Statistics

Hasan Baig

Lent 2022

Contents

0 Overview **3**
0.1 Parametric Inference 3

1 Review of Probability **3**
1.1 Maxima of Random Variables 5
1.2 Linear Transformations 5
1.3 Standardised Statistics 6
1.4 Moment Generating Functions 6
1.5 Limits of Random Variables 7
1.6 Conditioning 7
1.7 Change of Variables 8
1.8 Important Distributions 8

2 Estimation **9**
2.1 Bias-variance Decomposition 10
2.2 Sufficiency 12
2.2.1 Minimal sufficiency 13
2.3 Rao-Blackwell Theorem 15
2.4 Maximum Likelihood Estimation 17
2.5 Confidence Intervals 20
2.6 Bayesian Analysis 23
2.7 Point estimation 26

3 Hypothesis Testing **27**
3.1 Neyman-Pearson Lemma 28
3.2 Composite Hypothesis 31
3.3 Generalised Likelihood test 33
3.4 Wilk's Theorem 33
3.5 Goodness-of-fit Test 34
3.6 Pearson statistic 35
3.7 Goodness-of-Fit Test for Composite Null 36
3.8 Testing Independence in Contingency Tables 36
3.9 Problems With χ^2 Test of Independence 37
3.10 Testing Homogeneity 38
3.11 Relationship Between Tests and Confidence Sets 39
3.12 Multivariate Normal Distribution 40
3.13 Orthogonal Projections 41

4 Linear Models **45**
4.1 Matrix Formulation 45
4.2 Fitted Values and Residuals 47

4.3 Normal Linear Model 48
4.4 Inference in Normal Linear Model 48
4.5 Confidence Sets for β 51
4.6 F-test 52

0 Overview

Statistics is the science of making informed decisions. It can include:

- The design of experiments and studies
- Data visualisation
- Formal statistical inference
- Communication of uncertainty and risk
- Formal decision theory

In this course, we focus on formal statistical inference

0.1 Parametric Inference

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

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

Notation. We'll use the observed $X = (X_1, \ldots, X_n)$ to make inferences about θ:

(i) Point estimate $\hat{\theta}(X)$ of θ (hat usually denotes estimator)

(ii) Interval estimate of θ: $(\hat{\theta}_1(x), \hat{\theta}_2(x))$

(iii) Testing hypotheses about θ e.g. $H_0 : \theta = 1$. Testing is checking whether there is evidence in X against H_0

Remark. In general, we will assume that the distribution family of $X_1, \ldots, X_n$ is known and the parameter is unknown. However, some results (on m.s.e., bias, Gauss-Markov theorem) will make weaker assumptions.

1 Review of Probability

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

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

- $\mathbb{P}(\varnothing) = 0$
- $\mathbb{P}(\Omega) = 1$
- $$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_i \mathbb{P}(A_i)$$
  if $(A_i)_i$ is a sequence of disjoint events

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

Example. Tossing 2 coins: $\Omega = \{HH, HT, TH, TT\}$. $\mathcal{F}$ is the power set of Ω. We can let X be the number of heads.

$$X(HH) = 2 \quad X(HT) = X(TH) = 1 \quad X(TT) = 0$$

Definition. The **distribution function** of X is

$$F_X(x) = \mathbb{P}(X \leq x)$$

Definition. A **discrete** r.v. takes values in a countable set $\mathcal{X} \subset \mathbb{R}$

Definition. Its **probability mass function** is

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

We say that X has a continuous distribution if it has a **probability distribution function** p.d.f. $f_X(x)$ which satisfies:

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

for "nice" sets A

Definition. The **expectation** of X is

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

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

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

Definition. The **variance** of X is

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

Definition. We say $X_1, \ldots, X_n$ are **independent** if for all $x_1, \ldots, x_n$,

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

If $X_1, \ldots, X_n$ have pdfs (or pmfs) $f_{X_1}, \ldots, f_{X_n}$, the joint pdf (pmf) is

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

Note. Converse true

1.1 Maxima of Random Variables

Equation. If $Y = \max\{X_1, \ldots, X_n\}$ (indep), then

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

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

1.2 Linear Transformations

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

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

We let

$$\mathbb{E}X = \begin{bmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_n \end{bmatrix}$$

Remark. We do not require $X_1, \ldots, X_n$ to be independent

Equation.

$$\begin{aligned} \mathrm{Var}(a^T X) &= \sum_{i,j} a_i a_j \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i,j} \mathbb{E}[(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)] \\ &= a^T \mathrm{Var}(X) a \end{aligned}$$

where

$$(\mathrm{Var}(X))_{ij} = \mathrm{Cov}(X_i, X_j)$$

This is known as the "bilinearity of variance"

1.3 Standardised Statistics

Notation. Let $X_1, \ldots, X_n$ be iid r.v.s, $\mathbb{E}X_1 = \mu,\ \text{Var}(X_1) = \sigma^2$

$$S_n = \sum_i X_i, \quad \overline{X}_n = \frac{S_n}{n}$$

$\overline{X}_n$ is the **sample mean**. By linearity

$$\mathbb{E}\overline{X}_n = \mu \quad \text{Var}\overline{X}_n = \frac{\sigma^2}{n}$$

Define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \sqrt{n}\frac{X_n - \mu}{\sigma}$$

$$\mathbb{E}Z_n = 0 \quad \text{Var}Z_n = 1$$

1.4 Moment Generating Functions

Definition. The **mgf** of a r.v. X is

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

This is the Laplace transform of the pdf provided that it exists for t in some neighbourhood of 0. Relationship with moments:

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

Remarks.

- Under broad conditions $M_X = M_Y \iff F_X = F_Y$
- Moment generating functions are useful for finding the distribution of sums of indepented random variables

Example. Let $X_1, \ldots, X_n \sim \text{Poisson}(\mu)$

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

What is M_{S_n}?

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

Therefore, $S_n \sim \text{Poisson}(\mu)$

1.5 Limits of Random Variables

Theorem (Weak law of large numbers (WLLN))**.**

$$\forall \varepsilon > 0 \ \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \to 0 \text{ as } n \to \infty$$

we note our event depends only on $X_1, \ldots, X_n$

Theorem (Strong law of large numbers (SLLN))**.**

$$\mathbb{P}(\overline{X}_n \to \mu) = 1 \text{ as } n \to \infty$$

we note our event depends on the whole sequence

$$\overline{X}_n \to \mu \iff \forall \varepsilon > 0 \ \exists N \text{ s.t. } |\overline{X}_n - \mu| < \varepsilon \text{ if } n \geq N$$

Theorem (Central limit theorem)**.** $Z_n = (S_n - n\mu)/(\sigma\sqrt{n})$ is approximately $N(0,1)$ when n is large

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

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

1.6 Conditioning

Definition. If X, Y are discrete random variables

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

when the denominator is non-zero.

Definition. If X, Y are continuous, the **joint p.d.f.** of X, Y, $f_{X,Y}(x, y)$ satisfies:

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

The **conditional p.d.f.** of X given Y is

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

note that we can denote the denominator a $f_Y(y)$

Definition.

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

note that $\mathbb{E}[X|Y]$ is a function of Y so is itself a random variable. We define $\mathrm{Var}(X|Y)$ similarly.

Equation (Tower property).

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

Theorem (Law of total variance).

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

1.7 Change of Variables

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

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

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

1.8 Important Distributions

Examples.

- $X \sim \mathrm{Bin}(n, p)$: number of successes in n independent Bernoulli(p) trials
- $X \sim \mathrm{Multi}(n; p_1, \ldots, p_k)$: n independent trials, k types, p_j is the probability of type j in each trial. Note X takes values in $\mathbb{N}^k$. We let X_j be the number of trials with type j
- $X \sim \mathrm{Neg}(k, p)$: In iid Ber($p$) trials, X is the time where kth success occurs

$$\mathrm{Neg}(1, p) = \mathrm{Geometric}(p)$$

- $X \sim \mathrm{Poi}(\lambda)$: Limit of $\mathrm{Bin}(n, \lambda/n)$ as $n \to \infty$

Equation. If $X_i \sim \Gamma(\alpha_i, \lambda)$ for $i = 1, \ldots, n$ with $X_1, \ldots, X_n$ indep. What is the distribution of $S_n = X_1 + \cdots + X_n$?

$$M_{S_n}(t) = \prod_i M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\sum_i \alpha_i} \quad \text{for } t < \lambda$$

or ∞ if $t \geq \lambda$. Therefore, $S_n \sim \Gamma(\sum_i \alpha_i, \lambda)$. The first parameter is the "shape parameter". The second parameter is the rate parameter.
If $X \sim \Gamma(\alpha, \lambda)$, then $\forall b > 0$ $bX \sim \Gamma(\alpha, \lambda/b)$

Examples. Special cases:

- $\Gamma(1, \lambda) = \mathrm{Exp}(\lambda)$
- $\Gamma(k/2, 1/2) = \chi^2_k$ is the Chi-squared distribution with k degrees of freedom. This is the distribution of the sum of k independent squared $N(0, 1)$ random variables

2 Estimation

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

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

Example. $X_1, \ldots, X_n \sim N(\mu, 1)$

$$\hat{\mu} = T(X) = \frac{1}{n} \sum_i X_i = \overline{X}_n$$

The sampling distribution of $\hat{\mu}$ is $T(X) \sim N(\mu, \frac{1}{n})$

Definition. The **bias** of $\hat{\theta} = T(X)$

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

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

Remark. In general, the bias is a function of the true parameter θ, even though it is not explicit in notation "bias($\hat{\theta}$)"

Definition. We say $\hat{\theta}$ is **unbiased** if $\text{bias}(\hat{\theta}) = 0$ for all values of true parameter θ

Example (continued)**.** $\hat{\mu}$ is unbiased because

$$\mathbb{E}_\mu[\hat{\mu}] = \mathbb{E}_\mu[\overline{X}_n] \quad \forall \mu \in \mathbb{R}$$

Definition. The **mean squared error** (mse) of θ

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

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

Warning. The $\text{mse}(\hat{\theta})$ is a function of θ!

2.1 Bias-variance Decomposition

Equation.

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

There is a tradeoff between bias and variance

Example. $X \sim \text{Binomial}(n, \theta)$. Suppose n known, $\theta \in [0,1]$ is unknown parameter.

$$T_u = \frac{X}{n}$$

is the 'proportion of successes observed'. This is unbiased as $\mathbb{E}_\theta(T_u) = \mathbb{E}_\theta(X)/n = n\theta/n = \theta$. Therefore,

$$\begin{aligned}\text{mse}(T_u) &= \text{Var}_\theta(T_u) \\ &= \text{Var}_\theta\left(\frac{X}{n}\right) \\ &= \frac{\text{Var}_\theta}{n^2} \\ &= \theta(1-\theta)\end{aligned}$$

Consider another estimator

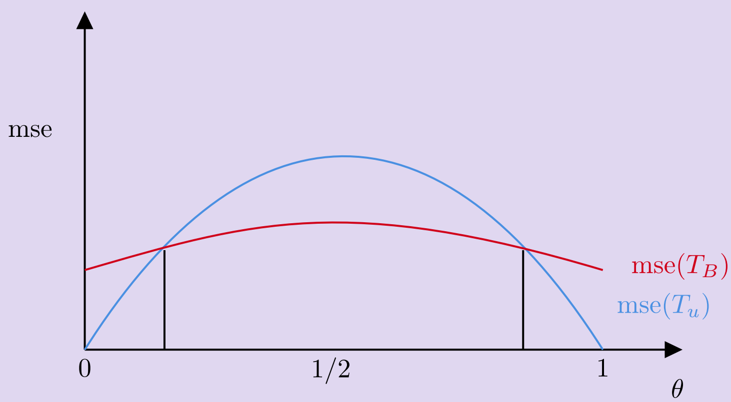
$$T_B = \frac{X+1}{n+1} = w\frac{X}{n} + (1-w)\frac{1}{2} \quad w := \frac{n}{n+2}$$

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

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

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

$$\text{mse}(T_B) = (1-w)^2(\frac{1}{2}-\theta)^2 + w^2\frac{\theta(1-\theta)}{n}$$



T_B is "better" than T_u

Remark. Prior judgement on true value of θ determines which estimator is better

Note. Unbiasedness is not necessarily desirable

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

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

The only function $T : \mathbb{N} \to \mathbb{R}$ satisfying this equality is

$$T(X) = (-1)^X$$

This makes no sense.

2.2 Sufficiency

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

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

Example. $X_1, \dots, X_n \sim \text{Bernoulli}(\theta)$ iid for some parameter $\theta \in [0, 1]$

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

Note: this only depends on x through $T(x) = \sum x_i$

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

If $\sum x_i = t$,

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

This does not depend on θ, hence $T(X)$ is sufficient

Theorem (Factorisation Criterion)**.** A statistic T is sufficient for θ iff $f_X(x|\theta) = g(T(x), \theta)h(x)$ for suitable functions g, h

Proof. We only prove in the discrete case. Suppose $f_X(x|\theta) = g(T(x), \theta)h(x)$. Then if $T(x) = t$:

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

does not depend on θ; hence $T(X)$ sufficient.
Conversely, suppose that $T(X)$ is sufficient

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

Note the first term does not depend on θ as T sufficient. The second term only depends on X through $T(x)$

Example. $X_1, \ldots, X_n \sim \text{Ber}(\theta)$ iid

$$f_X(x|\theta) = \underbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}_{g(T(x),\theta)} \cdot \underbrace{1}_{h(x)}$$

Let $T(X) = \sum X_i$. Then $T(X)$ is sufficient

Example. Let $X_1, \ldots, X_n \sim \text{Unif}([0\theta])$ for some $\theta > 0$. Then

$$\begin{aligned} f_X(x|\theta) &= \prod_{i=1}^n \frac{1}{\theta} 1_{\{x_i \in [0,\theta]\}} \\ &= \left(\frac{1}{\theta}\right) 1_{\{\max_i X_i \geq 0} 1_{\{\max X_i \leq \theta\}} \end{aligned}$$

$$g(T(x), \theta) = \left(\frac{1}{\theta}\right) 1_{\{\max_i X_i \geq 0}, \quad h(x) = 1_{\{\max X_i \leq \theta\}}$$

Therefore, $T(X)$ is sufficient

2.2.1 Minimal sufficiency

Note. Sufficient statistics are not unique

Remark. Any 1-to-1 function applied to a sufficient statistic yields another sufficient statistic. $T(X) = X$ is a trivial sufficient statistic. We want statistics which give us "maximal" copression of information in X

Definition. A sufficient statistic $T(X)$ is called **minimal** if it is a function of every other sufficient statistic. I.e. if T' is also sufficient, then

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

Remark. If S, T minimal sufficient, then they are in bijection, i.e.

$$T(x) = T(y) \iff S(x) = S(y)$$

Minimal sufficient statistics are unique "up to bijections"

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

Proof. For any value t of T let z_t be a representative from $\{x : T(x) = t\}$. Then

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

Call the first term $g(T(x), \theta)$ and second term does not depend on θ by hypothesis, call this $h(x)$. Then T is sufficient by factorisation criterion.
To prove T is minimal sufficient, let S be any other sufficient statistic. By factorisation criterion, $\exists$ functions g_S, h_S s.t.

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

Now suppose $S(x) = S(y)$ then

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

which is constant in θ, so $x \sim_1 y$. By hypothesis, $x \sim_2 y$ and $T(x) = T(y)$

Let $x \sim_1 y$ if $f_X(x|\theta)/f_Y(y|\theta)$ is constant in θ. It's easy to check that $\sim_1$ is an equivalence relation. Similarly, let $x \sim_2 y$ if $T(x) = T(y)$ also an equivalence relation. Hypothesis in theorem says equivalence classes of $\sim_1$, $\sim_2$ are the same

Note. We can always construct a statistic T which is constant on the equivalence classes of $\sim_1$. Hence, by the theorem a minial sufficient statistic exists

Example. Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

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

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

$$S(x) = (\overline{X}_n, S_{xx})$$

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

Note. In previous example, $\theta = (\mu, \sigma^2)$ has same dimension as $S(X)$. In general, they can differ

Example. Consider $X_1, \dots, X_n \sim N(\mu, \mu^2)$, $\mu \in \mathbb{R}$. In this case $S(X) = \overline{X}_n, S_{xx}$ is minimal sufficient

2.3 Rao-Blackwell Theorem

Notation. Up to now, we have used $\mathbb{E}_\theta, \mathbb{P}_\theta$ to denote expectations & probabilities under model $X_1, \dots, X_n$ are iid from $f_X(x|\theta)$. From now, we omit the subscript θ

Theorem. Let T be a sufficient statistic for θ and define an estimator $\tilde{\theta}$ with $\mathbb{E}[\tilde{\theta}^2] < \infty$ for all θ. Define a new estimator

$$\hat{\theta} = \mathbb{E}[\tilde{\theta}|T(X)]$$

Then for all $\theta \in \Theta$

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

The inequality is strict unless $\tilde{\theta}$ is a function of $T(x)$

Proof. By tower property

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\mathbb{E}[\tilde{\theta}|T]] = \mathbb{E}\tilde{\theta}$$

So $\text{bias}(\hat{\theta}) = \text{bias}(\tilde{\theta})$ for all $\theta \in \Theta$. By conditional variance formula

$$\text{Var}(\tilde{\theta}) = \underbrace{\mathbb{E}[\text{Var}(\tilde{\theta}|T)]}_{\geq 0} + \underbrace{\text{Var}(\mathbb{E}[\tilde{\theta}|T])}_{\text{Var}(\hat{\theta})}$$

So $\text{Var}(\tilde{\theta}) \geq \text{Var}(\hat{\theta})$, and by bias-variance decomposition

$$\text{mse}(\tilde{\theta}) \geq \text{mse}(\hat{\theta})$$

The inequality is strict unless $\text{Var}(\tilde{\theta}|T) = 0$ with probability 1, which would require $\tilde{\theta}$ is a function of T

Moral. Start from any estimator $\tilde{\theta}$ and by conditioning on sufficient statistic, we get a better one

Remark. As $T(X)$ is sufficient, $\hat{\theta}$ is a bona fide estimator of θ (i.e. it is a function of X but not of θ), because

$$\hat{\theta}(X) = \hat{\theta}(T) = \int \tilde{\theta}(x) f_{X|T}(x|T)\,\mathrm{d}x$$

Example. $X_1, \dots, X_n \sim \text{Poi}(\lambda)$. Let $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$

$$f_X(x|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_i x_i!}$$

$$\implies f_X(x|\theta = \frac{\theta^n(-\log\theta)^{\sum x_i}}{\prod_i x_i!})$$

$\therefore \sum x_i = T(x)$ is sufficient by factorisation. Recall $\sum x_i \sim \text{Poi}(\lambda)$. Let $\tilde{\theta} = 1_{\{X_1=0}$ (only depends on X_1). It's weak but unbiased

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

So $\hat{\theta} = (1 - 1/n)^{\sum x_i}$ is an estimator with $\text{mse}(\hat{\theta}) < \text{mse}(\hat{\theta})$ for all θ.
Sanity check: $\hat{\theta} = (1 - 1/n)^{n\overline{X}_n} \to e^{-\overline{X}_n}$ as $n \to \infty$ and by SLLN $\overline{X}_n \to \mathbb{E}X_1 = \lambda$ w.p. 1 so $\hat{\theta} \approx e^{-\lambda} = \theta$ when n is large

Example. Let $X_1, \dots, X_n$ be iid $\text{Unif}([0, \theta])$, want to estimate $\theta > 0$. We have seen previously that $T = \max_i X_i$ is sufficient.
Let $\tilde{\theta} = 2X_1$, an unbiased estimator of θ. Then,

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

By Rao-Blackwell $\text{mse}(\hat{\theta}) \leq \text{mse}(\tilde{\theta})$. Also, $\hat{\theta}$ is unbiased

2.4 Maximum Likelihood Estimation

Notation. Let $X_1, \dots, X_n$ iid with pdf (or pmf) $f_X(\cdot|\theta)$

Definition. The **likelihood** function $L : \Theta \to \mathbb{R}$ is given by

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

(we take x to be fixed observations)

Notation. We'll denote the log-likelihood

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

Definition. A **maximum likelihood estimator** (mle) is one that maximises L over Θ (or l)

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

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

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

This is equal to 0 $\iff$ $p = \sum X_i/n = \overline{X}_n$. We have $\mathbb{E}\hat{p} = \frac{n}{n}\mathbb{E}X_1 = p$. So the mle $\hat{p} = \overline{X}_n$ is unbiased

Example. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$

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

Maximised when $\frac{\partial l}{\partial \mu} = \frac{\partial l}{\partial \sigma^2} = 0$

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)$$

$\implies$ equal to 0 iff $\mu = \overline{X}_n = \frac{1}{n}\sum X_i$, for all $\sigma^2 > 0$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \mu)^2$$

If we set $\mu = \overline{X}_n$, $\frac{\partial l}{\partial \sigma^2}$ is 0 iff

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

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

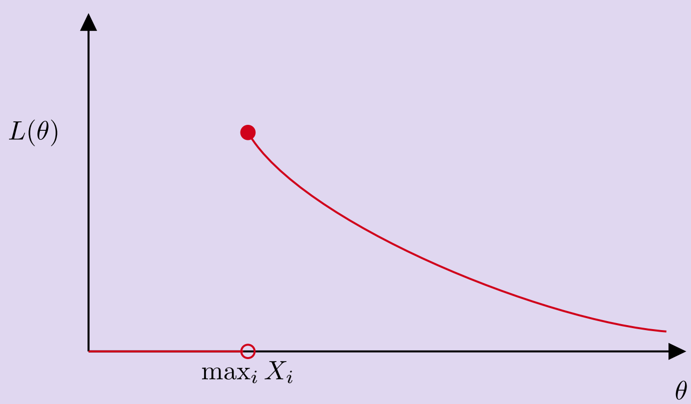
$$\frac{S_{xx}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

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

Hence $\hat{\sigma}^2$ is biased. But as $n \to \infty$, the bias converges to 0, so we say $\hat{\sigma}^2$ is "asymptotically unbiased"

Example. $X_1, \dots, X_n \sim \text{Unif}([0, \theta])$. Recall, we derived estimator $\hat{\theta} = \frac{n+1}{n}\max_i X_i$
What is the mle?

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



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

$$\mathbb{E}\hat{\theta}^{mle} = \frac{n}{n+1}\mathbb{E}\hat{\theta} = \frac{n}{n+1}\theta$$

Properties of the mle:

(i) If T is a sufficient statistic for θ, then mle is a function of T. Recall,

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

So the maximiser of L only depends on X through T

(ii) If $\phi = H(\theta)$ where H is a bijection and $\hat{\theta}$ is mle for θ, then $H(\hat{\theta})$ is the mle for ϕ

(iii) Asymptotic normality: under regularity conditions, as $n \to \infty$ the statistic $\sqrt{n}(\hat{\theta} - \theta)$ is approx $N(0, \Sigma)$, i.e. for some "nice" set A

$$\mathbb{P}(\sqrt{n}(\hat{\theta} - \theta) \in A) \to \mathbb{P}(z \in A)$$

where $z \sim N(0, \Sigma)$. The limiting covariance matrix Σ is a known function of l. In some sense, it is the "best" or "smallest" variance that any estimator can achieve asymptotically
(We prove this in Part II Principles of Statistics)

(iv) When the mle is not available analytically in closed form, it can be found numerically in many cases

2.5 Confidence Intervals

Definition. A $100 \cdot \gamma\%$ **confidence interval** (with $0 < \gamma < 1$) for a parameter θ is a random interval $(A(X), B(X))$ such that

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

A, B are random, θ is fixed.

We have a frequentist interpretation: if we repeat the experiment many times, on average $100 \cdot \gamma\%$ of the time, $(A(X), B(X))$ will contain θ

Warning. Misleading interpretation: Having observed $X = x$, there is now a probability γ that $\theta \in (A(x), B(x))$

Example. $X_1, \dots, X_n \sim N(\theta, 1)$. Find 95% C.I. for θ. We know

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\theta, \frac{1}{n})$$

and

$$Z = \sqrt{n}(\overline{X} - \theta) \sim N(0,1) \text{ for all } \theta \in \mathbb{R}$$

Let a, b be numbers s.t. $\Phi(b) - \Phi(a) = 0.95$
Then $\mathbb{P}(a \leq \sqrt{n}(\overline{X} - \theta) \leq b) = 0.95$. Rearrange:

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

Hence $(\overline{X} - b/\sqrt{n}, \overline{X} - a/\sqrt{n})$ isa 95% C.I. for θ.
Typically, we center the interval around some estimator $\hat{\theta}$ and aim to minimise its length. In this case, we want

$$-a = b = z_{0.025}$$

where z_α is equal to $\Phi^{-1}(1-\alpha)$ or the "upper α-point" of $N(0,1)$ distribution.
So C.I. is $(\overline{X} \pm 1.96/\sqrt{n})$

Method. Finding a C.I.:

(i) Find a quantity $R(X, \theta)$ whose $\mathbb{P}_\theta$-distribution does not depend on θ. This is called a pivot. e.g. $R(X,\theta) = \sqrt{n}(\overline{X} - \theta)$

(ii) Write down

$$\mathbb{P}(x_1 \leq R(X,\theta) \leq c_2) = \gamma$$

Given some γ, we find c_1, c_2 using the distribution function of $R(X, \theta)$

(iii) Rearrange to leave θ in the middle of two inequalities

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

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

Example. $X_1, \dots, X_n \sim N(0, \sigma^2)$ iid. Find a 95% C.I. for σ^2

(i) Note $X_1/\sigma \sim N(0,1)$.

$$\sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \chi_n^2()$$

(ii) Let

$$c_1 = F_{\chi_n^2}^{-1}(0.025), \quad c_2 = F_{\chi_n^2}^{-1}(0.975)$$

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

diagram

(iii)

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

(iv) Hence $\left(\frac{\sum x_i^2}{c_2}, \ \frac{\sum x_i^2}{c_1} \right)$ is a 95% C.I, for σ

Example. $X_1, \dots, X_n \sim \mathrm{Ber}(p)$ with n "large". Find approximate 95% C.I. for p

(i) The mle of p is $\hat{p} = \overline{X} = \frac{1}{n} \sum_i X_i$ By CLT, $\hat{p}$ is approx $N(p, \frac{p(1-p)}{n})$. Thus $\sqrt{n}(\hat{p}-p)/\sqrt{p(1-p)}$ is approx. $N(0,1)$

(ii)

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

(iii) Instead of directly rearranging the inequalities, we will approximate $\sqrt{p(1-p)} \approx \sqrt{\hat{p}(1-\hat{p})}$. And we argue that when n is large

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

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

Hence $\left(\hat{p} \pm z_{0.025} frac \sqrt{\hat{p}(1-\hat{p})} \sqrt{n} \right)$ is an approximate 95% C.I. for p

Remark. $p(1-p) \leq 1/4$ on $p \in (0,1)$ hence $\hat{p} \pm z_{0.025}/2\sqrt{n}$ is a "conservative" 95% C.I. for o

Moral. Interpreting C.I's: suppose X_1, X_2 are iid Unif$(\theta - 1/2, \theta + 1/2)$. What is a senseible 50% C.I. for θ? Note

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

Hence $(\min(X_1, X_2), \max(X_1, X_2)$ is a 50% C.I. for θ. The frequentist interpretaiton is exactly correct. But suppose $|X_1 - X_2| > 0.5$ then we know that θ is in $(\min(X_1, X_2), \max(X_1, X_2))$
The frequentist interpretation of the 50% C.I. is entirely correct. But it is not sensible tosay that having seen a particular X_1, X_2 (e.g. $X_1 = 0.1, X_2 = 0.9$) we are "50% certain that θ is in the C.I"

2.6 Bayesian Analysis

Remark. So far, we have talked about frequentist inference where we think of θ as fixed. Inferential statements interpreted in terms of repetitions of the experiment. Bayesian analysis is a different framework.
Bayesians treat θ as a r.v. taking values in Θ. The **prior distribution** $\pi(\theta)$ represents the investigator's beliefs or information about θ before observing data. Conditional on θ, the data X has pdf (or pmf) $f_X(\cdot|\theta)$
Having observed X, the information in X is combined with the prior to form the **posterior distribution** denoted $\pi(\theta|X)$, which is conditional distribution of θ given X. By Bayes' rule:

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

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

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

More simply,

$$\underbrace{\pi(\theta|X)}_{\text{post}} \propto \underbrace{\pi(\theta)}_{\text{prior}} \times \underbrace{f_X(X|\theta)}_{\text{likelihood}}$$

Often, it is easy to recognise that RHS is in some family of distributions up to normalising constant

Note. By factorisation criterion, if T is sufficien, then

$$\begin{aligned}\pi(\theta|X) &\propto \pi(\theta) \times g(T(X), \theta) \times h(X) \\ &\propto \pi(\theta) \times g(T(X), \theta)\end{aligned}$$

$\therefore$ posterior only depends on X through $T(X)$

Example (prior choice is clear). Patient walks into covid testing clinic (no information about them)

$$\theta = \begin{cases} 1 & \text{if patient infected} \\ 0 & \text{otherwise} \end{cases}$$

We observe $X = 1_{\{\text{positive covid text}}$. We know sensitivity of the test:

$$f_X(X = 1|\theta = 1)$$

and specificity of the test:

$$f_X(X = 0|\theta = 0)$$

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

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

Sometimes $\pi(\theta = 1) << \pi(\theta = 0)$ which can make $\pi(\theta = 1|X = 1)$ small (surprising!)

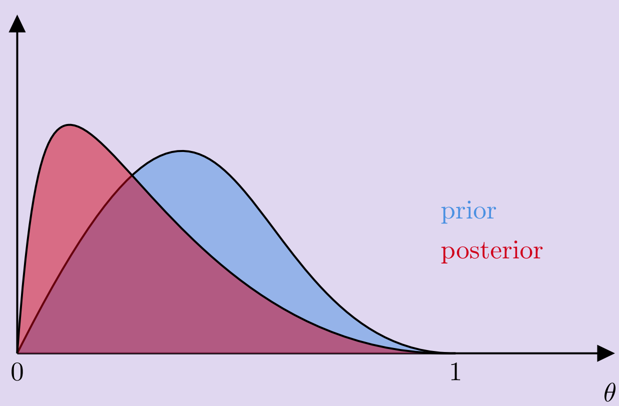
Example. θ taking values in $[0,1]$ is mortality rate for new surgery at Addenbrookes.
Data: in the first 10 operations, no deaths.
Model: $X \sim$ Binomial $(10, \theta)$, $X = 0$
Prior: in other hospitals, mortality ranges between 3% and 20%, with average of 10% e.g. take $\pi(\theta)$ is Beta(a, b)
Choose $a = 3, b = 27$ os that $\pi(\theta)$ has mean 0.1 and

$$\pi(0.03 < \theta < 0.2) \approx 0.9$$

Posterior:

$$\begin{aligned}\pi(\theta|X) &\propto \pi(\theta) \times f_X(X = 0|\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1} \times \theta^X(1-\theta)^{n-X} \\ &= \theta^{X+a-1}(1-\theta)^{b+n-X-1}\end{aligned}$$

for $\theta \in [0,1]$. We recognise this as a Beta$(X + a, n - X + b)$. In our example, Beta$(3, 10 + 27)$



Note. In the above example, prior and posterior are in the same family. This is known as conjugacy

Moral. What to do with posterior?
$\pi(\theta|X)$ represents info about θ after seeing X. This can eb used to make decisions under uncertainty

Method.
(i) We must pick some decision $\delta \in \Delta$
e.g. In first example, $\Delta = \{$ask patient to isolate, do not ask patient to isolate$\}$
(ii) Define loss function $L(\theta, \delta)$
e.g. $L(\theta = 1, \delta = 1)$ would be the loss incurred by asking patient to isolate if positive
(iii) Pick δ that minimises

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

in English, this is the "posterior expectation of loss" (see Von-Neumann-Morgenstern)

2.7 Point estimation

An example of a decision is a "best guess" for θ. The Bayes estimator $\hat{\theta}^(b)$ minimises

$$h(\delta) = \int_\Theta L(\theta, \delta)\pi(0|X)\,\mathrm{d}\theta$$

Example. Quadratic loss $L(\theta, \delta) = (\theta - \delta)^2$

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

$$h'(\delta) = 0 \text{ if } \int_\Theta (\theta - \delta)\pi(\theta|X)\,\mathrm{d}\theta = 0$$

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

This is $\hat{\theta}^{(b)}$ consider quadratic loss (posterior mean).

Example. Absolute error loss $L(\theta, \delta) = |\theta - \delta|$

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

Take derivative w.r.t. δ (invoke F.T.C.)

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

So $h'(\delta) = 0$ iff

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

hence $\hat{\theta}^{(b)}$ is median of posterior $\pi(\theta|X)$

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

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

Note. Unlike confidence intervals, credible intervals can be interpreted conditionally, i.e. "given a specific observation x, we are 95% certain that θ is in (a, b)"
Caveat: credible interval depends on choice of prior

Example. $X_1, \dots, X_n \sim N(\mu, 1)$
Prior: $\pi(\mu)$ is $N(0, \tau^{-2})$ with known τ^2

$$\pi(\mu|x) \propto f_X(x|\mu) \times \pi(\mu)$$
$$\propto \exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\} \times \exp\{-\mu\frac{\tau^2}{2}\}$$
$$\propto \exp\{-\frac{1}{2}(n + (\tau)^2)[\mu - \frac{\sum x_i}{n + \tau^2}]^2\}$$

$\implies$ posterior is $N(\frac{\sum x_i}{n+\tau^2}, \frac{1}{n+\tau^2})$
Bayes estimator under quadratic and mean absolute error loss is $\frac{\sum x_i}{n+\tau^2}$ (contrast this with mle $\hat{\mu}^{(mle)} = \frac{\sum x_i}{n}$)
Posterior variance decreases as $\frac{1}{n+\tau^2} \approx \frac{1}{n}$

How do credible intervals compare to confidence intervals?

Example. $X_1, \dots, X_n \sim \text{Poi}(\lambda)$
priot: $\pi(\lambda)$ is $\text{Exp}(1)$

$$\pi(\lambda|x) \propto f_X(x|\lambda) \times \pi(\lambda)$$
$$\propto \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_i x_i!} \times e^{-\lambda}, \quad \lambda > 0$$
$$\propto e^{-(n+1)\lambda}\lambda^{\sum x_i}$$

$\implies$ posterior is $\text{Gamma}(\sum x_i + 1, n + 1)$
Bayes estimator under quadratic loss is the posterior mean:

$$\hat{\lambda}^{(b)} = \frac{\sum x_i + 1}{n + 1}$$

3 Hypothesis Testing

Definition. A **hypothesis** is an assumption about the distribution of data X.
Scientific questions are often phrased as a decision between a **null hypothesis** H_0 and **alternative hypothesis** H_1

Example. (i) $X = (X_1, \ldots, X_n)$ are iid Bernoulli(θ)

$$H_0 : \theta = \frac{1}{2}, \quad H_1 : \theta = \frac{3}{4}$$

(ii)

$$H_0 : \theta = \frac{1}{2}, \quad H_1 : \theta \neq \frac{1}{2}$$

(iii) $X = (X_1, \ldots, X_n)$, x_i takes values in $\mathbb{N}_0$

$$H_0 X_i \sim \text{ Poi}(\lambda) \text{ for some } \lambda > 0$$

$$H_1 : X_i \sim f_1 \text{ for some other distribution } f_1$$

"Goodness-of-fit" test

Definition. A **simple hypothesis** is one which fully specifies the (pdf or pmf) of X.
Othewise, we say the hypothesis is **composite**
A test of the null H_0 is defined by a critical region $C \subset \chi$ when $X \in C$, we "reject the null". When $X \notin C$, we say we "fail to reject H_0" or "find no sufficient evidence against H_0"

Definition. Two types of error:

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

When H_0, H_1 are simple, define

$$\alpha = \mathbb{P}_{H_0}(H_0 \text{ is rejected}) = \mathbb{P}_{H_0}(X \in C)$$

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

The **size** of test is α, the **power** is $1 - \beta$

Note. What we typically do is choose an acceptable probability of type I errors (say 1%); set α to that, pick the test which minimises β (maximises power)

3.1 Neyman-Pearson Lemma

Definition. Let H_0 and H_1 be simple, with X having pdf (or pmf) f_i under H_i, $i = 0, 1$.
The **likelihood ratio statistic** is:

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

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

$$C = \{x : \Lambda_x(H_0; H_1) > k\}$$

for some k

Theorem. Suppose that f_0, f_1 are nonzero on some sets. Suppose there is $k > 0$ s.t. the LRT with critical region

$$C = \{x : \Lambda_x(H_0; H_1) > k\}$$

has size α. Then out of all tests with size $\leq \alpha$, this test has smallest β (largest power)

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

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

Let C^* be some other critical region with type I/ type II error probabilities α^*, β^*

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

Suppose $\alpha^* \leq \alpha$: want to prove $\beta \leq \beta^* \iff \beta - \beta^* \leq 0$

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

Notice we can cancel over $\overline{C} \cap \overline{C^*}$

$$\begin{aligned}
\beta - \beta^* &= \int_{\overline{C} \cap C^*} f_1(x)\,\mathrm{d}x - \int_{\overline{C^*} \cap C} f_1(x)\,\mathrm{d}x \\
&= \int_{\overline{C} \cap C^*} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\leq k} f_0(x)\,\mathrm{d}x - \int_{\overline{C^*} \cap C} \underbrace{\frac{f_1(x)}{f_0(x)}}_{> k} f_0(x)\,\mathrm{d}x \\
&\leq l \left[\int_{\overline{C} \cap C^*} f_0(x)\,\mathrm{d}x - \int_{\overline{C^*} \cap C} f_0(x)\,\mathrm{d}x\right] \\
&\leq l \left[\int_{C^*} f_0(x)\,\mathrm{d}z - \int_{C} f_0(x)\,\mathrm{d}x\right] \\
&\leq k\,[\alpha^* - \alpha] \leq 0
\end{aligned}$$

Remark. A LRT of size α does not always exist. Exercise: think of a (model,H_0, H_1, α)
But in general, we can find a "randomised test of size α"

Example. $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$, where σ^2 is known. Want the best size α test for

$$H_0 : \mu = \mu_0, \quad H_1 : \mu = \mu_1$$

for some fixed $\mu_1 > \mu_0$

$$\Lambda_X(H_0; H_1) = \frac{(2\pi\sigma_0)^{1/2} \exp\{-\frac{1}{2\sigma_0^2} \sum (x_i - \mu_1)^2\}}{(2\pi\sigma_0)^{1/2} \exp\{-\frac{1}{2\sigma_0^2} \sum (x_i - \mu_0)^2\}}$$
$$= \exp\{\frac{(\mu_1 - \mu_0)}{\sigma_0^2} n\overline{X} + n\frac{\mu_0^2 - \mu_1^2}{2\sigma_0^2}$$

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

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

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

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

i.e. the test $C = \{x : Z(x) > \Phi^{-1}(1 - \alpha)$. This is called a z-test

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

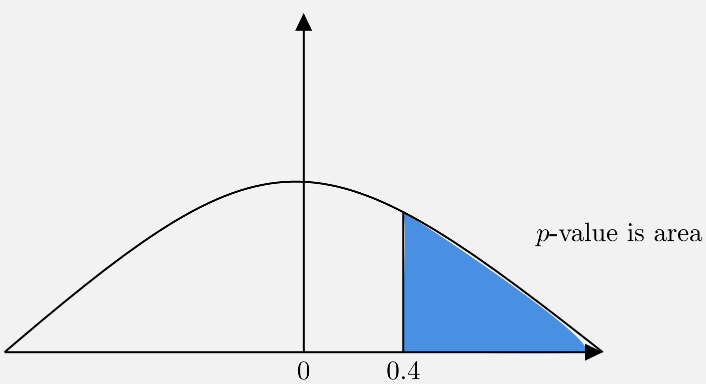
$$p = \mathbb{P}_{H_0}(T(X) > T(x^*))$$

where x^* is the observed data.

In the example above, suppose $\mu_0 = 5, \mu_1 = 6, \alpha = 0.005$
Data: $x^* = (5.1,\ 5.5,\ 4.9,\ 5.3)$

$$\overline{X^*} = 5.2,\ Z^* = 0.4$$

LRT is $\{x : Z(x) > \Phi^{-1}(0.95) = 1.645\}$.
Conclusion of LRT: we do not reject H_0



Prop. Under H_0, p-value is Unif$[0, 1]$

Proof. Let F be the distribution of T (which we assume to be continuous)

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

$$1 - F(F^{-1}(1 - u)) = u$$

3.2 Composite Hypothesis

$X \sim f_X(\cdot|\theta);\ \theta \in \Theta$

$$H_0 : \theta \in \Theta_0 \subset \Theta$$
$$H_1 : \theta \in \Theta_1 \subseteq \Theta$$

Now, the probabilities of type I or type II error may depend on the value within Θ_0 (or Θ_1) - not single numbers

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

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

The **size** of a test C is

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

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

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

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

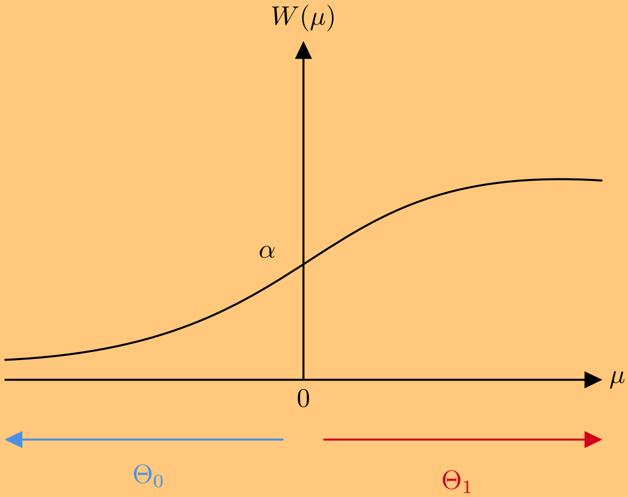
Example. One-sided test for normal location $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$, σ_0 is known

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

for some fixed μ_0 (e.g. $\mu_0 = 0$)

Claim. LRT for $H_0' : \mu = \mu_1,\ H_1' : \mu = \mu_1 > \mu_0$ derived earlier is UMP in the compound case. The power function is

$$\begin{aligned} W(\mu) &= \mathbb{P}_\mu(\text{reject } H_0) = \mathbb{P}_\mu(\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma_0} > z_\alpha) \\ &= \mathbb{P}_\mu(\sqrt{n}\frac{\bar{X}-\mu}{\sigma_0} > z_\alpha) + \sqrt{n}\frac{(\mu_0-\mu_1)}{\sigma_0}) \\ &= 1 - \Phi(z_\alpha + \sqrt{n}\frac{(\mu_0-\mu_1)}{\sigma_0}) \end{aligned}$$



Note: test has size α as $\sup_{\mu \in \Theta_0} W(\mu) = \alpha$

Proof. Indeed (i) is satisfied

$$\sup_{\mu \le \mu_0} W(\mu) = \alpha$$

Need to check that for any test C^* of size α, with power W^*

$$W(\mu) \ge W^*(\mu) \text{ for all } \mu > \mu_0$$

Note: Critical region C only depends on μ_0, not μ_1.
Take any $\mu_1 > \mu_0$ then C is LRT for $H_0' : \mu = \mu_0$ vs $H_1' : \mu = \mu_1$.
We can also see that C^* as a test of H_0' vs H_1'. And for these simple hypotheses C^* has size:

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

So by N-P lemma, C has power no smaller than C^* for H_0' vs H_1', i.e.

$$W(\mu_1) \ge W^*(\mu_1)$$

3.3 Generalised Likelihood test

Definition.

$$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$$

with $\Theta_0 \subset \Theta_1$, hypotheses are "nested"
The **GLR** is given by

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

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

Example. Two sided test for normal location $X_1, \dots, X_n \sim N(\mu, \sigma_0^2); \sigma_0^2$ known

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

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

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

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

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

3.4 Wilk's Theorem

The dimension of a hypothesis $H_0 : \theta \in \Theta_0$ is the number of "free parameters" in Θ_0 e.g.

(i) $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_1 = \dots = \theta_p = 0\}$ then $\dim(\Theta_0) = k - p$

(ii) Let $A \in \mathbb{R}^{p \times k}$ with linearly indep. rows $b \in \mathbb{R}^p$, $p < k$

$$\Theta_0 = \{\theta \in \mathbb{R}^k : A\theta = b\}$$

$$\dim \Theta_0 = k - p$$

(iii) Θ_0 is a Riemannian manifold

Theorem. Suppose $\Theta_0 \subset \Theta_1$ and $\dim(\Theta_1) - \dim(\Theta_0) = p$. Then if $X = (X_1, \dots, X_n)$ are iid under $f_X(\cdot|\theta)$ with $\theta \in \text{int}(\Theta_0)$, then [under some conditions] as $n \to \infty$, limiting distribution of $2 \log \Lambda_X$ is χ_p^2 i.e.

$$\mathbb{P}_\theta(2 \log \Lambda_X \leq l) \to \mathbb{P}(\Xi \leq l) \quad \forall l \in \mathbb{R}_+$$

where $\Xi \sim \chi_p^2$

Remark. This is very useful because it allows us to implement a GLR test even if we cant find the exact distribution of $2 \log \Lambda_X$ (assuming that n is large; any frequentist guarantee will be approximate)

Example. In 2 sided normal location example

$$\dim \Theta_0 = 0, \quad \dim \Theta_1 = 1$$

So theorem tells us $2 \log \Lambda_X$ is approximately χ_1^2 (in this example, this happens to be exact)

3.5 Goodness-of-fit Test

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

$$\sum_i N_i = n, \quad \sum_i p_i = 1$$

Parameters: $(p_1, \dots, p_k) := p$ parameter space has dimension $k-1$, because of constraint $\sum p_i = 1$
A G-o-F test has a null of form:

$$H_0 : p_i = \tilde{p}_i \quad i = 1, \dots, k$$

for some fixed distribution $\tilde{p}$. The alternative puts no constraints on p.
The model is $(N_1, \dots, N_k) \sim \text{Multinomial}(n; p_1, \dots, p_k)$

$$L(p) \propto p_1^{N_1} \dots p_k^{N_k}$$

$$l(p) = \log L(p) = \text{ const } + \sum +iN + i \log p_i$$

The GLR Λ_X has

$$2 \log \Lambda_X = 2 \left(\underbrace{\sup_{p \in \Theta_1} l(p)}_{l(\hat{p})} - \underbrace{\sup_{p \in \Theta_0} l(p)}_{l(\tilde{p})} \right)$$

To find $\hat{p}$ we use Lagrange multipliers

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

$\implies \hat{p}_i = N_i/n$ "fraction of samples equal to i"
After some computation, we get $\hat{p}_i = N_i/n$, so

$$2 \log \Delta_x = 2 \sum N_i \log \left(\frac{N_i}{n - \tilde{p}_i} \right)$$

Wilk's theorem tells us that when n is large, $2 \log \Delta_x$ is approximately χ_p^2

$$p = \dim(\Theta_1) - \dim(\Theta_0) = (k-1) - 0 = k-1$$

An approximate GLR test of size α rejectes when

$$N \in C = \{N_i 2 \sum N_i \log \left(\frac{N_i}{n - \tilde{p}_i} \right) \geq \chi_{k-1}^2(\alpha)$$

Let $o_i = N_i$ "observed number of type i"; $e_i = n\tilde{p}_i$ "expectation number null of nuber of type i"

$$2 \log \Lambda = 2 \sum_i o_i \log \left(\frac{o_i}{e_i} \right)$$

3.6 Pearson statistic

$$\delta_i = o_i - e_i$$

$$\begin{aligned} 2 \log \Lambda &= 2 \sum_i (e_i + \delta_i) \log \left(1 + \frac{\delta_i}{e_i} \right) \\ &\approx 2 \sum_i \left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i} \right) \\ &= \sum \frac{\delta_i^2}{e_i} = \sum_i \frac{(o_i - e_i)^2}{e_i} \end{aligned}$$

This is called Pearson's χ^2 statistic. It is also referred to a χ^2_{k-1} when we test H_0

Example. Mendel's experiment
Mendel crossed peas to obtain a sample of 556 descendents; each descentent is one of 4 types: SG, SY, WG, WY.
He observed $N = (315, 108, 102, 31)$.
Mendel's theory gives a null hypothesis

$$H_0 : p = \tilde{p} = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16} \right)$$

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

These are referred to a χ^2_3 distribution

$$\chi^2_3(0.05) = 7.05$$

so a test of size 5% does not reject H_0.
The p-value is $\mathbb{P}(\chi^2_4 > 0.6) \approx 0.96$

3.7 Goodness-of-Fit Test for Composite Null

$$H_0 : p_i = p_i(\theta) \text{ for some } \theta \in \Theta, \quad \forall i = 1, \dots, k$$
$$H_1 : p \text{ is any distribution on } \{1, \dots, k\}$$

$$2 \log \Lambda = 2(\sup_p l(p) - \sup_{\theta \in \Theta_0} l(p(\theta)))$$

We can sometimes compute $2 \log \Lambda$, and find a test which refers this test statistic to χ^2_p

$$p = \dim \Theta_1 - \dim \Theta_0 = (k-1) - \dim \Theta_0$$

Example.

$$p+1 = \theta^2, \quad p_2 = 2\theta(1-\theta), \quad p_3 = (1-\theta)^2$$

θ is the overall abundance of one type of gene.
In this example, we can find MLE $\hat{\theta}$ under null

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

So

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

where $\hat{p}_i = N_i/n$ can be computed and referred to a χ^2_2

Remark. We can check that in this model

$$2 \log \Lambda = \sum_i o_i \log \left(\frac{o_i}{e_i} \right) \approx \sum_i \frac{(o_i - e_i)^2}{e_i}$$

where $o_i = N_i$ "observed counts" and $e_i = n \cdot p_i(\hat{\theta})$ "expected counts under null"

3.8 Testing Independence in Contingency Tables

$(X_1, Y_1), \dots, (X_n, Y_n)$ are iid where X_i take values in $\{1, \dots, r\}$, Y_i take values in $\{1, \dots, c\}$
We wish to test whether X_i independent of Y_i
We shall summarise the data into a contingency table N

$$N_{ij} = \#\{l : 1 \leq l \leq n, (X_l, Y_l) = (i, j)\}$$

"number of samples of type (i, j)"

Example. Covid-19 death
Q: Have deaths decreased more rapidly for vaccinated groups?
Probability model: we observe n samples, each sample has probability p_{ij} of being of type (i, j)

$$(N_{ij})_{i,j} \sim \text{Multinomial}(n; (p_{ij})_{ij})$$

Null hypothesis:

$$H_0 : p_{ij} = p_{i+} \cdot p_{+j}$$

where $p_{i+} = \sum_{ij}, \; p_{+j} = \sum_i p_{ij}$.
Alternative: $H_1 : (p_{ij})_{1\leq i\leq r, 1\leq j\leq c}$ is any non-negative vector with $\sum_{i,j} p_{ij} = 1$.
As usual, we find $2 \log \Lambda$

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

where:

- $\hat{p}_{ij}$ is MLE under H_1
- $\hat{p}_{i+}, \hat{p}_{+j}$ is MLE under H_0

All of these MLEs can be found with Langranian method. We have

$$\hat{p}_{ij} = \frac{N_{ij}}{n}, \quad \hat{p}_{i+} = \frac{N_{i+}}{n}, \quad \hat{p}_{+j} = \frac{N_{+j}}{n}$$

writing $o_{ij} = N_{ij}, \; e_{ij} = n \cdot \hat{p}_{i+}p_{+j}$

$$\begin{aligned} 2 \log \Lambda &= \sum_{i,j} \log \left(\frac{o_{ij}}{e_{ij}} \right) \\ &\approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \end{aligned}$$

By Wilk's theorem, these test statistics have approximate χ^2_p

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

3.9 Problems With χ^2 Test of Independence

(i) χ^2 approximation requires n to be large.
Rule of Thumb: $N_{ij} \geq 5$ for all i, j
Solution: exact tests

(ii) Low power.
Why? The alternative H_1 is too large.
Solution: define a more specific H_1, lump categories

Remark. This test also applies when n is random with a Poisson

3.10 Testing Homogeneity

Example. 150 patients are randomly assigned to 3 groups of equal size. Two sets get a new drug with different doeses. Third set gets placebo.

	Improved	No Difference	Worse	
Placebo	18	17	15	50
Half-dose	20	10	20	50
Full-dose	25	13	12	50

Probability model: $N_{i1}, \dots, N)ic \sim$ Multinomial$(n_{i+}; p_{i1}, \dots, p_{ic})$ independently for$i = 1, \dots, r$
Null H_0: $p_{1j} = p_{2j} = \dots = p_{rj}$ $\forall j = 1, \dots, c$
Alternative H_1: $p_{i1}, \dots, p_{ic}$ is any probability vector for each row $i = 1, \dots, r$.
Under H_1:

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

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

To find the mle we use Lagranian method with $\sum_j p_{ij} = 1$ for each $1, \dots, r$

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

Under H_0: let $p_j = p_{ij}$

$$\begin{aligned} l(p) &= \text{const } + \sum_{i,j} N_{i,j} \log p_{ij} \\ &= \text{const } + \sum_{j} N_{+j} \log p_j \end{aligned}$$

Using Lagranian method with $\sum_j p_j = 1$

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

Hence

$$\begin{aligned} 2 \log \Lambda &= 2 \sum_{i,j} N_{ij} \log \left(\frac{\hat{p}_{ij}}{\hat{p}_j} \right) \\ &= 2 \sum_{i,j} N_{ij} \log \left(\frac{N_{ij}}{n_{i+} N_{+j} / n_{++}} \right) \end{aligned}$$

Same statistic as for χ^2 test for independence!
Furthermore if $o_{ij} = N_{ij}$ and $e_{ij} = n_{i+} \cdot \hat{p}_j = \frac{n_{i+} N_{+j}}{n_{++}}$, we have

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

By Wilk's theorem $2 \log \Lambda \sim \chi^2_p$ approx.

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

So limiting distribution of $2 \log \Lambda$ is $\chi^2_{(r-1) \times (c-1)}$ same as independence test!

Moral. Operationally χ^2 tests for independence and Homogeneity are identical

Example (continued).

$$2\log\Lambda = 5.129$$

$$\sum_{i,j} \frac{(o_{ij} - e_{ij})}{e_{ij}} = 5.173$$

we refer these to a $\chi^2_{(3-1)\times(3-1)} = \chi^2_4$

$$\chi^2_4(0.05) = 9.488\dots$$

Hence we do not reject H_0 with size 5%

3.11 Relationship Between Tests and Confidence Sets

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

Notation. Let $X \sim f_X(\cdot|\theta)$ for some $\theta \in \Theta$

Theorem. (i) Suppose for each $\theta_0 \in \Theta$ there is a test of size α with acceptance region $A(\theta_0)$ for the null $H_0: \theta = \theta_0$. Then

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

is a $100(1-\alpha)$ confidence set

(ii) Suppose $I(X)$ is a $100(1-\alpha)$ confidence set for θ. Then $A(\theta_0) : \{x : \theta_0 \in I(x)\}$ is the acceptence region of a size α test

Proof. Observe that for both (i) and (ii)

$$\underbrace{\theta_0 \in I(x)}_{A} \iff X \in A(\theta_0) \iff \underbrace{\text{“accept” } H_0 : \theta = \theta_0 \text{ in a test with data } X}_{B}$$

(i) Assume $\mathbb{P}_\theta(B) = 1 - \alpha$. Want to prove $\mathbb{P}_\theta(A) = 1 - \alpha$

(ii) Assume $\mathbb{P}_\theta(A) = 1 - \alpha$. Want to prove $\mathbb{P}_\theta(B) = 1 - \alpha$

Example. $X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$; σ_0^2 known. We found a $100(1-\alpha)\%$ C.I. for μ is

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

Using part (ii) of theorem we can find a test for $H_0 : \mu = \mu_0$ of size α

$$\begin{aligned} A(\mu_0) &= \{x : I(x) \ni \mu_0\} \\ &= \{x : \mu_0 \in [\overline{X} \pm \frac{Z_{\alpha/2}\sigma_0}{\sqrt{n}}] \end{aligned}$$

This is equivalent to rejecting H_0 when

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

This is what we call 2-sided test for a normal location

3.12 Multivariate Normal Distribution

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

$$\mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_n)^T, \quad \text{Var}(X) = (\mathbb{E}((X_1 - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)))_{i,j}$$

Linearity of expectation gave us:
Let $A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k$ be constant

$$\mathbb{E}(AX + b) = A\mathbb{E}X + b$$

$$\text{Var}(AX + b) = A\text{Var}(X)A^T$$

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

Prop. If X is MVN then $AX + b$ is MVN

Proof. Take any $t \in \mathbb{R}^k$, then

$$t^T(AX + b) = (A^Tt)^TX + t^Tb$$

This is $N(\mu + t^Tb, \sigma^2)$ where (μ, σ^2) are the mean and variance of $(A^Tt)^TX$

Prop. A MVN is fully specified by its mean and covariance

Proof. Let X_1, X_2 be MVN, both with mean μ and variance Σ. We'll show they have the same MGF, hence the same distribution

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

This is only a function of μ, Σ. A similar argument yields some MGF for X_2

3.13 Orthogonal Projections

Definition. We say $P \in \mathbb{R}^{n \times n}$ is an **orthogonal projection** onto $\text{col}(P)$ if for all $v \in \text{col}(P)^\perp$, $Pw = 0$

Prop. P is an orthogonal projection if and only if
- Symmetry: $P = P^T$
- Idempotency: $PP = P$

Proof. $\Longleftarrow$: Take $v \in \text{col}(P),\ v = Pa$ for some $a \in \mathbb{R}^n$
Then

$$Pv = PPa = Pa = v$$

Take $w \in \text{col}(P)^\perp$, by definition $P^T w = 0$ so

$$Pw = P^T w = 0$$

$\Longrightarrow$: We can write any $a \in \mathbb{R}^n$ uniquely as $a = v + w$ where $v \in \text{col}(P),\ w \in \text{col}(P)^\perp$. Then

$$P^2 a = PP(v + w) = Pv = P(v + w) = Pa$$

Since this holds for all a, $P^2 = P$. For symmetry, take $u_1, u_2 \in \mathbb{R}^n$, note

$$(Pu_1)^T((I - P)u_2) = 0$$

Since this holds for all $u_1,\ u_2$, we have

$$u_1^T(P^T(I - P))u_2 = 0$$

$$\Longrightarrow P^T(I - P) = 0$$

$$\Longrightarrow P^T - P^T P = 0 \Longrightarrow P^T = P^T P$$

Hence P^T (and P) are symmetric

Corollary. If P is orthogonal projection, so is $(I - P)$

Proof. If P is symmetric, so is $I - P$. Also

$$(I - P)(I - P) = I - 2P + P^2 = I - P$$

Prop. If P is an orthogonal projection, then

$$P = UU^T$$

where columns of U are an orthonormal basis for $\text{col}(P)$

Proof. Check that UU^T is projection. It is clearly symmetric, and

$$UU^T UU^T = UU^T$$

Furthermore, by definition, $\text{col}(P) = \text{col}(UU^T)$

Prop. $\text{rank}(P) = \text{Tr}(P)$

Proof. $\text{rank}(P) = \text{Tr}(U^T U) = \text{Tr}(UU^T) = \text{Tr}(P)$

Theorem. If X is MVN, $X \sim N(0, \sigma^2 I)$ and P is an orthogonal projection, then

- $PX \sim N(0, \sigma^2 P),\ (I-P)X \sim N(0, \sigma^2(I-P))$ are independent
- $$\frac{\|PX\|^2}{\sigma^2} = \chi^2_{\text{rank}(P)}$$

Proof. The vector $\begin{bmatrix} P \\ I-P \end{bmatrix} X$ is MVN as it is a linear function of X as it is a linear function of X. Its distribution is fully specified by the mean and variance:

$$\mathbb{E}\begin{bmatrix} PX \\ (I-P)X \end{bmatrix} = \begin{bmatrix} P \\ I-P \end{bmatrix} \mathbb{E}X = 0$$

$$\begin{aligned} \text{Var}\begin{bmatrix} PX & (I-P)X \end{bmatrix} = \begin{bmatrix} P \\ I-P \end{bmatrix} \sigma^2 I \begin{bmatrix} P & I-P \end{bmatrix} &= \sigma^2 \begin{bmatrix} P & P(I-P) \\ P(I-P) & I-P \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} P & 0 \\ 0 & I-P \end{bmatrix} \end{aligned}$$

Let $Z \sim N(0, \sigma^2 P),\ Z' \sim N(0, \sigma^2(I-P))$ independent. Then we can see that

$$\begin{bmatrix} Z \\ Z' \end{bmatrix} \sim N(0, \sigma^2 \begin{bmatrix} P & 0 \\ 0 & I-P \end{bmatrix})$$

Hence $\begin{bmatrix} PX \\ (I-P)X \end{bmatrix} = \begin{bmatrix} Z \\ Z' \end{bmatrix}$ hence $PX \perp\!\!\!\perp (I-P)X$.
For (ii) note that

$$\begin{aligned} \frac{\|PX\|^2}{\sigma^2} &= \frac{X^T P^T P X}{\sigma^2} \\ &= \frac{X^T (UU^T)^T (UU^T) X}{\sigma^2} \\ &= \frac{\|U^T X\|^2}{\sigma^2} \end{aligned}$$

where cols of U are orthonormal basis of col(P). But $U^T X \sim N(0, \sigma^2 U^T U) = N(0, \sigma^2 U_{\text{rank}(P)})$ so

$$\frac{(U^T X)_i}{\sigma} \sim N(0,1) \text{ iid for } i = 1, \dots, \text{ rank}(P)$$

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

Example. $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ for some unknown $\mu \in \mathbb{R},\ \sigma^2 > 0$. Recall that the mles are

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

Theorem. (i) $\overline{X} \sim N(\mu, \sigma^2/n)$
(ii) $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$
(iii) $\overline{X},\ S_{XX}$ are independent

Proof. Let

$$P = \begin{bmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

Its easy to check P is symmetric and idempotent, hence a projection matrix.

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

We'll write

$$X = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} + \varepsilon \text{ where } \varepsilon \sim N(0, \sigma^2 I)$$

Note:

- $\overline{X}$ is a function of $P\varepsilon$

$$\overline{X} = (PX)_1 = (P \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} + P\varepsilon)_1$$

-

$$\begin{aligned} S_{XX} &= \sum_i (X_i - \overline{X})^2 \\ &= \|X - \begin{bmatrix} \overline{X} \\ \vdots \\ \overline{X} \end{bmatrix}\|^2 \\ &= \|(I-P)X\|^2 \\ &= \|(I-P)\varepsilon\|^2 \end{aligned}$$

Hence S_{XX} is a function of $(I-P)\varepsilon$. Therefore, $\overline{X}$ and S_{XX} are independent

Remark. Noting that $I - P$ is a projection with

$$\text{rank}(I-P) = \text{Tr}(I-P) = n-1$$

we can apply the previous theorem to obtain

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

4 Linear Models

Data $(x_1, Y_1), \dots, x_n Y_n)$ wher $Y_i \in \mathbb{R},\ x_i \in \mathbb{R}^p$.
Y_i: response or dependent variable
$x_{i1}, \dots, x_{ip}$: predictors or independent random variables.
Goal: model $\mathbb{E}Y_i$ as a function of $(x_{i1}, \dots, x_{ip})$
We assume

$$Y_i = \alpha + \beta_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- α intercept
- $\beta \in \mathbb{R}^p$: coefficients
- ε_i is a random variable, "the noise"

$\alpha,\ \beta$ are the parameters of interest

Remarks.

(i) We will eliminate the intercept by making $x_{i1} = 1$ for all i, so β_1 plays the role of the intercept
(ii) A linear model can also model non-linear relationships
e.g. $Y_i = a + bz_i + cz_i^2 + \varepsilon_i$. We can rephrase this as a linear model with $x_i = (1,\ z_i,\ z_i^2)$
(iii) β_j can be interpreted as the effect on Y_i of increasing x_{ij} by 1, while keeping $x_{i1}, \dots, x_{i,j-1}, x_{i,j+1}, \dots, x_{ip}$ fixed. This effect cannot be interpreted causally, unless this is a randomised control experiment.

4.1 Matrix Formulation

Equation.

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} x_{i1} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

Y is random, $X\beta$ is fixed, and ε is random

Moment assumptions:
(i) $\mathbb{E}\varepsilon = 0 \implies \mathbb{E}Y_i = x_i^T \beta$
(ii) $\text{Var}\varepsilon = \sigma^2 I \iff$
- $\text{Var}\varepsilon_i = \sigma^2$ "homokedasticity"
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$

Initially, we won't assume anything else about the distribution of ε.
We will always assume that X has full rank p. Since $X \in \mathbb{R}^{n \times p}$, this requires $n \geq p$ (we need at least as many samples as predictors)

Method. Least squares estimation: The least squares estimator $\hat{\beta}$ minimises the residual sum of squares

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

This is a P.D quadratic polynomial in β, so it is minimised at point where

$$\left.\frac{\partial S(\beta)}{\partial \beta_k}\right|_{\beta=\hat{\beta}} = 0 \text{ for all } k = 1, \dots, p$$

$$\implies -2\sum_{i=2}^{n} x_{ik}(Y_i - \sum_j x_{ij}\hat{\beta}_j) = 0 \quad \forall k = 1, \dots, p$$

$$\implies X^T X \hat{\beta} = X^T Y$$

as X has full rank, $X^T X$ is invertible

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

Note.

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

  $\therefore \hat{\beta}$ is unbiased

- $$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T \text{Var}(Y)[(X^T X)^{-1} X^T]^T \\ &= \sigma^2 (X^T X^{-1}) X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

Theorem (Gauss-Markov)**.** Let $\beta^* = CY$ be any other linear estimator, which is unbiased, $\mathbb{E}\beta^* = \beta$ $(\forall \beta)$. Then for any fixed $t \in \mathbb{R}^p$

$$\text{Var}(t^T\hat{\beta}) \leq \text{Var}(t^T\beta^*)$$

We say $\hat{\beta}$ is the Best Linear Unbiased Estimator

Proof. Want to prove:

$$\text{Var}(t^T\beta^*) - \text{Var}(t^T\hat{\beta}) = t^T(\text{Var}\beta^* - \text{Var}\hat{\beta})t \geq 0 \quad \forall t \in \mathbb{R}^p$$

$\iff$ $\text{Var}(\beta^*) - \text{Var}(\hat{\beta})$ is P.S.D.
Let $A = C - (X^TX)^{-1}X^T$.
Note $\forall \beta$

$$\mathbb{E}AY = \mathbb{E}\beta^* - \mathbb{E}\hat{\beta} = 0$$

$$\mathbb{E}AY = A\mathbb{E}Y = AX\beta = 0$$

Thus

$$AX = 0$$

Now

$$\begin{aligned}\text{Var}(\beta^*) = \text{Var}((A + (X^TX)^{-1}X^T)Y) &= (A + (X^TX)^{-1}X^T)\underbrace{\text{Var}Y}_{\sigma^2 I}[A + (X^TX)^{-1}X^T]^T \\ &= \sigma^2(AA^T + (X^TX)^{-1} + AX(X^TX)^{-1} + (X^TX)^{-1}X^TA^T) \\ &= \sigma^2 AA^T + \text{Var}(\hat{\beta})\end{aligned}$$

$$\implies \text{Var}(\beta^*) - \text{Var}(\hat{\beta}) = \sigma^2 AA^T$$

which is P.S.D

Remark. Think of $t \in \mathbb{R}^p$ as vector of predictors for a ew sample. Then $t^T\hat{\beta}$ is a prediction for $\mathbb{E}Y_i$ for this new sample, when we use $\hat{\beta}$, and $t^T\beta^*$ is prediction with β^*.
Note $t^T\hat{\beta}, t^T\beta^*$ are both unbiased

4.2 Fitted Values and Residuals

Definition. **Fitted values** are

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

Residuals are

$$Y - \hat{Y} = (I - P)Y$$

Prop. P is orthogonal projection onto col(X)

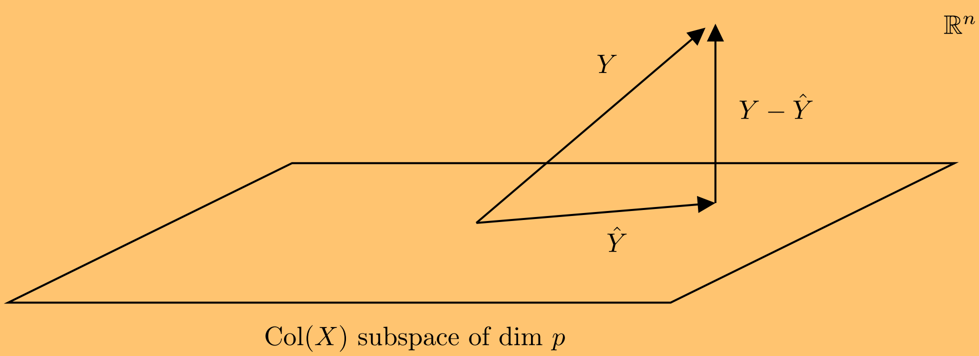
Proof. If $v \in \text{col}(X)$, i.e. $v = Xb$

$$Pv = X(X^TX)^{-1}X^TXb = Xb = v$$

If $w \in \text{col}(X)^\perp$

$$Pw = X(X^TX^{-1})X^Tw = 0$$

Corollary. $\hat{Y} = PY$ is orthogonal projection of Y onto col(X), and residuals $Y - \hat{Y} = (I-P)Y$ is a perpendicular vector



4.3 Normal Linear Model

From now on, we will assume

$$\varepsilon \sim N(0, \sigma^2 I)$$

parameters in model are (β, σ^2)
Likelihood:

$$\begin{aligned} L(\beta, \sigma^2) &= f_Y(y|\beta, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - x_i^T\beta)^2\right\} \end{aligned}$$

4.4 Inference in Normal Linear Model

MLE for σ^2?
Take

$$\frac{\partial l(\hat{\beta, \sigma^2})}{\partial \sigma^2} = 0$$

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

Theorem. (i) $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$
(ii) $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$
(iii) $\hat{\beta}, \hat{\sigma}^2$ are independent

Proof. For (i), we already know $\mathbb{E}\hat{\beta} = \beta$ and $\text{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. So enough to show that $\hat{\beta}$ is MVN. Have

$$\hat{\beta} = (X^TX)^{-1}X^TY \text{ where } Y \sim N(X\beta, \sigma^2 I)$$

hence $\hat{\beta}$ is MVN.
For (ii)

$$\begin{aligned}\frac{n\hat{\sigma}^2}{\sigma^2} &= \frac{\|(I-P)Y\|^2}{\sigma^2} = \frac{\|(I-P)(X\beta+\varepsilon)\|^2)}{\sigma^2} \\ &= \frac{\|(I-P)\varepsilon\|^2}{\sigma^2} \sim \chi^2_{\text{Tr}(I-P)} \text{ as } (I-P)X = 0\end{aligned}$$

where $\text{Tr}(I-P) = n - \text{Tr}(P) = n - p$ since $X \in \mathbb{R}^{n\times p}$ has rank p
For (iii) observe that $\hat{\sigma}^2$ is a function of $(I-P)\varepsilon$, and also

$$\begin{aligned}\hat{\beta} = (X^TX)^{-1}X^TY &= (X^TX)^{-1}X^T(X\beta+\varepsilon) \\ &= \beta + \underbrace{(X^TX)^{-1}X^T\varepsilon}_{=(X^TX)^{-1}X^TP\varepsilon}\end{aligned}$$

so $\hat{\beta}$ is a function of $P\varepsilon$. But by Thm 1, $P\varepsilon \perp (I-P)\varepsilon$, hence $\hat{\beta} \perp \hat{\sigma}^2$

Equation.

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

$$\implies \mathbb{E}\hat{\sigma}^2 = \sigma^2\frac{n-p}{n} < \sigma^2$$

So $\hat{\sigma}^2$ is a biased estimator. It is asymptotically unbiased if p is fixed as $n \to \infty$.

Example (Student-t distribution)**.** Let $U \sim N(0,1)$, $V \sim \chi^2_n$, $U \perp V$. Then we say $T = U/\sqrt{V/n}$ has a t_n distribution

Examples (F distribution)**.** If $V \sim \chi^2_n$, $W \sim \chi^2_m$, $V \perp W$ then we say that $F = \frac{V/n}{W/m}$ has an $F_{n,m}$ distribution.

Method. Confidence interval for β_1: We'd like to find a $100 \cdot (1-\alpha)\%$ for one of the coefficients in β, WLOG take β_1.
Note:

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

taking matrix inverse first, then index. We construct a pivot

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

Then

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

Rearrange to obtain:

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

Hence

$$I = [\beta_1 \pm t_{n-p}(\alpha/2)\frac{\sqrt{(X^T X)^{-1}_{11}\sigma^2}}{\sqrt{(n-p)/n}}]$$

is a $100 \cdot (1-\alpha)$ CI for β_1

Example. Test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$?
By connection between tests and C.I.s, we can test H_0 with size α if we reject H_0 whenever $0 \notin I$

Example (Q10, ES2)**.** We have special case: $Y_1, \dots, Y_n \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$ are both unknown. Want to do inference on μ.
Note: this is a normal linear model with

$$X = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \beta = \begin{bmatrix} \mu \end{bmatrix}$$

i.e. $\beta_1 = \mu$

4.5 Confidence Sets for β

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

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

Hence

$$\underbrace{\frac{\|(X^T X)^{-1/2}(\beta \hat{-} \beta)\|^2}{\sigma^2}}_{=\|X(\hat{\beta}-\beta)\|^2/\sigma^2} \sim \chi_p^2$$

This is independent of $\hat{\sigma}^2 n/\sigma^2 \sim \chi_{n-p}$ by Theorem 1. Form a pivot

$$\frac{\|X(\hat{\beta} - \beta)\|^2/\sigma^2 p}{\sigma^2 n/(\sigma^2(n-p))} \sim \frac{\chi_p/p}{\chi_{(n-p)}/(n-p)} \sim F_{p,n-p}$$

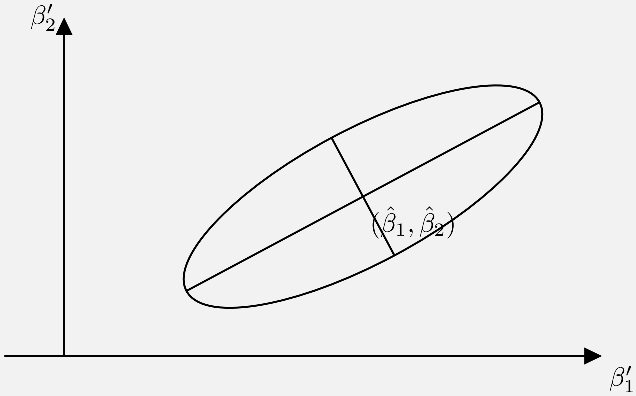
Therefore for all β, σ^2,

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

But we can say

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

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



principal axes are given by eigenvectors of $X^T X$

4.6 F-test

Method. We wish to test whether a whole collection of predictors has no effect on the response. WLOG take the first $p_0 \leq p$ predictors

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p_0} = 0$$
$$H_1 : \beta \in \mathbb{R}^p$$

Write $X = (\underbrace{X_0}_{n \times p_0}, \underbrace{X_1}_{n(p-p_0)})$

$$\beta = \begin{bmatrix} \beta^0 \\ \beta^1 \end{bmatrix} \quad {\beta^0}^T = (\beta_1, \dots, \beta_{p_0})$$

The null model has $\beta^0 = 0$, so it is a linear model:

$$Y = X\beta + \varepsilon = X_1\beta^1 + \varepsilon$$

We'll write

$$P = X(X^TX)^{-1}X^T \quad P_1 = X_1(X_1^TX_1)^{-1}X_1^T$$

Note that as X, P have full rank, so must X_1, P_1

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

Proof. $P - P_1$ is clearly symmetric. Also idempotent:

$$\begin{aligned} (P - P_1)(P - P_1) &= P^2 - PP_1 - P_1P + P_1^2 \\ &= P - P_1 - P_1 + P_1 \\ &= P - P_1 \end{aligned}$$

$$\text{rank}(P - P_1) = \text{Tr}(P - P_1) = \text{Tr}(P) - \text{Tr}(P_1) = p - (p - p_0) = p_0$$

Also

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

Method (continued)**.** Recall that the maximum log-likelihood in the normal linear model

$$\begin{aligned}\max_{\beta\in\mathbb{R}^p,\sigma^2>0} l(\beta,\sigma^2) &= l(\hat\beta,\hat\sigma^2)\\ &= -\frac{n}{2}\log(\hat\sigma^2) - \frac{n}{2} + \text{ const.}\\ &= -\frac{n}{2}\log\left(\frac{\|(I-P)Y\|^2}{n}\right) + \text{ const.}\end{aligned}$$

The generalised LRT statistic is

$$\begin{aligned}2\log\Lambda &= 2\{\max_{\beta\in\mathbb{R}^p,\sigma^2>0} l(\beta,\sigma^2) - \max_{\beta^0=0,\beta^1\in\mathbb{R}^{p-p_0},\sigma^2>0} l(\beta,\sigma^2)\\ &= n\{-\log\left(\frac{\|(I-P)Y\|^2}{n}\right) + \log\left(\frac{\|(I-P_1)Y\|^2}{n}\right)\}\end{aligned}$$

Wilk's theorem says this is approximately $\chi^2_{p_0}$ if $n\to\infty$ with p, p_0 fixed.
Note that $2\log\Lambda$ is monotone in

$$\frac{\|(I-P_1)Y\|^2}{\|(I-P)Y\|^2} = \frac{\|(I-P)Y\|^2 + \|(P-P_1)Y\|^2}{\|(I-P)Y\|^2}$$

So generalised LRT rejects when the follorring statistic is large

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

Theorem. F has an $F_{p_0,n-p}$ distribution under the null hypothesis

Proof. Recall

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

Need to show that this is indep from $\|(P-P_1)Y\|^2 \sim \chi_{p_0}\cdot\sigma^2$. Under the null,

$$\begin{aligned}(P-P_1) &= (P-P_1)(X\beta+\varepsilon)\\ &= (P-P_1)(X_1\beta^1+\varepsilon)\\ (P-P_1)\varepsilon\end{aligned}$$

So indeed

$$\frac{\|(P-P_1)Y\|^2}{\sigma^2} = \frac{\|(P-P_1)\varepsilon\|^2}{\sigma^2} \sim \chi^2_{\text{rank}(P-P_1)} = \chi_{p_0}$$

To show independence of $\|(I-P)Y\|^2$ and $\|(P-P_1)Y\|^2$ note that these depend on $(I-P)\varepsilon$ and $(P-P_1)\varepsilon$, respectively and these are independent as $\begin{bmatrix}(I-P)\varepsilon\\(P-P_1)\varepsilon\end{bmatrix}$ is MVN and $\sim N(0,\begin{bmatrix}I-P & (I-P)(P-P_1)\\(I-P)(P-P_1) & P-P_1\end{bmatrix}) = N(0,\begin{bmatrix}I-P & 0\\0 & P-P_1\end{bmatrix})$ by lemma.
Hence $(I-P)\varepsilon$ and $(P-P_1)\varepsilon$ are normal, uncorrelated, therefore independent.

So the generalised LRT of size α rejects H_0 when

$$F > F_{p_0,n-p}(\alpha)$$

Remarks.

- This is exact for every n, p, p_0
- Previously, we found test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. This is a special case of the current setting where $p_0 = 1$.
  The test we found (of size α) rejects when
  $$|\hat{\beta}_1| > t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2 n(X^T X_{11}^{-1})}{n-p}}$$
  We will show this is some critical region as the F-test. We have above iff
  $$\hat{\beta}_1^2 > t_{n-p}\left(\frac{\alpha}{2}\right)^2 \frac{\hat{\sigma}^2 n(X^T X_{11}^{-1})}{n-p}$$
  Recall
  $$T = \frac{U}{\sqrt{W/n}}, \quad U \sim N(0,1) \perp\!\!\!\perp W \sim X_n^2$$
  $$T^2 = \frac{U^2}{W/n} \sim \frac{\chi_1^2/1}{W/n} \sim F_{1,n}$$
  So previously reject when
  $$\frac{\hat{\beta}_1/(X^T X)_{11}^{-1}}{\hat{\sigma}^2 n/(n-p)} > F_{1,n-p}(\alpha)$$
  Enough to show that
  $$\frac{\hat{\beta}_1}{(X^T X)_{11}^{-1}} = \frac{\|(P-P_1)Y\|^2}{p_0}, \quad \frac{\hat{\sigma}^2 n}{n-p} = \frac{\|(I-P)Y\|^2}{n-p}$$
  Note $P - P_1$ is rank-1 projection onto the 1dim subspace spanned by
  $$(I-P)X_0 = v$$
  $$\begin{aligned}\|(P-P_1)Y\|^2 &= \|\frac{v}{\|v\|}\left(\frac{v}{\|v\|}\right)^T Y\|^2 \\ &= \frac{(v^T Y)^2}{\|v\|^2} \\ &= \frac{(X_0^T(I-P_1)Y)^2}{\|(I-P_1)X_0\|^2} = \frac{(X_0^T(I-P_1)PY)^2}{\|(I-P_1)X_0\|^2} \\ &= \frac{(X_0^T(I-P_1)X\hat{\beta})^2}{\|(I-P_1)X_0\|^2}\end{aligned}$$
  $$(I-P_1)X = [(I-P_1X_0), 0, 0, \ldots, 0]$$
  So
  $$\|(P-P_1)Y\|^2 = \|(I-P_1)X_0\|^2\hat{\beta}_1$$
  Finally, we show
  $$(X^T X)_{11}^{-1} = \frac{1}{\|(I-P_1)X_0\|^2}$$
  (exercise, apply woodbury identity to $X^T X$)