2 = n^{-1}\left\|Y - X\hat{\beta}\right\|^2 = n^{-1}\left\|\hat{Y} - Y\right\|^2 = n^{-1}\|(I-P)Y\|^2$.

Theorem. In the normal linear model,
(i) $\hat{\beta} \sim N(\beta, \sigma^2(X^\intercal X)^{-1})$;
(ii) $n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$;
(iii) $\hat{\beta}, \hat{\sigma}^2$ are independent.

Proof. We prove each part separately.

(i) We already know that $\mathbb{E}\left[\hat{\beta}\right] = \beta$, and $\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2(X^\intercal X)^{-1}$. So it suffices to show that $\hat{\beta}$ is a normal vector. Since $\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$, it is a linear function of a normal vector, so is a normal vector.

(ii) Observe that

$$n\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|(I-P)Y\|^2}{\sigma^2} = \frac{\|(I-P)(X\beta + \varepsilon)\|^2}{\sigma^2}$$

Since $(I-P)X = 0$ as P is the orthogonal projection onto the column space of X,

$$n\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\|(I-P)\varepsilon\|^2}{\sigma^2} \sim \chi^2_{\mathrm{tr}(I-P)}$$

where $\mathrm{tr}(I-P) = \mathrm{tr}\, I - \mathrm{tr}\, P = n - p$ since $X \in \mathbb{R}^{n\times p}$ is assumed to have full rank.

(iii) Note that $\hat{\sigma}^2$ is a function of $(I-P)\varepsilon$, and

$$\begin{aligned}\hat{\beta} &= (X^\intercal X)^{-1} X^\intercal Y \\ &= (X^\intercal X)^{-1} X^\intercal (X\beta + \varepsilon) \\ &= \beta + (X^\intercal X)^{-1} X^\intercal \varepsilon \\ &= \beta + (X^\intercal X)^{-1} X^\intercal P\varepsilon\end{aligned}$$

is a function of $P\varepsilon$. Since $(I-P)\varepsilon$ and $P\varepsilon$ are independent, so are $\hat{\beta}, \hat{\sigma}^2$.

□

Note,

$$\mathbb{E}\left[\frac{n\hat{\sigma}^2}{\sigma^2}\right] = \mathbb{E}\left[\chi^2_{n-p}\right] = n - p \implies \mathbb{E}\left[\hat{\sigma}^2\right] = \sigma^2 \cdot \frac{n-p}{n} < \sigma^2$$

Hence this $\hat{\sigma}^2$ is a biased estimator, but asymptotically unbiased.

6.9 Inference

Definition. Let $U \sim N(0,1)$ and $V \sim \chi^2_n$ be independent random variables. Then

$$T = \frac{U}{\sqrt{\frac{V}{n}}}$$

has a t_n-*distribution*.

As $n \to \infty$, this approaches the standard normal distribution.

Definition. Let $V \sim \chi^2_n$ and $W \sim \chi^2_m$ be independent random variables. Then

$$F = \frac{V/n}{W/m}$$

has an $F_{n,m}$-*distribution*.

Example. We consider a $100(1-\alpha)\%$ confidence interval for one of the coefficients β in the normal linear model $Y = X\beta + \varepsilon$. Without loss of generality, we will consider β_1.

We begin by finding a *pivot*, which is a distribution that does not depend on the parameters of the model. By standardising the above form of $\hat{\beta}$,

$$\frac{\beta_1 - \hat{\beta}_1}{\sqrt{\sigma^2 (X^\intercal X)^{-1}_{11}}} \sim N(0,1)$$

where M^{-1}_{11} is the top left entry in the matrix M^{-1}. This random variable is independent from $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Now, to construct a pivot, we find

$$\frac{\frac{\beta_1 - \hat{\beta}_1}{\sqrt{\sigma^2 (X^\intercal X)^{-1}_{11}}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2} \cdot \frac{n}{n-p}}} \sim \frac{U}{\sqrt{\frac{V}{n}}} \sim t_{n-p}$$

The σ^2 terms cancel, so the statistic is a function only of β_1 and functions of the data. Then,

$$\mathbb{P}_{\beta,\sigma^2}\left(-t_{n-p}\left(\frac{\alpha}{2}\right) \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{(X^\intercal X)^{-1}_{11}}} \sqrt{\frac{n-p}{n\hat{\sigma}^2}} \leq t_{n-p}\left(\frac{\alpha}{2}\right) \right) = 1 - \alpha$$

since the t distribution is symmetric about zero. Rearranging to find an interval for β_1,

$$\mathbb{P}_{\beta,\sigma^2}\left(\hat{\beta}_1 - t_{n-p}\left(\frac{\alpha}{2}\right) \frac{\sqrt{(X^\intercal X)^{-1}_{11}\hat{\sigma}^2}}{\sqrt{(n-p)/n}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-p}\left(\frac{\alpha}{2}\right) \frac{\sqrt{(X^\intercal X)^{-1}_{11}\hat{\sigma}^2}}{\sqrt{(n-p)/n}} \right) = 1 - \alpha$$

Hence,

$$I = \left[\hat{\beta}_1 \pm t_{n-p}\left(\frac{\alpha}{2}\right) \frac{\sqrt{(X^\intercal X)^{-1}_{11}\hat{\sigma}^2}}{\sqrt{(n-p)/n}} \right]$$

is a $100(1-\alpha)\%$ confidence interval for β_1.

Consider a test for $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$. By connecting tests and confidence intervals, we can test H_0 with size α by rejecting this null hypothesis when zero is not contained within the confidence interval I.

Consider a special case where $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where μ, σ^2 are unknown, and we want to infer results about μ. Note that this is a special case of the normal linear model where

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}; \quad \beta = (\mu)$$

So we can infer a confidence interval for μ using the above statistic.

Example. Consider a $100(1-\alpha)\%$ confidence set for β as a whole. Note that

$$\hat{\beta} - \beta \sim N(0, \sigma^2 (X^\intercal X)^{-1})$$

Then,

$$(X^\intercal X)^{1/2}(\hat{\beta} - \beta) \sim N(0, \sigma^2 (X^\intercal X)^{1/2}(X^\intercal X)^{-1}(X^\intercal X)^{1/2}) \sim N(0, \sigma^2 I)$$

where $(X^\intercal X)^{1/2}$ is obtained using the eigendecomposition of the positive definite matrix $X^\intercal X$. Hence,

$$\frac{\left\| (X^\intercal X)^{1/2}(\hat{\beta} - \beta) \right\|^2}{\sigma^2} \sim \chi^2_p$$

We can also write this as

$$\frac{\left\| (X^\intercal X)^{1/2}(\hat{\beta} - \beta) \right\|^2}{\sigma^2} = \frac{\left\| X(\hat{\beta} - \beta) \right\|^2}{\sigma^2}$$

Since this is a function of $\hat{\beta}$, this is independent of any function of $\hat{\sigma}^2$. In particular, it is independent of $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$. Thus, we can form a pivot by

$$\frac{\left\| X(\hat{\beta} - \beta) \right\|^2 / (\sigma^2 p)}{\hat{\sigma}^2 n / (\sigma^2 (n-p))} \sim \frac{\chi^2_p / p}{\chi^2_{n-p}/(n-p)} \sim F_{p,n-p}$$

This does not depend on σ^2. For all β, σ^2,

$$\mathbb{P}_{\beta,\sigma^2}\left(\frac{\left\| X(\hat{\beta} - \beta) \right\|^2 / p}{\hat{\sigma}^2 n/(n-p)} \le F_{p,n-p}(\alpha) \right) = 1 - \alpha$$

because the F distribution has support only on the positive real line. It is nontrivial to express this as a region for β since it is vector-valued. We can say, however, that

$$\left\{ \beta' \in \mathbb{R}^p : \frac{\left\| X(\hat{\beta} - \beta) \right\|^2 / p}{\hat{\sigma}^2 n/(n-p)} \le F_{p,n-p}(\alpha) \right\}$$

is a $100(1-\alpha)\%$ confidence set for β.

This set is an ellipsoid centred at $\hat{\beta}$. The shape of the ellipsoid depends on the design matrix X; the principal axes are given by eigenvectors of $X^\intercal X$.

The above two results are exact; no approximations were made.

6.10 F-tests

We wish to test whether a collection of predictors β_i are equal to zero. Without loss of generality, we will take the first $p_0 \leq p$ predictors. We have $H_0 : \beta_1 = \cdots = \beta_{p_0} = 0$, and $H_1 = \beta \in \mathbb{R}^p$. We denote $X = (X_0, X_1)$ as a block matrix with $X_0 \in \mathbb{R}^{n \times p_0}$ and $X_1 \in \mathbb{R}^{n \times (p-p_0)}$, and we denote $\beta = (\beta^0, \beta^1)^\intercal$ similarly. The null model has $\beta^0 = 0$. This is a linear model $Y = X\beta + \varepsilon = X_1\beta^1 + \varepsilon$. We will write $P = X(X^\intercal X)^{-1}X^\intercal$ and $P_1 = X_1(X_1^\intercal X_1)^{-1}X_1^\intercal$. Note that as X and P have full rank, so must X_1, P_1.

Lemma. $(I - P)(P - P_1) = 0$, and $P - P_1$ is an orthogonal projection with rank p_0.

Proof. $P - P_1$ is symmetric since P and P_1 are symmetric. It is also idempotent, since

$$(P - P_1)(P - P_1) = P^2 - P_1P - PP_1 + P_1^2 = P - P_1 - P_1 + P_1 = P - P_1$$

since P_1 projects onto the column space of X_1. Hence $P - P_1$ is indeed an orthogonal projection matrix. The rank is $\operatorname{rank}(P - P_1) = \operatorname{tr}(P - P_1) = \operatorname{tr} P - \operatorname{tr} P_1 = p - (p - p_0) = p_0$. Also,

$$(I - P)(P - P_1) = P - P_1 - P + PP_1 = P - P_1 - P + P_1 = 0$$

□

Recall that the maximum log-likelihood in the normal linear model is given by

$$\ell(\hat{\beta}, \hat{\sigma}^2) = \frac{-n}{2} \log \hat{\sigma}^2 - \frac{n}{2} \cdot \text{constant} = \frac{-n}{2} \log \frac{\|(I - P)Y\|^2}{n} + \text{constant}$$

The generalised likelihood ratio statistic is

$$\begin{aligned} 2 \log \Lambda &= 2 \sup_{\beta \in \mathbb{R}^p, \sigma^2 > 0} \ell(\beta, \sigma^2) - 2 \sup_{\beta_0 = 0, \beta_1 \in \mathbb{R}^{p-p_0}, \sigma^2 > 0} \ell(\beta, \sigma^2) \\ &= n \left[-\log \frac{\|(I - P)Y\|^2}{n} + \log \frac{\|(I - P_1)Y\|^2}{n} \right] \end{aligned}$$

Wilks' theorem applies here, showing that $2 \log \Lambda \sim \chi^2_{p_0}$ asymptotically as $n \to \infty$ with p, p_0 fixed. However, we can find an exact test, so using Wilks' theorem will not be necessary. $2 \log \Lambda$ is monotone in

$$\begin{aligned} \frac{\|(I - P_1)Y\|^2}{\|(I - P)Y\|^2} &= \frac{\|(I - P + P - P_1)Y\|^2}{\|(I - P)Y\|^2} \\ &= \frac{\|(I - P)Y\|^2 + \|(P - P_1)Y\|^2 + 2Y^\intercal(I - P)(P - P_1)}{\|(I - P)Y\|^2} \\ &= \frac{\|(I - P)Y\|^2 + \|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \\ &= 1 + \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \end{aligned}$$

The generalised likelihood ratio test rejects when the F-statistic

$$F = \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \cdot \frac{1/p_0}{1/(n - p)}$$

is large.

Theorem. Under $H_0 : \beta_1 = \cdots = \beta_{p_0} = 0$, in the normal linear model,

$$F = \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \cdot \frac{1/p_0}{1/(n-p)} \sim F_{p_0, n-p}$$

Proof. Recall that

$$\|(I - P)Y\|^2 = \|(I - P)\varepsilon\|^2 \sim \chi^2_{n-p} \cdot \sigma^2$$

Therefore it suffices to show that $\|(P - P_1)Y\|^2$ is an independent $\chi^2_{p_0} \cdot \sigma^2$ random variable. Under H_0, we have that

$$(P - P_1)Y = (P - P_1)(X\beta + \varepsilon) = (P - P_1)(X_1\beta^1 + \varepsilon) = (P - P_1)\varepsilon$$

since P, P_1 preserve X_1. Hence, $\|(P - P_1)Y\|^2 = \|(P - P_1)\varepsilon\|^2 \sim \chi^2_{\operatorname{rank}(P-P_1)} \cdot \sigma^2 = \chi^2_{p_0} \cdot \sigma^2$. We must now show independence between $(I - P)Y$ and $(P - P_1)Y$. The vectors $(I - P)\varepsilon, (P - P_1)\varepsilon$ are independent; indeed,

$$E = \begin{pmatrix} (I - P)\varepsilon \\ (P - P_1)\varepsilon \end{pmatrix}$$

is a multivariate normal vector, and

$$\mathbb{E}[E] = 0; \quad \operatorname{Var}(E) = \begin{pmatrix} I - P & (I - P)(P - P_1) \\ (I - P)(P - P_1) & P - P_1 \end{pmatrix} = \begin{pmatrix} I - P & 0 \\ 0 & P - P_1 \end{pmatrix}$$

and since $(I - P)\varepsilon$ and $(P - P_1)\varepsilon$ are elements of a multivariate normal vector and are uncorrelated, they are independent as required. □

The generalised likelihood ratio test of size α rejects H_0 when $F > F^{-1}_{p_0, n-p}(\alpha)$. This is an exact test for all n, p, p_0. Previously, we found a test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. This is a special case of the F-test derived above, where $p_0 = 1$. The previous test of size α rejects H_0 when

$$|\hat{\beta}| > t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2 n (X^\intercal X)^{-1}_{11}}{n-p}}$$

We will show that these two tests are equivalent; they reject H_0 in the same critical region. The t-test rejects if and only if

$$\hat{\beta}_1^2 > t_{n-p}\left(\frac{\alpha}{2}\right)^2 \frac{\hat{\sigma}^2 n (X^\intercal X)^{-1}_{11}}{n-p}$$

Note that $t_{n-p}\left(\frac{\alpha}{2}\right)^2 = F_{1,n-p}(\alpha)$, since

$$U \sim N(0,1);\ W \sin \chi^2_n \implies T = \frac{U}{\sqrt{W/n}} \implies T^2 = \frac{U^2}{W/n} = \frac{V/1}{W/n} \sim F_{1,n}$$

where $V \sim \chi^2_1$. Hence,

$$\frac{\hat{\beta}_1 / (X^\intercal X)^{-1}_{11}}{\hat{\sigma}^2 n/(n-p)} > F_{1,n-p}(\alpha)$$

It suffices to show that

$$\frac{\hat{\beta}_1}{(X^\intercal X)^{-1}_{11}} = \frac{\|(P-P_1)Y\|^2}{\underbrace{p_0}_{=1}}; \quad \frac{\hat{\sigma}^2 n}{n-p} = \frac{\|(I-P)Y\|^2}{n-p}$$

We have already shown the latter part. For $\hat{\beta}_1$, note that in this case, $P - P_1$ is a projection of rank 1 onto the one-dimensional subspace spanned by the vector $v = (I-P)X^0$ where X^0 is the first column in the matrix X. First, note the following identity.

$$X_0^\intercal(I-P_1) = v^\intercal = v^\intercal(P-P_1) = X_0^\intercal(I-P_1)(P-P_1) = X_0^\intercal(I-P_1)P$$

Then,

$$\begin{aligned}
\|(P-P_1)Y\|^2 &= \left\| \frac{v}{\|v\|}\left(\frac{v}{\|v\|}\right)^\intercal Y \right\|^2 \\
&= \frac{(v^\intercal Y)^2}{\|v\|^2} = \frac{(X_0^\intercal(I-P_1)Y)^2}{\|(I-P_1)X_0\|^2} \\
&= \frac{(X_0^\intercal(I-P_1)PY)^2}{\|(I-P_1)X_0\|^2} \\
&= \frac{(X_0^\intercal(I-P_1)X\hat{\beta})^2}{\|(I-P_1)X_0\|^2}
\end{aligned}$$

Note that $(I-P_1)X = [(I-P_1)X_0, 0, \dots, 0]$. Hence,

$$\begin{aligned}
\|(P-P_1)Y\|^2 &= \frac{\|(I-P_1)X_0\|^4 \hat{\beta}_1}{\|(I-P_1)X_0\|^2} \\
&= \|(I-P_1)X_0\|^2 \hat{\beta}_1
\end{aligned}$$

Finally, we show that

$$(X^\intercal X)^{-1}_{11} = \frac{1}{\|(I-P_1)X_0\|^2}$$

using the Woodbury identity for blockwise matrix inversion. Hence,

$$\frac{\hat{\beta}_1^2}{(X^\intercal X)^{-1}_{11}} = \|(P-P_1)Y\|^2$$

as required.

6.11 Analysis of variance

Suppose we investigate responses of patients after receiving one of three treatments, including a control, which will be given index 1. We will consider only one predictor, denoting which treatment a given patient received. Consider the linear model

$$Y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$$

where $j = 1, 2, 3$ is the treatment index, and $i = 1, \dots, N$ is the index of a patient in a given group. Let $(\varepsilon_{ij}) \sim N(0, \sigma^2)$ be independent. Without loss of generality, we can set $\mu_1 = 0$, since we have an additional parameter α; this is known as a *corner point* constraint. Then, μ_j should be interpreted as the effect of treatment j relative to treatment 1, which in this case is the control.

Definition. The *analysis of variance (ANOVA)* test on the linear model

$$Y_{ij} = \alpha + \mu_j + \varepsilon_{ij}$$

where $\mu_1 = 0$ is given by

$$H_0 : \mu_2 = \mu_3 = \cdots = 0; \quad H_1 : \mu_2, \mu_3, \cdots \in \mathbb{R}$$

In particular, H_0 gives $\mathbb{E}\left[Y_{ij}\right] = \alpha$.

In our example, $H_0 : \mu_2 = \mu_3 = 0$ and $H_1 : \mu_2, \mu_3 \in \mathbb{R}$. This is a special case of the F-test, since we are testing whether the coefficients μ_i are equal to zero.

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} X_1 & X_0 \end{pmatrix}$$

The first column of X, denoted X_1, represents α, and the other columns, denoted X_0, represent μ_2, μ_3. X_0 is eliminated under the null hypothesis. The predictor can be called *categorical*; it is discrete, and entirely dependent on which treatment category a given patient is placed in. Note that X has $3N$ rows, where each block of N consecutive rows is identical. Recall that the F-test uses the test statistic

$$F = \frac{\|(P - P_1)Y\|^2}{\|(I - P)Y\|^2} \cdot \frac{1/p_0}{1/(n-p)} \sim F_{p_0, n-p}$$

For this test, P projects onto the space of vectors in $\mathbb{R}^{3N}$ which are constant over treatment groups. In other words, let

$$\overline{Y}_j = \frac{1}{N} \sum_{i=1}^{N} Y_{ij}$$

Then,

$$PY = \left(\underbrace{\overline{Y}_1, \ldots, \overline{Y}_1}_{N \text{ entries}}, \underbrace{\overline{Y}_2, \ldots, \overline{Y}_2}_{N \text{ entries}}, \underbrace{\overline{Y}_3, \ldots, \overline{Y}_3}_{N \text{ entries}} \right)^\intercal$$

P_1 projects onto the subspace of constant vectors in $\mathbb{R}^{3N}$, so

$$\overline{Y} = \frac{1}{3N} \sum_{i=1}^{N} \sum_{j=1}^{3} Y_{ij} \implies P_1 Y = \left(\underbrace{\overline{Y}, \ldots, \overline{Y}}_{3N \text{ entries}} \right)^\intercal$$

Hence, we can write the F statistic as

$$F = \frac{\sum_{j=1}^{3} N\left(\overline{Y}_j - \overline{Y}\right)^2 / 2}{\sum_{i=1}^{N} \sum_{j=1}^{3} \left(Y_{ij} - \overline{Y}_j\right)^2 / (3N - 3)}$$

We can generatlise this to the case where there are $J > 3$ treatment groups:

$$F = \frac{\sum_{j=1}^{J} N\big(\overline{Y}_j - \overline{Y}\big)^2/(J-1)}{\sum_{i=1}^{N}\sum_{j=1}^{J}\big(Y_{ij} - \overline{Y}_j\big)^2/(JN-J)} = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

Remark. This test is sometimes called *one-way* analysis of variance. *Two-way* analysis of variance is a similar analysis in an experiment where groups are defined according to two variables. For instance, the response could be a student's performance in an exam, where the treatments are

(i) completion of supervisions (zero representing not complete, one representing complete); and

(ii) whether a monetary incentive was given (zero representing no incentive, one representing an incentive).

Here, we would have the result Y_{ijk} as the number of marks of student i in group (j,k). The model would be

$$Y_{ijk} = \alpha + \mu_j + \lambda_k + \varepsilon_{ijk}$$

with a constraint without loss of generality that $\mu_0 = \lambda_0 = 0$. The two-way analysis of variance test is then

$$H_0 : \mu_1 = \lambda_1 = 0; \quad H_1 : \mu_1, \lambda_1 \in \mathbb{R}$$

6.12 Simple linear regression

In a linear regression model, we often centre predictors to simplify certain expressions.

$$Y_i = \alpha + \beta(x - \overline{x}) + \varepsilon_i$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the ε_i independently have the usual $N(0,\sigma^2)$ distribution. In this case, the maximum likelihood estimator $(\hat{\alpha},\hat{\beta})$ takes a simple form. Recall that $(\hat{\alpha},\hat{\beta})$ minimises

$$S(\alpha,\beta) = \sum_{i=1}^{n}\big(Y_i - \alpha - \beta(x_i - \overline{x})\big)^2$$

Hence,

$$\frac{\partial S(\alpha,\beta)}{\partial \alpha} = \sum_{i=1}^{n} -2\big(Y_i - \alpha - \beta(x_i - \overline{x})\big) = \sum_{i=1}^{n} -2(Y_i - \alpha)$$

This gives the simple expression

$$\alpha = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$$

Now,

$$\frac{\partial S(\alpha,\beta)}{\partial \beta}\bigg|_{\alpha=\hat{\alpha}} = \sum_{i=1}^{n} -2\big(Y_i - \overline{Y} - \beta(x_i - \overline{x})\big)(x_i - \overline{x})$$

This vanishes when

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Note that $\frac{S_{xy}}{n}$ is the sample covariance of X and Y, and $\frac{S_{xx}}{n}$ is the sample variance of X.