Statistics

June 5, 2023

Contents

Lectures

Start of lecture 1

0 Introduction

Statistics: The science of making informed decisions. Can include:

- Design of experiments
- Graphical exploration of data
- *Formal statistical inference* $\in$ Decision theory
- Communication of results.

Let $X_1, X_2, \dots, X_n$ be independent observations from some distribution $f_X(\bullet \mid \theta)$, with parameter θ. We wish to infer the value of θ from $X_1, \dots, X_n$.

- Estimating θ
- Quantifying uncertainty in estimator
- Testing a hypothesis about θ.

0.1 Probability Review

Let Ω be the *sample space* of outcomes in an experiment. A "nice" or measurable subset of Ω is called an *event*, we denote the set of events $\mathcal{F}$. A function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a *probability measure* if:

- $\mathbb{P}(\phi) = 0$
- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ if (A_i) are disjoint.

A *random variable* is a (measurable) function $X : \mathbb{R} \to \mathbb{R}$. For example: tossing a coin twice $\Omega = \{HH, HT, TH, TT\}$. X: number of heads.

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

The *distribution function* of X is

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

A *discrete* random variable takes values in a countable $\mathcal{X} \in \mathbb{R}$, its *probability mass function* or pmf is $p_X(x) = \mathbb{P}(X = x)$. We say X has *continuous* distribution if it has a *probability density function* or pdf satisfying

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

for any "nice" set A.

The *expectation* of X is

$$\mathbb{E}X = \begin{cases} \sum_{x \in \mathcal{X}} x p_X(x) & \text{if } X \text{ is discrete} \\ \int x f_X(x) \mathrm{d}x & \text{if } X \text{ is continuous} \end{cases}$$

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

$$\mathbb{E}g(x) = \int g(x) f_X(x) \mathrm{d}x$$

The *variance* of X is

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

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

$$\mathbb{P}(X_1 \le x_2, \ldots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n)$$

If the variables have pdf's, then

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

$(x = (x_1, \ldots, x_n),\ X = (X_1, \ldots, X_n))$.

Linear transformations

If $a_1, \ldots, a_n \in \mathbb{R}$

$$\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1 \mathbb{E}X_1 + \cdots + a_n \mathbb{E}X_n$$

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

$(\mathrm{Cov}(X_i, X_i) = \mathbb{E}((X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)))$. If $X = (X_1, \ldots, X_n)^\top$

$$\mathbb{E}X = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)^\top$$

$$\mathbb{E}(a^\top X) = a^\top \mathbb{E}X$$

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

Moment generating functions

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

This may only exist for t in some neighbourhood of 0.

- $\mathbb{E}(X^n) = \frac{\mathrm{d}^n}{\mathrm{d}t^n} M_X(0)$
- $M_X = M_Y \implies F_X = F_Y$
- Makes it easy to find the distribution function of sums of IID variables.

Example. Let $X_1, \ldots, X_n$ be IID Poisson(μ)

$$\begin{aligned} M_{X_1}(t) &= \mathbb{E}e^{tX_1} \\ &= \sum_{x=0}^{\infty} e^{tx} \cdot \frac{e^{-\mu}\mu^x}{x!} \\ &= e^{-\mu} \sum_{x=0}^{\infty} \frac{(e^t \mu)^x}{x!} \\ &= e^{-\mu} e^{\mu \exp t} \\ &= e^{-\mu(1-e^t)} \end{aligned}$$

$S_n = X_1 + \cdots + X_n$.

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

Observe this is Poisson(μn) mgf. So $S_n \sim$ Poisson(μn).

Limit Theorems

Weak law of large numbers (WLLN). $X_1, \ldots, X_n$ are IID with $\mathbb{E}X_1 = \mu$.

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the "sample mean". For all $\varepsilon > 0$,

$$\mathbb{P}(\underbrace{|\overline{X}_n - \mu| > \varepsilon}_{\text{event that depends only on } X_1, \ldots, X_n}) \to 0 \qquad \text{as } n \to \infty$$

Strong law of large numbers (SLLN)

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

(This event depends on *whole* sequence $X_1, X_2, \ldots$. $\overline{X}_n \to \mu \iff \forall \varepsilon > 0 \exists N \forall n > N |\overline{X}_n - \mu| < \varepsilon$.

Central Limit Theorem

$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ where $\sigma^2 = \text{Var}(X_i)$. Then Z_n is approximately N$(0,1)$ as $n \to \infty$.

$$\mathbb{P}(Z_n \le z) \to \Phi(z) \qquad \text{as } n \to \infty \qquad \forall z \in \mathbb{R}$$

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

Conditioning

Let X and Y be discrete random variables. Their joint pmf is

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

The marginal pmf

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y \in Y} p_{X,Y}(x,y)$$

Conditional pmf of X given $Y = y$ is

$$\begin{aligned} p_{X|Y}(x \mid y) &= \mathbb{P}(X = x \mid Y = y) \\ &= \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)} \end{aligned}$$

(defined $= 0$ if $p_Y(y) = 0$). If X, Y are continuous, the joint pdf $f_{X,y}$ has

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

The marginal pdf of Y is

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

The conditional pdf of X given Y is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Conditional expectation:

$$\mathbb{E}(X \mid Y) = \begin{cases} \sum_x x p_{X|Y}(x \mid y) \\ \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) \mathrm{d}x \end{cases}$$

(this is treated as a random variable, which is a function of Y).

Tourer property:

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}X$$

Conditional variance formula:

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

Start of lecture 2

Change of Variables (in 2D)

Let $(x, y) \mapsto (u, v)$ is a differentiable bijection. Then

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

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

Important Distributions

$X \sim \text{Negbin}(k, p)$: In successive IID $\text{Ber}(p)$ trials X is the time at which k-th success occurs.

$X \sim \text{Poisson}(\lambda)$ is the limit of a $\text{Bin}(n, \lambda/n)$ as $n \to \infty$.

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

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

This is the MGF of a $\Gamma(\sum \alpha_i, \lambda)$. Hence $S_n \sim \Gamma(\sum \alpha_i, \lambda)$. Also, if $X \sim \Gamma(a, \lambda)$, then for any $b \in (c, \infty)$, $bX \sim \Gamma(\alpha, \lambda/b)$.

Special cases

$\Gamma(1, \lambda) = \text{Exp}(\lambda)$, $\Gamma(k/2, 1/2) = \chi_k^2$ "Chi-squared with k degrees of freedom." Sum of k independent squared $\text{N}(0, 1)$ random variables.

0.2 Estimation

Suppose we observe data $X_1, X_2, \dots, X_n$ which are IID from some PDF (pmf) $f_X(x \mid \theta)$, with θ unknown.

Definition (Estimator)**.** An *estimator* is a statistic or a function of the data $T(X) = \hat{\theta}$, which we use to approximate the true parameter θ. The distribution of $T(X)$ is called the *sampling distrbution.*

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{N}(\mu, 1)$.

$$\hat{\mu} = T(X) = \frac{1}{h} \sum_{i=1}^{n} X_i$$

The sampling distribution of $\hat{\mu}$ is $\text{N}\left(\mu, \frac{1}{n}\right)$.

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

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

Note. In general, the bias is a function of θ, even if notation $\text{bias}(\hat{\theta})$ does not make it explicit.

Definition. We say that $\hat{\theta}$ is *unbiased* if $\text{bias}(\hat{\theta}) = 0$ for all $\theta \in \Theta$.

Example (Continuing from previous)**.** $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased because $\mathbb{E}_\mu(\hat{\mu}) = \mu$ for all $\mu \in \mathbb{R}$.

Definition. The *mean squared error* (mse) of $\hat{\theta}$ is

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

Note. Like the bias, $\text{mse}(\hat{\theta})$ is a function of θ!

Bias-variance decomposition

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

The two terms on the RHS are ≥ 0.

There is a trade off between bias and variance.

Example. $X \sim \text{Bin}(n, \theta)$. Suppose n known, we wish to estimate θ. Standard estimator $T_u = \frac{X}{n}$, then $\mathbb{E}_\theta T_u = \frac{\mathbb{E}_\theta X}{n} = \theta$ (holds for all θ). Hence T_u is unbiased.

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

Consider a second estimator

$$T_B = \frac{X+1}{n+2} = \omega \frac{X}{n} + (1-\omega)\frac{1}{2}$$

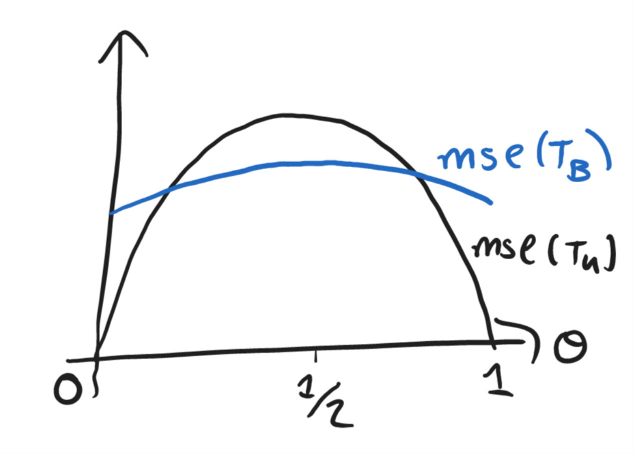
with $\omega = \frac{n}{n+2}$. If $X = 8$, $n = 10$ (8 successes in 10 trials), then $T_u = 0.8$, $T_B = \frac{9}{12} = 0.75$.

$$\begin{aligned}\text{bias}(T_B) &= \mathbb{E}_\theta T_B - \theta \\ &= \mathbb{E}\left(\frac{X+1}{n+2}\right) - \theta \\ &= \frac{n}{n+2}\theta + \frac{1}{n+2} - \theta\end{aligned}$$

This is $\neq 0$ for all but one value of θ. Hence T_b is biased.

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

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



Message: Our prior judgements about θ affect our choice of estimator (for example in this previous example, if we knew the X_i represent coin flips, then we expect θ to be near $\frac{1}{2}$, so we should use $\mathrm{mse}(T_B)$).

Unbiasedness is not necessarily desirable. Consider this pathological example:

Example. Suppose $X \sim \mathrm{Poisson}(\lambda)$. We wish to estimate $\theta = \mathbb{P}(X = 0)^2 = e^{-2\lambda}$. For an estimator $T(X)$ to be unbiased we must have for all λ

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

$$\iff \sum_{x=0}^{\infty} T(x) \frac{\lambda^x}{x!} = e^{-\lambda} = \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!}$$

for this to hold $\forall \lambda \ge 0$, we need

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

This estimator makes no sense!

Start of lecture 3

0.3 Sufficiency

$X_1, \dots, X_n$ are IID random variables from a distribution with pdf (or pmf) $f_X(\bullet \mid \theta)$. Let $X = (X_1, \dots, X_n)$.

Question: Is there a statistic $T(X)$ which contains all information in X needed to estimate θ?

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

Remark. θ and $T(X)$ could be vector-valued.

Example. $X_1, \ldots, X_n \overset{\text{IID}}{\sim} \text{Ber}(\theta)$ for $\theta \in [0, 1]$.

$$\begin{aligned} f_X(\bullet \mid \theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \end{aligned}$$

Note. This only depends on X through $T(X) = \sum_{i=1}^{n} X_i$.

For x with $\sum x_i = t$,

$$\begin{aligned} f_{X|T=t}(x \mid T(x) = t) &= \frac{\mathbb{P}_\theta(X = x, T(X) = t)}{\mathbb{P}_\theta(T(X) = t)} \\ &= \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta(T(X) = t} \\ &= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \binom{n}{t}^{-2} \end{aligned}$$

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

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

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

for suitable functions g, h.

Proof. (Discrete case)

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

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

As this doesn't depend on θ, $T(X)$ is sufficient.

Conversely, suppose $T(X)$ is sufficient, then

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

Then by sufficiency of T, $h(x)$ doesn't depend on θ (so it is a function of x). Thus the pmf of X, $f_X(\bullet \mid \theta)$ factorises as in the statement of the theorem. □

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Ber}(\theta)$.

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

Take $g(t,\theta) = \theta^t(1-\theta)^{n-t}$, $h(x) = 1$. This immediately implies $T(X) = \sum x_i$ is sufficient.

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Unif}([0,\theta])$, $\theta > 0$. Then

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

$T(x) = \max_i x_i$. Then by factorisation lemma, $T(x) = \max_i x_i$ is sufficient for θ.

Minimal Sufficiency

Sufficient stats are *not* unique. Indeed any 1-to-1 function of a sufficient statistic is also sufficient. Also $T(X) = X$ is always sufficient by not very useful.

Definition. A sufficient statistic T is *minimal sufficient* if it is a function of any other sufficient statistic. That is, if T' is also sufficient, then

$$T'(x) = T'(y) \implies T(x) = T(y)$$

for all $x, y \in \mathcal{X}^n$.

Remark. Any two minimal sufficient statistics, T, T' are "in bijection with each other":

$$T(x) = T(y) \iff T'(x) = T'(y)$$

Useful condition to check minimal sufficiency.

Theorem (Minimal Sufficiency Theorem)**.** Suppose that $T(X)$ is a statistic such that $f_X(x \mid \theta)/f_X(y \mid \theta)$ is constant as a function of θ if and only if $T(x) = T(y)$. Then T is minimal sufficient.

Let $x \overset{1}{\sim} y$ if $\frac{f_X(x|\theta)}{f_X(y|\theta)}$ is constant in θ. It's easy to check that $\overset{1}{\sim}$ is an equivalence relation.

Similarly, for a given statistic T, $x \overset{2}{\sim} y$ if $T(x) = T(y)$ defines another equivalence relation. The condition of theorem says $\overset{1}{\sim}$ and $\overset{2}{\sim}$ are the same.

Note. We can always construct a statistic T which is constant on the equivalence classes of $\overset{1}{\sim}$, which by the theorem is minimal sufficient.

Proof. For any value t of T, let z_t be a representative from the equivalence class

$$\{x \mid T(x) = t\}$$

Then

$$f_X(x \mid \theta) = \underbrace{f_X(z_{T(x)} \mid \theta)}_{g(T(x),\theta)} \underbrace{\frac{f_X(x \mid \theta)}{f_X(z_{T(x)} \mid \theta)}}_{h(x)}$$

Where $h(x)$ does not depend on θ by the hypothesis, as $x \overset{1}{\sim} z_{T(x)}$. By factorisation criterion, T is sufficient.

To prove that T is minimal, take any other sufficient statistic S. Want to prove that if $S(x) = S(y)$ then $T(x) = T(y)$. By factorisation criterion, there are functions g_S, h_S such that

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

Suppose $S(x) = S(y)$. Then

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

which doesn't depend on θ. Hence $x \overset{1}{\sim} y$. By hypothesis, $x \overset{2}{\sim}$, hence $T(x) = T(y)$. $\square$

Remark. Sometimes the range of X depends on θ (for example $X_1, \dots, X_n \overset{\text{IID}}{\sim}$ Unif$([0, \theta])$. In this case we can interpret

$$\text{“}\frac{f_X(x|\theta}{f_X(y|\theta)} \text{ is constant in } \theta\text{”}$$

to mean that $f_X(x \mid \theta) = c(x, y) f_X(y \mid \theta)$ for some function c which does not depend on θ.

Start of lecture 4

Example. Suppose that $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma^2)$, with parameters (μ, σ^2) unknown.

$$\begin{aligned}\frac{f_X(x \mid \theta)}{f_X(y \mid \theta)} &= \frac{(2\pi\sigma^2)^{-\pi/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\}}{(2\pi\sigma^2)^{-\pi/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2\right\}} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) \frac{\mu}{\sigma^2}(\sum_{i=1}^n x_I - \sum_{i=1}^n y_i\right\}\end{aligned}$$

If $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, this ratio does not depend on (μ, σ^2). The converse is also true: if the ratio does not depend on (μ, σ^2) then we must have $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. By the theorem, $T(X) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is minimal sufficient.

Recall that bijections of T are also minimal sufficient. A more common way of expressing a minimal sufficient statistic in this model is

$$S(X) = (\overline{X}, S_{XX})$$

$$\overline{X} = \frac{1}{n}\sum_i X_i \qquad S_X X = \sum_i (X_i - \overline{X})^2$$

In this example, (μ, σ^2) and $T(X)$ are both 2-dimensional. In general, the parameter and sufficient statistic can have different dimensions.

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \mu^2)$, $\mu \geq 0$. Here, the minimal sufficient statistic is $S(X) = (\overline{X}, S_{XX})$.

Rao-Blackwell Theorem

Note. So far we've written $\mathbb{E}_\theta$, $\mathbb{P}_\theta$ to denote expectations and probabilities in the model where $X_1, \dots, X_n \overset{\text{IID}}{\sim} f_X(\bullet \mid \theta)$. From now on, I'll drop the subscript θ.

Theorem (Rao-Blackwell)**.** Let T be a sufficient statistic for θ. Let $\tilde{\theta}$ be some estimator for θ, with $\mathbb{E}(\tilde{\theta}^2) < \infty$, for all θ. Define a new estimator $\hat{\theta} = \mathbb{E}_\theta(\tilde{\theta} \mid T(X))$. Then, for all θ,

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

($\text{mse}(\hat{\theta}) \le \text{mse}(\tilde{\theta})$). The inequality is strict unless $\tilde{\theta}$ is a function of $T(X)$.

Remark. $\hat{\theta}$ is a valid estimator, i.e. it does not depend on θ, only depends on X, because T is sufficient.

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

Moral. We can improve the mse of any estimator $\tilde{\theta}$ by taking a conditional expectation given $T(X)$.

Proof. By the tower property:

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

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

$$\begin{aligned} \text{Var}(\tilde{\theta}) &= \mathbb{E}(\text{Var}(\tilde{\theta} \mid T)) + \text{Var}(\mathbb{E}(\tilde{\theta} \mid T)) \\ &= \mathbb{E}[\underbrace{\text{Var}(\tilde{\theta} \mid T)}_{\ge 0 \text{ with } \mathbb{P}=1}] + \text{Var}(\hat{\theta}) \\ \implies \text{Var}(\tilde{\theta}) &\ge \text{Var}(\hat{\theta}) \end{aligned}$$

for all θ. Therefore $\text{mse}(\tilde{\theta}) \ge \text{mse}(\hat{\theta})$.

Note: $\text{Var}(\tilde{\theta} \mid T) > 0$ with some positive probability unless $\tilde{\theta}$ is a function of $T(X)$. So $\text{mse}(\tilde{\theta}) > \text{mse}(\hat{\theta})$ unless $\tilde{\theta}$ is a function of $T(X)$. □

Example. Say $X_1, \dots, X_n \overset{\text{IID}}{\sim}$ Poisson(λ). We wish to estimate $\theta = \mathbb{P}(X_1 = 0) = e^{-\lambda}$.

$$\begin{aligned} f_X(x \mid \lambda) &+ \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_i x_i!} \\ \implies f_X(x \mid \theta) &= \frac{\theta^n(-\log\theta)^{\sum x_i}}{\prod_i x_i!} \end{aligned}$$

Letting $h(x) = \frac{1}{\prod x_i!}$, $g(T(X), \theta) = \theta^n(-\log\theta)^{T(X)}$, then by factorisation criterion, $T(X) = \sum X_i$ is a sufficient statistic. Let $\tilde{\theta} = \mathbb{1}_{\{X_1=0\}}$ (unbiased: only uses one observation X_1).

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

So $\hat{e} = \left(1 - \frac{1}{n}\right)^{\sum x_i}$ is an estimator which by the Rao-Blackwell theorem has

$$\operatorname{mse}(\hat{\theta}) < \operatorname{mse}(\tilde{\theta})$$

Sanity check: What happens as $n \to \infty$?

$$\hat{\theta} = \left(1 - \frac{1}{n}\right)^{n\overline{x}} \overset{n\to\infty}{\longrightarrow} e^{-\overline{x}}$$

and by the Strong Law of Large Numbers, $\overline{X} \to \mathbb{E}X_1 = \lambda$ so $\theta^n \approx e^{-\lambda} = \theta$ as h grows large.

Example. Let $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Unif}([0, \theta])$, θ unknown. $\theta \ge 0$. Recall $T(X) = \max_i X_i$ is sufficient for θ. Let $\tilde{\theta} = 2X_1$, which is unbiased. Then

$$\begin{aligned}
\hat{\theta} &= \mathbb{E}[\tilde{\theta} \mid T = t] \\
&= 2\mathbb{E}[X_1 \mid \max_i X_i = t] \\
&= 2\mathbb{E}[X_1 \mid \max_i X_i = t, \max_i X_i = X_1]\mathbb{P}[\max_i X_i = X_1 \mid \max X_i = t] \\
&\quad + \mathbb{E}[X_1 \mid \max_i X_i = t, \max_i X_i \ne X_1]\mathbb{P}[\max_i X_i \ne X_1 \max X_i = t] \\
&= \frac{2t}{n} + \frac{2(n-1)}{n}\mathbb{E}[X_1 \mid X_1 \le t, \max_{1 \le i \le n} X_i = t] \\
&= \frac{2t}{n} + \frac{2(n-1)}{n}\frac{t}{2}
\end{aligned}$$

So $\hat{\theta} = \frac{n+1}{n} \max_i X_i$ is a valid estimator with

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

Start of lecture 5

0.4 Maximum likelihood Estimation

Let $X = (X_1, \dots, X_n)$ have f=joint pdf (or pmf) $f_X(x \mid \theta)$.

Definition (Likelihood function)**.** The likelihood function is

$$L : \theta \mapsto f_X(X \mid \theta)$$

The maximum likelihood estimator (mle) is any value of θ maximising $L(\theta)$.

If $X_1, \dots, X_n$ are IID each with pdf (or pmf) $f_X(\bullet \mid \theta)$, then

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

We'll denote the logarithm

$$l(\theta) = \log L(\theta) = \sum_{i=1}^n \log f_X(x_i \mid \theta)$$

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Ber}(\theta)$.

$$l(\theta) = \left(\sum X_i\right) \log \theta + \left(n - \sum X_i\right) \log(1 - \theta)$$

$$\frac{\partial l}{\partial \theta} = \frac{\sum X_i}{\theta} - \frac{n - \sum X_i}{1 - \theta}$$

This is equal to 0 if and only if $\theta = \frac{1}{n} \sum X_i = \overline{X}$. Hence $\overline{X}$ is the mle for θ. This is unbiased as $\mathbb{E}\overline{X} = \theta$.

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{N}(\mu, \sigma^2)$

$$l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

This is maximised when $\frac{\partial l}{\partial \mu} = \frac{\partial l}{\partial \sigma^2} = 0$

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

equal to 0 when $\mu = \overline{X}$ $(\forall \sigma^2)$

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

This is equal to 0 when $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n} S_{XX}$. Hence $(\hat{\mu}, \hat{\sigma}^2) = (\overline{X}, S_X X/n)$ are the mle in this model.

Note that $\overline{\mu} = \overline{X}$ is unbiased. Is $\hat{\sigma}^2$ biased? We could compute $\mathbb{E}\hat{\sigma}^2$ directly. Later in the course, we'll show that

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

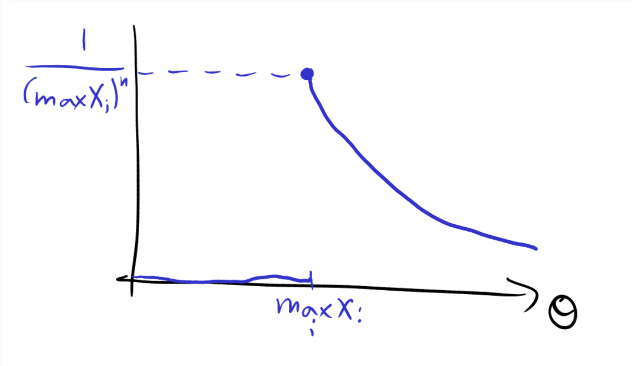
$$\mathbb{E}\hat{\sigma}^2 = \mathbb{E}(\chi^2_{n-1}) \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

So $\hat{\sigma}^2$ is biased, but asymptotically unbiased:

$$\text{bias}(\hat{\sigma}^2) \overset{n \to \infty}{\longrightarrow} 0 \qquad \forall \sigma^2$$

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Unif}[0, \theta]$

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



We an see from the plot that $\hat{\theta} = \max_i X_i$ is the mle for θ. Last time we started from unbiased estimator $\tilde{\theta} = 2X_1$ and using the R-B theorem we found an estimator

$$\hat{\theta} = \frac{n+1}{n} \max_i X_i$$

This is also unbiased. So in this model the mle is biased as

$$\mathbb{E}\hat{\theta}_{mle} = \mathbb{E}\left[\frac{n+1}{n}\hat{\theta}\right] = \frac{n}{n+1}\theta$$

but it is asymptotically unbiased.

Properties of the mle

(1) If T is a sufficient statistic then the mle is a function of $T(X)$. By the factorisation criterion:

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

If $T(x) = T(y)$ the likelihood function with data x or y is the same up to a multiplicative constant. Hence, the mle in each case is the same.

(2) If $\phi = h(\theta)$ where h is a bijection, then the mle of ϕ is $\hat{\phi} = h(\hat{\theta})$ where $\hat{\theta}$ is the mle of θ.

(3) Asymptotic normality: $\sqrt{n}(\hat{\theta} - \theta)$ is approximately normal with mean 0 when n is large. Under some regularity conditions, for a "nice set" A,

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

where $z \sim \text{N}(0, \underline{\Sigma})$. This holds for all "regular" values of θ.

Here Σ is some function of l, and there is a theorem (Cramer-Rao) which says this is the smallest variable attainable.

(4) Sometimes if the mle is not available analytically, we can find it numerically.

Confidence Intervals

Example. Vaccine has 76% efficacy in a 3-month period, with a 95% confidence interval $(59\%, 86\%)$

Definition (Confidence Interval)**.** A $(100 \cdot \gamma)\%$-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 values of θ. (A and B are random, and θ is fixed).

Correct or frequentist interpretation:

There exists some fixed true parameter θ. We repeat the experiment many times. On average, $100 \cdot \gamma\%$ of the time the interval $(A(X), B(X))$ contains θ.

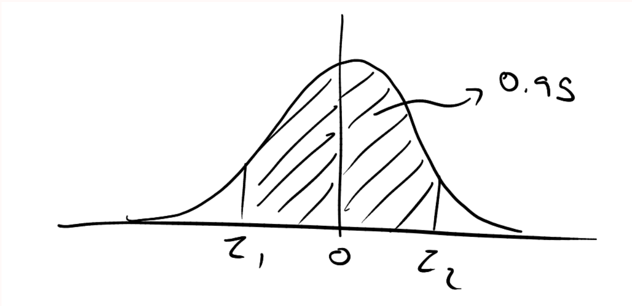
Misleading interpretation:
"Having observed $X = x$, there is a probability γ that θ is in $(A(x), B(x))$."

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\theta, 1)$. Find a 95% confidence interval for θ. We know that

$$\overline{X} = \frac{1}{n}\sum X_i \sim \mathrm{N}\left(\theta, \frac{1}{n}\right)$$
$$\implies z := \sqrt{n}(\overline{X} - \theta) \sim \mathrm{N}(0,1)$$

z has this distribution for all θ.

Let z_1, z_2 be any two numbers such that $\Phi(z_2) - \Phi(z_1) = 0.95$.



Then

$$\mathbb{P}(z_1 \le \sqrt{n}(\overline{X} - \theta) \le z_2) = 0.95$$

Rearrange:

$$\mathbb{P}\left(\overline{X} - \frac{z_2}{\sqrt{n}} \le \theta \le \overline{X}\frac{z_1}{\sqrt{n}}\right) = 0.95$$

Then $\left(\overline{X} - \frac{z_2}{\sqrt{n}}, \overline{X} + \frac{z_2}{\sqrt{n}}\right)$ is a 95% confidence interval. How to choose z_1, z_2? Usually we minimise the width of interval. In this case this is achieved by

$$z_1 = \Phi^{-1}(0.025), \qquad z_2 = \Phi^{-1}(0.975)$$

Start of lecture 6

Recipe for Confidence Interval

(1) Find some quantity $R(X, \theta)$ such that the $\mathbb{P}_\theta$-distribution of $R(X,\theta)$ does not depend on θ. This is called a *pivot*. For example

$$z = \sqrt{n}(\overline{X} - \mu) \sim \mathrm{N}(0,1) \qquad \forall \mu$$

(2) Write down a probability statement about the pivot of the form

$$\mathbb{P}(c_1 \le R(X,\theta) \le c_2) = \gamma$$

by using the quantities c_1, c_2 of the distribution of $R(X,\theta)$ [typically a $\mathrm{N}(0,1)$ or χ^2_p distribution).

(3) Rearrange the inequalities to leave θ in the middle.

Proposition. If T is a monotone increasing function $T : \mathbb{R} \to \mathbb{R}$, and $(A(x), B(X))$ is a $100\gamma\%$ confidence interval for θ, then $(T(A(X)), T(B(X)))$ is a confidence interval for $T(\theta)$.

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

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(0, \sigma^2)$. Find a 95% confidence interval for σ^2.

(1) Note that $\frac{X_i}{\sigma} \sim \mathrm{N}(0,1)$

$$\implies \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi_n^2$$

Hence $R(X, \sigma^2) = \sum_i \frac{X_i^2}{\sigma^2}$ is a pivot.

(2) Let $C_1 = F_{\chi_n^2}^{-1}(0.025)$, $c_2 = F_{\chi_n^2}^{-1}(0.975)$. Then

$$\mathbb{P}\left(c_1 \le \frac{1}{\sigma^2} \sum_i X_i^2 \le c_2\right) = 0.95$$

(3) Rearranging:

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

Hence $\left[\frac{\sum X_i^2}{c_2}, \frac{\sum X_i^2}{c_1}\right]$ is a 95% confidence interval for σ^2.

Hence, using the proposition above, $\left[\sqrt{\frac{\sum X_i^2}{c_2}}, \sqrt{\frac{X_i^2}{c_1}}\right]$ is a 95% confidence interval for σ.

Example. $X_1, \ldots, X_n \stackrel{\text{IID}}{\sim} \text{Ber}(p)$, n is large. Find an approximate 95% confidence interval for p.

(1) The mle for p is $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$. By the Central limit theorem when n is large, $\hat{p}$ is approximately $\text{N}\left(p, \frac{p(1-p)}{n}\right)$. Therefore $\sqrt{n}\frac{(\hat{p}-p)}{\sqrt{p(1-p)}}$ is approximately $\text{N}(0,1)$.

(2) $z = \Phi^{-1}(0.975)$

$$\mathbb{P}\left(-z \le \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}} \le z\right) \approx 0.95$$

(3) Rearranging this is tricky. Argue that as $n \to \infty$, $\hat{p}(1-\hat{p}) \to p(1-p)$. So replace denominator:

$$\mathbb{P}\left(-z \le \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \le z\right) \approx 0.95$$

Now it's easier to rearrange:

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

So $\left[\hat{p} \pm z\frac{\sqrt{\hat{p}(1-\hat{p}j}}{\sqrt{n}}\right]$ is an approximate 95% confidence interval for p.

Note.

- $z \approx 1.95$
- $\sqrt{\hat{p}(1-\hat{p})} \le \frac{1}{2}$ for all $\hat{p} \in (0,1)$

So a "conservative" confidence interval is $\left[\hat{p} \pm 1.96 \cdot \frac{1}{2} \cdot \frac{1}{\sqrt{n}}\right]$.

0.5 Interpreting Confidence intervals

Suppose $X_1, X_2 \stackrel{\text{IID}}{\sim} \text{Unif}\left[\theta - \frac{1}{2}, \theta + \frac{1}{2}\right]$. What is a sensible 50% confidence interval for θ? Consider

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

Immediately conclude that $(\min(X_1, X_2), \max(X_1, X_2))$ is a 50% confidence interval for θ.

But we observe $X_1 = x_1$, $X_2 = x_2$ with $|x_1 - x_2| > \frac{1}{2}$. In this case we can be sure that θ is in $(\min(x_1, x_2), \max(x_1, x_2)$.

Frequentist interpretation of confidence interval is entirely correct! If we repeat the experiment many times $\theta \in (\min(X_1, X_2), \max(X_1, X_2))$ *exactly* 50% of the time. However, we cannot say that *given* a *specific* observation (x_1, x_2) we are "50% certain that $\theta \in$ C.I..

Bayesian Inference

So far, we have assume that there is some true parameter θ. That data X has pdf (or pmf) $f_X(\bullet \mid \theta)$.

Bayesian analysis is a different framework, where we treat θ as a random variable taking values in Θ.

We being by assigning to θ a *prior distribution* $\pi(\theta)$, which represents the investigator's opinions or information about θ *before* seeing any data. Conditional on θ, the data X has pdf (or pmf) $f_X(x \mid \theta)$. Havign observed a specific value of $X = x$, this information is combined with the prior to form the *posterior distribution.* $\pi(\theta \mid x)$ which is the conditional distribution of θ given $X = x$.

By Bayes rule:

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

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

$$f_X(x) = \begin{cases} \int_\Theta f_X(x \mid \theta)\pi(\theta)\mathrm{d}\theta & \text{if } \theta \text{ is constant} \\ \sum_{\theta \in \Theta} f_X(x \mid \theta)\pi(\theta) & \text{if } \theta \text{ is discrete} \end{cases}$$

Start of lecture 7

Bayesian Analysis

Idea: treat θ as a random variable.
Prior distribution: $\pi(\theta)$ (Info about θ before seeing data)
Joint distribution of X, θ:

$$f_X(x \mid \theta) \cdot \pi(\theta)$$

Posterior distribution:

$$\begin{aligned} \pi(\theta \mid x) &= \frac{f_X(f \mid \theta)\pi(\theta)}{\int f_X(x \mid \theta)\pi(\theta)\mathrm{d}\theta} \\ &\propto f_X(x \mid \theta)\pi(\theta) \end{aligned}$$

(likelihood times prior).

Example (Prior choice clear)**.** Patient gets a COVID test:

$$\theta = \begin{cases} 0 & \text{patient does not have COVID} \\ 1 & \text{patient does have COVID} \end{cases}$$

Data:

$$X = \begin{cases} 0 & \text{negative test} \\ 1 & \text{positive test} \end{cases}$$

We know: Sensitivity of test:

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

Specificity of test:

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

What prior? Suppose we don't know anything about patient but we know that a proportion p of people in the UK are infected today. Natural choice:

$$\pi(\theta = 1) = p$$

Chance of infection given true test?

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

If $\pi(\theta = 0) \gg \pi(\theta = 1)$, this posterior can be small.

Example. $\theta \in [0,1]$ mortality rate for new surgery at addenbrookes. In the first 10 operations, there were no deaths. Model: $X_i \sim \text{Ber}(\theta)$, $X_i = 1$ if i-th operation is death, 0 otherwise.

$$f_X(x \mid \theta) = \theta^{\sum X_i}(1-\theta)^{10-\sum X_i}$$

Prior: We're told that the surgery is performed in other hospitals with a mortality rate ranging from 3% to 20%, with an average of 10%. We'll say that $\pi(\theta)$ is $\text{Beta}(a,b)$. We choose $a = 3$, $b = 27$, so that the mean of $\pi(\theta)$ is 0.1 and

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

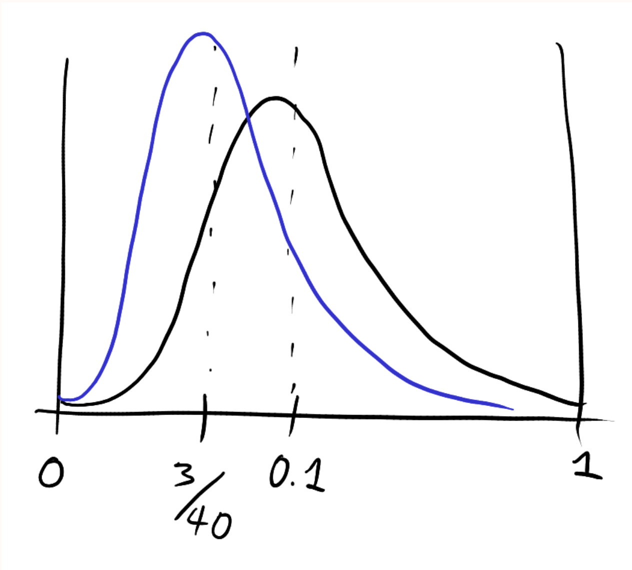
Posterior:

$$\begin{aligned}\pi(\theta \mid x) &\propto \pi(\theta) \times f_X(x \mid \theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{\sum x_i}(1-\theta)^{10-\sum x_i} \\ &= \theta^{\sum x_i + a - 1}(1-\theta)^{b+10-\sum x_i - 1}\end{aligned}$$

(we ommitted the normalising constant of $\text{Beta}(a,b)$ because it does not depend on θ). We deduce this is a $\text{Beta}\left(\sum x_i + a, 10 - \sum x_i + b\right)$ distribution. In our case

$$\sum_{i=1}^{10} x_i = 0, \quad a = 0, \quad b = 27$$

$\implies \text{Beta}(3, 37)$



Note. Here prior and posterior are in the same family of distrbutions. This is known as conjugacy.

What to do with posterior? The information in $\pi(G \mid x)$ can be used to make decisions under uncertainty.

Formal Process

(1) We must pick a decision $\delta \in D$.

(2) The loss function $L(\theta, \delta)$ is the loss incurred when we make decision δ and true parameter has value θ. For example $\delta = \{0, 1\}$, $\delta = 1$ means we ask the patient to self isolate. Then, $L(\theta = 0, \delta = 1)$ is the loss incurred when we ask a non-infected patient to self-isolate.

(3) We pick decision which minimises the posterior expected loss:

$$\delta^* = \arg\min_{\delta \in D} \int_\Theta L(\theta, \delta)\pi(\theta \mid x)\mathrm{d}\theta$$

(Von Neumann-Morgenstern theorem)

Point estimation:
The decision is a "best guess" for the true parameter, so $\delta \in \Theta$. The *Bayes estimator* $\hat{\theta}^{(b)}$ minimises

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

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

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

$h'(\delta) = 0$ if

$$\int (\theta - \delta)\pi(\theta \mid x)\mathrm{d}\theta = 0$$

$$\iff \int \theta\pi(\theta \mid x)\mathrm{d}\theta = \delta \underbrace{\int \pi(\theta \mid x)\mathrm{d}\theta}_{=1}$$

Hence $\hat{\theta}^{(b)}$ equals the posterior mean of θ.

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

$$\begin{aligned} h(\delta) &= \int |\theta - \delta| \pi(\theta \mid x) \mathrm{d}\theta \\ &= \int_{-\infty}^{\delta} -(\theta - \delta)\pi(\theta \mid x)\mathrm{d}\theta + \int_{\delta}^{\infty} (\theta - \delta)\pi(\theta \mid x)\mathrm{d}\theta \\ &= -\int_{-\infty}^{\delta} \theta\pi(\theta \mid x)\mathrm{d}\theta + \int_{\delta}^{\infty} \theta\pi(\theta \mid x)\mathrm{d}\theta + \delta\int_{-\infty}^{\delta} \pi(\theta \mid x)\mathrm{d}\theta - \delta\int_{\delta}^{\infty} \pi(\theta \mid x)\mathrm{d}\theta \end{aligned}$$

Take derivative with respect to δ. By the FTC,

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

So $h'(\delta) = 0$ if and only if

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

So in this case

$$\hat{\theta}^{(b)} = \text{median of the posterior}$$

Credible Interval

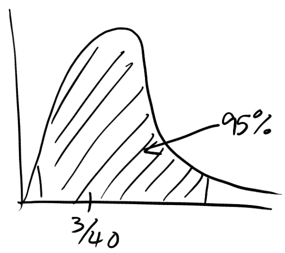
A $100\gamma\%$ *credible interval* $(A(x), B(x))$ is one which satisfies

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

(A and B are fixed at the observed data x, but θ is random).

$$\int_{A(x)}^{B(x)} \pi(\theta \mid x)\mathrm{d}\theta = \gamma$$

In example sheet 2:



Note. We *can* interpret intervals conditionally ("given x, we are $100\gamma\%$ sure that $\theta \in [A(x), B(x)]$").

Note. If T is a sufficient statistic, $\pi(\theta \mid x)$ only depends on x through $T(x)$.

$$\begin{aligned}\pi(\theta \mid x) &\propto \pi(\theta) \times f_X(x \mid \theta) \\ &= \pi(\theta) g(T(x), \theta) h(x) \\ &\propto \pi(\theta) g(T(x), \theta)\end{aligned}$$

Start of lecture 8

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, 1)$. Prior: $\pi(\mu)$ is $\mathrm{N}\left(0, \frac{1}{\tau^2}\right)$

$$\begin{aligned}\pi(\mu \mid x) &\propto f_X(x \mid \mu) \cdot \pi(\mu) \\ &\propto \exp\left[-\frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right] \exp\left[-\frac{\mu^2\tau^2}{2}\right] \\ &\propto \exp\left[-\left(\frac{1}{2}\right)^{(n+\tau^2)} \left\{\mu - \frac{\sum x_i}{n + \tau^2}\right\}^2\right]\end{aligned}$$

we recognise this as a

$$\mathrm{N}\left(\frac{\sum x_i}{n + \tau^2}, \frac{1}{n + \tau^2}\right)$$

distribution. The Bayes estimator $\hat{\mu}^{(b)} = \frac{\sum x_i}{n + \tau^2}$ for both quadratic loss and absolute error loss ($\hat{\mu}^{\text{mle}} = \frac{\sum x_i}{n}$). A 95% credible interval is

$$\left(\hat{\mu}^{(b)} - \frac{1.96}{\sqrt{n + \tau^2}}, \hat{\mu}^{(b)} + \frac{1.96}{\sqrt{n + \tau^2}}\right)$$

This is close to a 95% confidence interval when $n \gg \tau^2$.

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$. Prior: $\pi(\lambda)$ is Exp(1), $\pi(\lambda) = e^{-\lambda}$, $\lambda > 0$.

$$\begin{aligned} \pi(\lambda \mid x) &\propto f_X(x \mid \lambda) \cdot \pi(\lambda) \\ &\propto \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod_i x_i!} e^{-\lambda} & \lambda > 0 \\ &= e^{-(n+1)\lambda}\lambda^{\sum x_i} & \lambda > 0 \end{aligned}$$

THis is a $\Gamma\left(1 + \sum x_i, n+1\right)$ distribution. The Bayes estimator under quadratic loss is the posterior mean

$$\hat{\lambda}^{(b)} = \frac{\sum x_i + 1}{n+1} \xrightarrow{n\to\infty} \frac{\sum x_i}{n} = \hat{\lambda}^{\text{mle}}$$

Under the absolute error loss the bayes estimator $\tilde{\lambda}^{(b)}$ has

$$\int_0^{\tilde{\lambda}^{(b)}} \frac{(n+1)^{\sum x_i - 1}}{(\sum x_i)!} x^{\sum x_i} e^{-(n+1)\lambda} \mathrm{d}\lambda = \frac{1}{2}$$

Simple Hypothesis

A *hypothesis* is some assumption about the distribution of the data X. Scientific questions are phrased as a choice between a *null hypothesis* H_0 (base case, simple model, no effect) and an *alternative hypothesis* H_1 (complex model, interesting case, positive or negative effect).

Examples and non-examples of simple hypotheses (no explanation yet)

(1) $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{Ber}(\theta)$, H_0: $\theta = \frac{1}{2}$ (fair coin), H_1: $\theta = \frac{3}{4}$. This is a valid pair.

(2) As in the previous but H_0: $\theta = \frac{1}{2}$ and H_1: $\theta \neq \frac{1}{2}$. This is not a valid pair.

(3) $X_1, \dots, X_n$ takes values in $\mathbb{N}_0$. H_0: $X_i \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$ for some $\lambda > 0$, H_2: $X_i \overset{\text{IID}}{\sim} f_1$ for some other f_1. This is not a valid pair.

(4) X has pdf $f(\bullet \mid \theta)$, $\theta \in \Theta$. H_0: $\theta \in \Theta_0 \subset \Theta$, H_1: $\theta \notin \Theta_0$. This is simple if $\Theta_0 = \{\theta_0\}$.

A hypothesis is said to be *simple* if it fully specifies the distribution of X. Otherwise we say it is *composite*.

A test of H_0 is defined by a *critical region* $C \subseteq \mathcal{X}$. When $X \in C$ we "reject" H_0 and when $X \notin C$ we say we "fail to reject" or "find no evidence against" H_0.

Type I error: we reject H_0 when H_0 is true.
Type II error: we fail to reject H_0 when H_0 is false.

When H_0 and H_1 are simple, we define

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

"probability of type I error".

$$\beta = \mathbb{P}_{H_2}(H_0 \text{ is not rejected}) = \mathbb{P}_{H_1}(X \notin C)$$

"probability of type II error".

The *size* of the test is α. The *power* of the test is $1 - \beta$. Tradeoff between minimising size and maximising power. Usually we fix an acceptable size (say $\alpha = 1\%$), then pick test of size α which maximises the power.

Neyman-Pearson Lemma

Let H_0, H_1 be simple. Let X have pdf f_i under H_i, $i = 0, 1$. The likelihood ratio statistic

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

A likelihood ratio test (LRT) rejects H_0 when

$$X \in C = \{x : \Lambda_x(H_0, H_1) > k\}$$

for some threshold or "critical value" k.

Theorem (Neyman-Pearson Lemma)**.** Suppose that f_0, f_1 are non-zero on the same sets. Suppose there exists k such that the LRT with critical region

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

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

Remark. A LRT of size α need not exist (try to think of an example). Even then, there is a "randomised LRT" with size α.

Proof. Let $\overline{C}$ be complement of C. The LRT has

$$\alpha = \mathbb{P}_{H_0}(X \in C) \qquad = \int_C f_0(x)\mathrm{d}x$$

$$\beta = \mathbb{P}_{H_1}(X \notin C) \qquad = \int_{\overline{C}} f_1(x)\mathrm{d}x$$

Let C^* be critical region of another test with size α^*, power $1-\beta^*$, with $\alpha^* \le \alpha$. Want to prove that $\beta \le \beta^*$ or $\beta - \beta^* \le 0$.

$$\begin{aligned}
\beta - \beta^* &= \int_{\overline{C}} f_1(x)\mathrm{d}x - \int_{\overline{C^*}} f_1(x)\mathrm{d}x \\
&= \int_{\overline{C}\cap C^*} f_1(x)\mathrm{d}x - \int_{\overline{C^*}\cap C} f_1(x)\mathrm{d}x \\
&= \int_{\overline{C}\cap C^*} \underbrace{\frac{f_1(x)}{f_0(x)}}_{\le R \text{ on } \overline{C}} f_0(x)\mathrm{d}x - \int_{\overline{C^*}\cap C} \underbrace{\frac{f_1(x)}{f_(x)}}_{> R \text{ on } \overline{C}} f_0(x)\mathrm{d}x \\
&\le k\left[\int_{C\cap C^*} f_0(x)\mathrm{d}x - \int_{\overline{C^*}\cap C} f_0(x)\mathrm{d}x\right] \\
&= k\left[\int_{C^*} f_0(x)\mathrm{d}x - \int_{C} f_0(x)\mathrm{d}x\right] \\
&= k(\alpha^* - \alpha) \\
&\le 0
\end{aligned}$$

□

Start of lecture 9

Lemma. If C is a LRT with size α, and C^* is another test of size $\le \alpha$, then C is more powerful than C^*, i.e.

$$\beta = \mathbb{P}_{H_1}(x \notin C) \le \mathbb{P}_{H_1}(x \notin C^*) = \beta^*$$

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, σ_0^2 is known. Want the best size α test for H_0: $mu = \mu_0$, H_1: $\mu = \mu_1$ for some fixed $\mu_1 > \mu_0$

$$\begin{aligned}\Lambda_x(H_0; H_1) &= \frac{\cancel{(2\pi\sigma^2)^{-\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_1)^2\right)}{\cancel{(2\pi\sigma_0^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_0)^2\right)} \\ &= \exp\left(\frac{(\mu_1 - \mu_0)}{\sigma_0^2} n\overline{x} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma_0^2}\right)\end{aligned}$$

$\Lambda_x(H_0; H_1)$ is monotone increasing in $\overline{x} = \frac{1}{n}\sum x_i$. Hence, for any k, there is a c, such that $\Lambda_x(H_0; H_1) > k \iff \overline{x} > c$. Thus the LRT critical region is $\{x : \overline{x} > a\}$ for some constant c. By the same logic the LRT is of the form

$$C = \{\sqrt{n}\frac{(\overline{x} - \mu_0)}{\sigma_0} < c'\}$$

want to pick c' such that

$$\mathbb{P}_{H_0}\left(\sqrt{n}\frac{(\overline{x} - \mu_0)}{\sigma_0} > c'\right) = \alpha$$

But $\sqrt{n}\frac{(\overline{x}-\mu_0)}{\sigma_0} \sim \mathrm{N}(0, 1)$ (this is a pivot). So if we take $c' = \Phi^{-1}(1 - \alpha) \cdot z_\alpha$. Finally the LRT has critical region

$$\left\{x : \frac{\sqrt{n}(\overline{x} - \mu_0)}{\sigma_0} > z_\alpha\right\}$$

By N-D lemma, this is the most powerful test of size α. This is called a "z-test" because we use a z statistic $z = \sqrt{n}\left(\frac{\overline{x}-\mu_0}{\sigma_0}\right)$ to define the critical region.

P-value

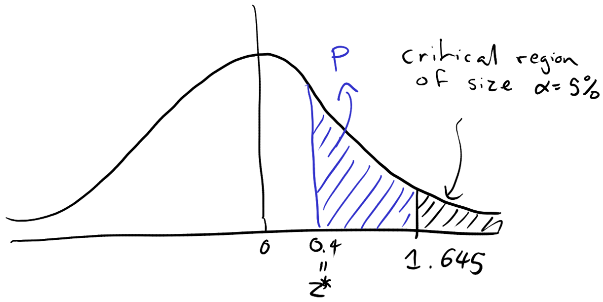
For any test with critical region of the form $\{x : T(x) > k\}$ for some statistic T, a *p-value* or observed significance level is

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

where x^* is the observed date. In example we just saw, let $\mu_0 = 5$, $\mu_1 = 6$, $\sigma_0 = 1$, $\alpha = 0.05$, observe

$$x^* = (5.1, 5.5, 4.9, 5.3)$$

$\overline{x^*} = 5.2$, $z^* = 0.4$. $z_\alpha = \Phi^{-1}(1 - \alpha) = 1.645$



Here, we fail to reject H_0: $\mu_0 = 5$, $p = 0.35$.

Proposition. Under H_0, p has a $\text{Unif}(0,1)$ distribution. p is a function of x^*; null distribution assumes $x^* \sim \mathbb{P}_{H_0}$.

Proof.

$$\mathbb{P}_{H_0}(p < u) = \mathbb{P}_{H_0}(1 - F(T) < u)$$

where F is the cdf of T.

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

for all $u \in [0,1]$. Thus $p \sim \text{Unif}(0,1)$. □

Composite Hypotheses

$X \sim f_X(\bullet \mid \theta)$, $\theta \in \Theta$. H_0: $\theta \in \Theta_0 \subset \Theta$, H_1: $\theta \in \Theta_1 \subset \Theta$. Type I, II error probabilities depend on the value of θ within Θ_0 or Θ_1 respectively. Let C be some critical region.

Definition (Power Function and UMP test)**.** The *power function* of the test C is

$$W(\theta) = \mathbb{P}_\theta(\underbrace{x \in C}_{H_0 \text{ rejected}})$$

The *size* of c is the worst case Type I error probability:

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

We say that C is *uniformly most powerful* (UMP) of size α for H_0 against H_1 if:

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

(2) For any other test C^* of size $\leq \alpha$, with power function W^*, we have $W(\theta) \geq W^*(\theta)$ for all $\theta \in \Theta_1$.

Note. UMP test need not exist. But, in some simple cases, the LRT is UMP.

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \text{N}(\mu, \sigma_0^2)$: σ_0^2 known. We wish to test H_0: $\mu \le \mu_0$ against H_1: $\mu > \mu_0$ for some fixed μ_0. We just studied the simple hypothesis:

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

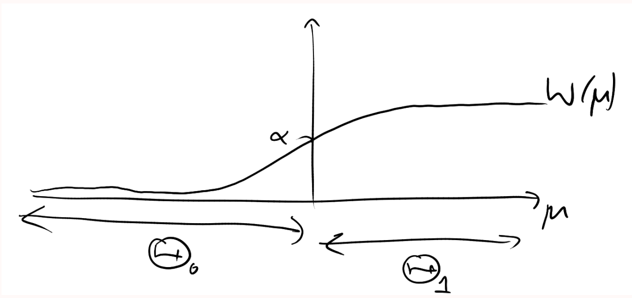
LRT was:

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

Claim: the same test C is UMP for H_0 against H_1. The power function for C is

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

This is monotone increasing in $\mu \in (-\infty, \infty)$



The test has size α as $\sup_{\mu \in \Theta_0} W(\mu) = \alpha$. It remains to show that if C^* is another test of size $\le \alpha$ with power function W^* then $W(\mu_1) \ge W^*(\mu_1)$ for all $\mu_1 > \mu_0$. Main observation: critical region only depends on μ_0. And C is the LRT for the simple hypothesis H_0': $\mu = \mu_0$, H_1': $\mu = \mu_1$. Any test C^* of H_0 vs H_1 of size $\le \alpha$ also has size $\le \alpha$ for H_0' vs H_1'.

$$W^*(\mu_0) \le \sup_{\mu \in \Theta_0} W^*(\mu) \le \alpha$$

Hence by N-D lemma, we know $W(\mu_1) \ge W(\mu_2)$. As we can apply this argument for any $\mu_1 > \mu_0$, we have

$$W^*(\mu_1) \le W(\mu_1) \qquad \forall \mu_1 > \mu_0$$

Start of lecture 10

Generalised Likelihood Ratio Tests

$X \sim f_X(\bullet \mid \theta)$, H_0: $\theta \in \Theta_0$, H_1: $\theta \in \Theta_1$. The generalised likelihood ratio statistic:

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

Large values of Λ_x indicate larger departure from H_0.

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, σ_0 is known. Wish to test H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$ for fixed μ_0. Here $\Theta_0 = \{\mu_0\}$, $\Theta_1 = \mathbb{R} \setminus \{\mu_0\}$. The GLR is

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

Taking $2 \cdot \log$ of Λ_x (monotone increasing transformation)

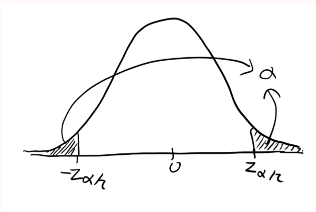
$$2 \log \Lambda x = \frac{n}{\sigma_0^2} (\overline{x} - \mu_0)^2$$

The GLR test rejects H_0 when Λ_x is large (or when $2 \log \Lambda_x$ is large), i.e. when

$$\left| \sqrt{n} \frac{(\overline{x} - \mu_0}{\sigma_0} \right|$$

is large. (Under H_0, the expression in the modulus has a $\mathrm{N}(0,1)$ distribution). For a test of size α, reject when

$$\left| \sqrt{n} \frac{(\overline{x} - \mu_0)}{\sigma_0} \right| > z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$



This is called a 2-sided test.

Note. $2 \log \Lambda_x = n \frac{(\overline{x} - \mu_0)}{\sigma_0^2} \sim \chi_1^2$ under H_0.

We can also define the critical region of the GLR test as

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

In general, we can approximate the distribution of $2\log\Lambda_x$ with a χ_1^2 distribution when n is large(!)

Wilks' Theorem

Suppose θ is k-dimensional $\theta = (\theta_1, \ldots, \theta_k)$. The dimension of a hypothesis H_0: $\theta \in \Theta_0$ is the number of "free parameters" in Θ_0.

(1) $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_1 = \theta_2 = \cdots = \theta_p = 0\}$ for some $p < k$. Here $\dim(\theta_0) = k - p$.

(2) Let $A \in \mathbb{R}^{p\times k}$, $b \in \mathbb{R}^p$, $p < k$¿

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

$\dim(\Theta_0) = k - p$ if rows of A are linearly independent (Θ_0 is a hyperplane).

(3) $\Theta_0 = \{\theta \in \mathbb{R}^k : \theta_0 = f_i(\phi), \phi \in \mathbb{R}^p\}$, $p < l$. Here ϕ are the free parameters; f_i need not be linear. Under regularity conditions $\dim(\theta_0) = p$.

Theorem (Wilk's Theorem)**.** Suppose $\Theta_0 \subset \Theta_1$ ("nested hypotheses")

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

If $X_1, \ldots, X_n$ are iid from $f_X(\bullet \mid \theta 0$, then as $n \to \infty$, the limiting distribution of $2\log\Lambda_x$ under H_0 is χ_p^2. That is, for any $\theta \in \Theta_0$, any $l > 0$,

$$\mathbb{P}_\theta(z\log\Lambda_x \leq l) \overset{n\to\infty}{\longrightarrow} \mathbb{P}(Z \leq l)$$

where $Z \sim \chi_p^2$.

How to use this? If we reject H_0 when $2\log\Lambda_x \geq \chi_p^2(\alpha)$ then when n is large, the size of the test is $\approx \alpha$. (!!!)

Example. In the two-sided normal mean test

$$\Theta_0 = \{\mu_0\}, \qquad \Theta_1 = \mathbb{R} \setminus \{\mu_0\}$$

we found $2\log\Lambda_x \sim \chi_1^2$. If we take $\Theta_1 = \mathbb{R}$, the GLR statistic doesn't change, so $2\log\Lambda_x \sim \chi_1^2$.

$$\dim(\theta_1) - \dim(\Theta_0) = 1 - 0 = 1$$

The prediction of Wilk's theorem is exact.

Proof. Wait for Part II Principles of Statistics :(□

Tests of goodness of fit

$X_1, \ldots, X_n$ are iid samples from a distribution on $\{1, 2, \ldots, k\}$. Let $p_i = \mathbb{P}(X_1 = i)$, let N_i be the number of observations equal to i. So,

$$\sum_{i=1}^{k} p_i = 1, \qquad \sum_{i=1}^{k} N_i = n$$

Goodness of fit test: H_0: $p = \tilde{p}$ for some fixed distribution $\tilde{p}$ on $\{1, \ldots, k\}$. H_1: p is *any* distribution with $\sum_{i=1}^{k} p_i = 1$, $p_i \geq 0$.

Example. Mendel crossed $n = 556$ smooth yellow peas with wrinkled green peas. Each member of the progeny can have any combination of the 2 features: SY, SG, WY, WG. Let (p_1, p_2, p_3, p_4) be the probabilities of each type, and $(N_1, \dots, N_4)$ are the number of progeny of each type, $\sum N_i = n = 556$.

Mendel's hypothesis:

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

Is there any evidence in $N_1, \dots, N_4$ to reject H_0? The model can be written $(N_1, \dots, N_k) \sim \text{Multinomial}(n; p_1, \dots, p_k)$. Likelihood: $L(p) \propto p_1^{N_1} \cdots p_k^{N_k}$

$$\implies l(p) = \text{const} + \sum_i N_i \log p_i$$

We can test H_0 against H_1 using a GLR test:

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

Since $\Theta_0 = \{\tilde{p}\}$, $\sup_{p \in \Theta_0} l(p) = l(\tilde{p})$. In the alternative p must satisfy $\sum p_i = 1$.

$$\sup_{p \in \Theta_1} l(p) = \sup_{p : \sum p_i = 1} \sum_i N_i \log p_i$$

Use Lagrangian $\mathcal{L}(p, \lambda) = \sum_i N_i \log p_i - \lambda \left(\sum_i p_i - 1\right)$. We find that $\hat{p}_i = \frac{N_i}{n}$ (the observed propoertion of samples of type i).

$$\begin{aligned} 2 \log \Lambda &= 2(l(\hat{p}) - l(\tilde{p})) \\ &= 2 \sum_i N_i \log \left(\frac{N_i}{n \cdot \tilde{p}_i} \right) \end{aligned}$$

Wilk's theorem tells us that $2 \log \Lambda_x$ is approximately χ^2_p with

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

So we can reject the H_0 with size $\approx \alpha$ when

$$2 \log \Lambda_x > \chi^2_{k-1}(\alpha)$$

Start of lecture 11

Tests of Goodness of fit and Independence

It's common to write

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

where $o_i = N_i$ "observed number of type i" and $e_i = n \cdot \tilde{p}_i$ "expected number of type i under null".

Pearson's statistic: Let $\delta_i = o_i - e_i$. Then

$$\begin{aligned} 2\log\Lambda &= 2\sum_i (e_i + \delta_i) \underbrace{\log\left(1 + \frac{\delta_i}{e_i}\right)}_{=\frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i^2} + O\left(\frac{\delta_i^3}{e_i^3}\right)} \\ &\approx 2\sum_i \left(\underbrace{\cancel{\delta_i}}_{\sum_i \delta_i = \sum_i (o_i - e_i) = n - n = 0} + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i} \right) \\ &= \sum \frac{\delta_i^2}{e_i} \\ &= \sum_i \frac{(o_i - e_i)^2}{e_i} \end{aligned}$$

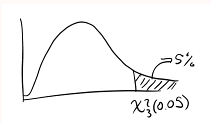
This is called Preason's statistic. This is also referred to a χ^2_{k-1} distribution when n is large.

Example. Mendel's data:

$$(n_1, n_2, n_3, n_4) = (315, 108, 102, 31)$$

$2\log\Lambda \approx 0.618$, $\sum_i \frac{(o_i - e_i)^2}{e_i} \approx 0.604$. We refer each statistic to a $\chi^2_{k-1} = \chi^2_3$ distribution.

$$\chi^2_3(0.05) = 7.815$$



We don't reject H_0 at size 5%. The p-value is $\mathbb{P}(\chi^2_3 > 0.6) \approx 0.96$. The data fir the null model almost too well.

Goodness of fit test for composite null

H_0: $p_i = p_i(\theta)$ for some parameter θ. H_1: p can be any distribution on $\{1, \dots, k\}$.

Example. Individuals can have 3 genotypes. H_0: $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta)$, $p_3 = (1-\theta)^2$, for some $\theta \in [0,1]$.

$$\begin{aligned} 2\log\Lambda &= 2\left(\sup_{p:\sum p_i=1} l(p) - \sup_\theta l(p(\theta))\right) \\ &= 2(l(\hat{p}) - l(p(\hat{\theta})) \end{aligned}$$

where $\hat{p}$ is the mle in the alternative H_1; $\hat{\theta}$ is the mle in null H_0. Last time we found $\hat{p}_i = \frac{N_i}{n}$. $\hat{\theta}$ would need to be computed for the null model in question.

$$\begin{aligned} 2\log\Lambda &= 2\sum_i N_i \log\left(\frac{N_i}{np_i(\hat{\theta})}\right) \\ &= 2\sum_i o_i \log\left(\frac{o_i}{e_i}\right) \end{aligned}$$

$o_i = N_i$ "observed number of type i", $e_i = n \cdot p_i(\hat{\theta})$ "expected number of type i under H_0". We can define a Pearson statistic $\sum_i \frac{(o_i-e_i)^2}{e_i}$ using the same argument as before.

Each statistic can be referred to a χ^2_d when n is large by Wilke's theorem.

$$\begin{aligned} d &= \dim(\Theta_1) - \dim(\Theta_0) \\ &= (k-1) - \dim(\Theta_0) \end{aligned}$$

Example. $l(\theta) = \sum_i N_i \log p_i(\theta) = 2N_1 \log\theta + N_2 \log(2\theta(1-\theta)) + 2N_3 \log(1-\theta)$. Maximising over $\theta \in [0,1]$ gives $\hat{\theta} = \frac{2N_1+N_2}{2n}$ (exercise). In this model $2\log\Lambda$ and $\sum_i \frac{(o_i-e_i)^2}{e_i}$ have a χ^2_d distribution with $d = (k-1) - \dim(\Theta_0) = (k-1) - 1 = k-2 = 3-2 = 1$.

Testing independence in contingency tables

$(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid with X_i taking values in $\{1, \ldots, r\}$, Y_i taking values in $\{1, \ldots, c\}$. The entries in a contingency table are

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

($\#$ samples of type (i, j))

Example. COVID-19 deaths. X_i: age of i-th death. Y_i: week on which it fell. Question: are deaths decreasing faster for older age grou that had been vaccinated?

Probability Model

We'll assume n is fixed. A sample (X_l, Y_l) has probability p_{ij} of falling in (i, j) entry of table.

$$(N_{11}, \dots, N_{1c}, N_{21}, \dots, N_{2c}, \dots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, \dots, p_{1c}, \dots, p_{rc})$$

> **Remark.** Fixing n may not be natural; we'll consider other models later.

Null hypothesis

Week of death is independent of age. X_i independent of Y_i for each sample. Let

$$p_{i+} = \sum_{j=1}^{n} p_{ij} \qquad p_{+j} = \sum_{i=1}^{r} p_{ij}$$

H_0: $p_{ij} = p_{i+}p_{+j}$. $(\mathbb{P}(X_l = i, Y_l = j) = \mathbb{P}(X_l = i)\mathbb{P}(Y_l = j))$. H_1: (p_{ij}) is unconstrained except for $p_{ij} \ge 0$, $\sum_{i,j} p_{ij} = 1$. The generalised LRT:

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

$o_{ij} = N_{ij}$, $e_{ij} = n\hat{p}_{ij}$, where $\hat{p}$ is the mle under independence model H_0. Using Lagrange multipliers we can find

$$\hat{p}_{ij} = \hat{p}_{i+}\hat{p}_{+j}$$

where

$$\hat{p}_{i+} = \frac{N_{i+}}{n} \qquad \hat{p}_{+j} = \frac{N_{+j}}{n}$$
$$N_{i+} = \sum_j N_{ij} \qquad N_{+j} = \sum_i N_{ij}$$

$$\implies 2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \log \left(\frac{N_{ij}}{n \cdot \hat{p}_{i+}\hat{p}_{+j}} \right) \approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Wilke's: The asymptotic distribution of these statistics is χ^2_d with

$$\begin{aligned} d &= \dim(\Theta_1) - \dim(\Theta_0) \\ &= (rc - 1) - [(r - 1) + (c - 1)] \\ (r-1)(c-1) \end{aligned}$$

($(r - 1)$ and $(c - 1) \to$ degrees of freedom in $(p_{1+}, \dots, p_{r+})$ and $(p_{+1}, \dots, p_{+c})$)

Start of lecture 12

Testing independence in contingency tables

N_{ij}: number of samples of type (i, j).

$$(N_{ij}) \sim \text{Multinomial}(n, (p_{ij}))$$

H_0: $p_{ij} = p_{i+} \times p_{+j}$
H_1: (p_{ij}) unconstrained.
Found $2 \log \Lambda$, which has asymptotic $\chi^2_{(r-1)(c-1)}$ distribution.

Example (COVID-19 deaths)**.** Problems with χ^2 independence test:

(1) χ^2 approximation can be bad when we have large tables. Rule of thumb: Need $N_{ij} \geq 5$ for all i, j.

Solution (non-examinable): exact testing. Idea: under H_0, the margins of N (N_{i+}), (N_{+j}) are sufficient statistics for p. therefore 2 tables N, $\tilde{N}$ with the same margins are equally likely under H_0. An exact test contrasts the test statistic observed $2 \log \Lambda(N)$ with the distribution of this statistic for the set of tables with the same margins as N. This gives a test of *exact* size α.

(2) $2 \log \Lambda$ can detect deviations from H_0 in any direction. $\implies$ Low power, especially when r, c is large. This is why H_0 is not rejected in a test of size 1% in COVID-19 example. Solutions:

   (1) Define a parametric alternative H_1 with fewer degrees of freedom.

   (2) Lump categories in the table.

Tests of Homogeneity

Instead of assuming $\sum_{i,j} N_{ij}$ fixed, we assume row totals are fixed.

Example. 150 patients, split into groups of 50 for placebo, half-dose, full-dose. We record whether each patient improved, showed no difference or got worse.

	I	N.D.	W
Placebo			
Half			
Full			

Now row totals are fixed. Null of homogeneity: probability of each outcome is the same in each treatment group.

Model:

$$(N_i 1, \dots, N_{ic}) \sim \text{Multinomial}(n_{i+}, p_{i1}, \dots, p_{ic})$$

independent for $i = 1, \dots, r$. Paramters satisfy $\sum_j p_{ij} = 1$ for all i. H_0: $p_{1j} = p_{2j} = \cdots p_{rj}$ for all $j = 1, \dots, c$. H_1: $(p_{i1}, \dots, p_{ic})$ is a probability vector for all i.

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

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

To find $2 \log \Lambda$ we need to maximise $l(p)$ over H_0, H_1. H_1: use Lagrange multipliers with constraints $\sum_j p_{ij} = 1$ for all i. Then the mle is

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

H_0: let $p_j = p_{1j} = \cdots = p_{+j}$.

$$l(p) = \text{const} + \sum_{j=1}^{c} N_{+j} \log p_j$$

hence the mle is $\hat{p}_j = \frac{N_{+j}}{n_{++}}$, $n_{++} = \sum_i n_{i+}$. Thus

$$2 \log \Lambda = 2 \sum_{i,j} N_{ij} \log \left(\frac{N_{ij}}{n_{i+} N_{+j} / n_{++}} \right)$$

This is exactly the same statistic as $2 \log \Lambda$ for the independence test. Let $o_{ij} = N_{ij}$, $e_{ij} = n_{i+} \hat{p}_j = n_{i+} \frac{N_{+j}}{n_{++}}$

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

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

This is also the same as Pearson's statistic for independence test.

Wilk's implies $2 \log \Lambda$ is approximately χ^2_d,

$$\begin{aligned} d &= \dim(\Theta_1) - \dim(\Theta_0) \\ &= (c-1)r - (c-1) \\ &= (c-1)(r-1) \end{aligned}$$

Asymptotic distribution of $2 \log \Lambda$ is also the same as in the independence test.

Testing independence or homogeneity with size α always has the same conclusion.

Relationship between tests and confidence sets

Define the *acceptance ragion* A of a test to be the complement of the critical region. Let $X \sim f_X(\bullet \mid \theta)$ for some $\theta \in \Theta$.

Theorem. (1) Suppose that for each $\theta_0 \in \Theta$ there is a test of H_0: $\theta = \theta_0$ of size α with acceptance region $A(\theta_0)$. Then, the set

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

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

(2) Suppose $I(X)$ 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 size α test for H_0: $\theta = \theta_0$.

Proof. In each part:

$$\theta_0 \in I(X) \iff X \in A(\theta_0)$$

For part (1), we calculate:

$$\begin{aligned} \mathbb{P}_{\theta_0}(I(X) \ni \theta_0) &= \mathbb{P}_{\theta_0}(x \in A(\theta_0)) \\ &= 1 - \mathbb{P}_{\theta_0}(x \in C(\theta_0)) \\ &= 1 - \alpha \end{aligned}$$

as desired. For part (2):

$$\begin{aligned} \mathbb{P}_{\theta_0}(X \in C(\theta_0)) &= \mathbb{P}_{\theta_0}(X \notin A(\theta_0)) \\ &= \mathbb{P}_{\theta_0}(I(X) \not\ni \theta_0) \\ &= 1 - \mathbb{P}_{\theta_0}(I(x) \ni \theta_0) \\ &= 1 - (1-\alpha) \\ &= \alpha \end{aligned}$$

as desired. □

Example. $X_1, \dots, X_n \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, σ^2 known.

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

confidence interval. Test: H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$. Critical region:

$$\left\{x : \left|\sqrt{n}\frac{(X - \overline{X}}{\sigma_0}\right| > z_{\alpha/2}\right\}$$

Start of lecture 13

Multivariate Normal Theory

Recall: if X is a random vector, then

$$\begin{aligned}\mathbb{E}[AX + b] &= A\mathbb{E}X + b \\ \mathrm{Var}(AX + b) &= A\,\mathrm{Var}(X)A^\top\end{aligned}$$

Definition. We say X has a multivariate normal distribution if for any $t \in \mathbb{R}^n$, $t^\top X$ is normal.

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

Proof. Say $AX + b$ is in $\mathbb{R}^m$. Take $t \in \mathbb{R}^m$.

$$t^\top(X + b) = (A^\top t)^\top X + t^\top b$$

Since X is MVN, $A^\top t)^\top X$ is a normal distribution, and since $t^\top b$ is a constant, this means that $t^\top(AX + b)$ is normal. □

Proposition. A MVN distribution is fully specified by its mean and variance.

Proof. Take X_1, X_2 both MVN with mean μ and variance Σ. We'll show that their mgf's are equal, hence X_1 and X_2 have the same distribution.

$$\begin{aligned}\mathbb{E}e^{1 \cdot t^\top X_1} &= M_{t^\top X_1}(1) && t^\top X_1 \text{ is Normal} \\ &= \exp\left(1 \cdot \mathbb{E}(t^\top X_1) + \frac{1}{2}\mathrm{Var}(t^\top X_1) \cdot 1^2\right) \\ &= \exp\left(t^\top \mu + \frac{1}{2}t^\top \Sigma t\right)\end{aligned}$$

This just depends on μ, Σ, so it is the same for X_1, X_2. □

Orthogonal projections

Definition. (1) We say $P \in \mathbb{R}^{n\times n}$ is an *orthogonal projection* if it is:

- Idempotent: $PP = P$.
- Symmetric: $P^\top = P$.

(2) Or equivalently, $P \in \mathbb{R}^{n\times n}$ is an *orthogonal projection* if for any $v \in \operatorname{col}(P)$, $Pv = v$, and for any $w \in \operatorname{col}(P)^\perp$, $Pw = 0$.

Proposition. (1) and (2) are equivalent.

Proof. (1) $\implies$ (2) Take $v \in \operatorname{col}(P)$, so $v = Pa$ for some $a \in \mathbb{R}^n$. Then

$$Pv = PPa = Pa = v$$

Take $w \in \operatorname{col}(P)^\perp$. Then $P^\top w = 0$. Hence

$$Pw = P^\top w = 0$$

(2) $\implies$ (1) We can write any $a \in \mathbb{R}^n$ uniquely as $a = v + w$, $w \in \operatorname{col}(P)^\perp$, $v \in \operatorname{col}(P)$. Then

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

As a was arbitrary, $P = P^2$. For symmetry, take $u_1, u_2 \in \mathbb{R}^n$. Then

$$\underbrace{(Pu_1)}_{\in \operatorname{col}(P)}^\top \underbrace{((I-P)u_2)}_{\in \operatorname{col}(P)^\perp} = 0$$

$\implies u_1^\top (P^\top - P^\top P)u_2 = 0$. Since this holds for all $u_1, u_2 \in \mathbb{R}^n$, $P^\top = P^\top P$. But $P^\top P$ is symmetric, hence $P^\top$ is symmetric, hence P symmetric. □

Corollary. If P is orthogonal projection, then $I - P$ is as well.

Proof.

$$(I-P)^\top = I - P^\top = I - P$$

and

$$(I-P)(I-P) = I - 2P + PP = I - P$$ □

Proposition. If $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection then

$$P = UU^\top$$

where the columns of U form an orthogonal basis for $\mathrm{col}(P)$. (if $k = \mathrm{rank}(P)$, then $U \in \mathbb{R}^{n\times k}$).

Proof. $UU^\top$ is cleraly symmetric and also idempotent

$$U \underbrace{U^\top U}_{I_k} U^\top = UU^\top$$

So $UU^\top$ is an orthogonal projection. To show it is equal to P, note $\mathrm{col}(P) = \mathrm{col}(UU^\top)$ by construction. □

Corollary.

$$k = \mathrm{rank}(P) = \mathrm{Tr}(\underbrace{U^\top U}_{I_k}) = \mathrm{Tr}(UU^\top) = \mathrm{Tr}(P)$$

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

(1) $PX \sim \mathrm{N}(9, \sigma^2 P)$, $(I-P)X \sim \mathrm{N}(0, \sigma^2(I-P))$, PX, $(I-P)X$ independent.

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

Proof. The vector

$$\begin{pmatrix} P \\ I-P \end{pmatrix} X$$

is MVN, because it is a linear function of X. The distribution is specified by the mean and variance:

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

and:

$$\begin{aligned} \mathrm{Var}\begin{pmatrix} PX \\ (I-P)X \end{pmatrix} &= \begin{pmatrix} P \\ I-P \end{pmatrix} \mathrm{Var}(X) \begin{pmatrix} P \\ I-P \end{pmatrix}^\top \\ &= \begin{pmatrix} P \\ I-P \end{pmatrix} \sigma^2 I \begin{pmatrix} P \\ I-P \end{pmatrix}^\top \\ &= \sigma^2 \begin{bmatrix} P & \cancel{P(I-P)} \\ \cancel{(I-P)P} & I-P \end{bmatrix} \end{aligned}$$

Let $Z \sim \mathrm{N}(0, \sigma^2 P)$, $Z' \sim \mathrm{N}(0, \sigma^2(I-P))$, Z, Z' independent. Then

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

So

$$\begin{pmatrix} PX \\ (I-P)X \end{pmatrix} \stackrel{d}{=} \begin{pmatrix} Z \\ Z' \end{pmatrix}$$

hence PX, $(I-P)X$ independent. This proves (1).

For (2):

$$\frac{\|PX\|^2}{\sigma^2} = \frac{(PX)^\top PX}{\sigma^2} = \frac{X^\top (UU^\top)^\top UU^\top X}{\sigma^2} = \frac{X^\top UU^\top X}{\sigma^2}$$

Cols of U form orthogonal basis for $\mathrm{col}(P)$

$$\implies \frac{\|PX\|^2}{\sigma^2} = \frac{\|U^\top X\|^2}{\sigma^2} = \sum_{i=1}^{\mathrm{rank}(P)} \frac{(U^\top X)_i^2}{\sigma^2}$$

But $U^\top X \sim \mathrm{N}(0, \sigma^2 I)$

$$\mathrm{Var}(U^\top X) = U^\top \mathrm{Var}(X) U = \sigma^2 U^\top U = \sigma^2 I$$

Therefore $(U^\top X)_i$, $i = 1, \ldots, \mathrm{rank}(P)$ are IID $\mathrm{N}(0, \sigma^2)$

$$\implies \frac{(U^\top X)_i}{\sigma} \stackrel{\text{IID}}{\sim} \mathrm{N}(0,1)$$

Hence $\frac{\|PX\|^2}{\sigma^2}$ is the sum of $\mathrm{rank}(P)$ squared independent $\mathrm{N}(0,1)$ variables, i.e. $\chi^2_{\mathrm{rank}(P)}$. □

Application

$X_1, \ldots, X_n \stackrel{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma^2)$. Both μ, σ^2 unknown. Recall that the mle for μ is $\overline{X} = \frac{1}{n} \sum X_i$. The mle for σ^2 is $\hat{\sigma}^2 = \frac{S_{XX}}{n}$, where $S_{XX} = \sum_i (X_i - \overline{X})^2$.

Theorem. (i) $\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n)$

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

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

Proof. Let $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^n$. Let $P = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be an orthogonal projection onto $\mathrm{span}(\mathbf{1})$. Easy to check that $P = P^\top = P^2$. We can write

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \mu\mathbf{1} + \varepsilon$$

where $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$. Note:

- $\overline{X}$ is a function of PX
$$PX = \mu \mathbf{1} + P\varepsilon$$
because $\overline{X} = (PX)_1$. In particular, $\overline{X}$ is function of $P\varepsilon$.

- $$\begin{aligned} S_{XX} &= \sum_i (X_i - \overline{X})^2 \\ &= \|X - \mathbf{1}\overline{X}\|^2 \\ &= \|(I-P)X\|^2 \\ &= \|(I-P)\varepsilon\|^2 \end{aligned}$$
so S_{XX} is a function of $(I-P)\varepsilon$. By previous theorem, $P_\varepsilon \perp\!\!\!\perp (I-P)\varepsilon$. Hence $\overline{X} \perp\!\!\!\perp S_{XX}$. Part (i) we've shown before. Also,
$$\frac{S_{XX}}{\sigma^2} = \frac{\|(I-P)\varepsilon\|^2}{\sigma^2} \sim \chi^2_{\underbrace{\mathrm{Tr}(I-P)}_{n-1}} \qquad \square$$

Start of lecture 14

0.6 The linear Model

Data are pairs (x_1, Y_1), ..., (x_n, Y_n). $Y_i \in \mathbb{R}$: "responses", random. $x_i \in \mathbb{R}^p$: "predictors", fixed.

Example. Y_i: number of insurance claims for client i. x_i: (age, number of claims in 2-21, years with driver's license, ...).

In a linear model, we assume
$$Y_i = \not{\alpha} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- α is an intercept.
- $\beta_1, \ldots, \beta_p$ are coefficients.
- $\varepsilon_1, \ldots, \varepsilon_n$ are random noise variables.

Remark. We normally remove intercept by including a dummy predictor which is equal to 1 for all i, i.e. $x_{i1} = 1$ for all $i = 1, \ldots, n$.

Remark. We can also model non-linear relationships between Y_i and x_i using a linear model, for example by using $x_i = (\text{age}, \text{age}^2, \log(\text{age}))$.

Remark. β_j is the effect on Y_i of increasing x_{ij} by a unit, whilst keeping all other predictors constant. Estimates of β should not be interpreted causally, unless we have a randomised experiment.

Matrix formulation:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots x_{2p} & \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}}_{\text{“design matrix”}}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

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

Moment assumptions on ε:

(1) $\mathbb{E}\varepsilon = 0 \implies \mathbb{E}Y = X\beta$.

(2) $\operatorname{Var}\varepsilon = \sigma^2 I \implies \operatorname{Var}(\varepsilon_i) = \sigma^2$ for all i “homoscedasticity”. $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

We'll assume throughout that $x \in \mathbb{R}^{k \times p}$ has full rank. In particular, $p \leq n$ (more samples than predictors).

Least squares estimator

$\hat{\beta}$ minimises the residual sum of squares

$$\begin{aligned} S(\beta) &= \|Y - X\beta\|^2 \\ &= \sum_{i=1}^{n} (Y_i - x_i^\top \beta)^2 \end{aligned}$$

This is a quadratic (positive definite) polynomial in β so $\hat{\beta}$ satisfies

$$\nabla S(\beta)|_{\beta = \hat{\beta}} = 0$$

$$\implies \left.\frac{\partial S(\beta)}{\partial \beta_k}\right|_{\beta=\hat{\beta}} = -2\sum_{i=1}^n x_{ik}\left(Y_i - \sum_{j=1}^p x_{ij}\hat{\beta}_j\right) = 0$$

for each $k = 1, \dots, p$. Equivalent matrix form:

$$X^\top X\hat{\beta} = X^\top Y$$

As X has rank p, the matrix $X^\top X \in \mathbb{R}^{p\times p}$ is invertible, hence

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

(linear in Y!). Check:

$$\begin{aligned}\mathbb{E}\hat{\beta} &= \mathbb{E}[(X^\top X)^{-1}X^\top Y] \\ &= (X^\top X)^{-1}X^\top \mathbb{E}Y \\ &= \cancel{(X^\top X)^{-1}X^\top X}\beta \\ &= \beta\end{aligned}$$

Hence $\hat{\beta}$ is unbiased. We can also calculate:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}((X^\top X)^{-1}X^\top Y) \\ &= (X^\top X)^{-1}X^\top \mathrm{Var}(Y)X(X^\top X)^{-1} \\ &= (X^\top X)^{-1}X^\top \sigma^2 I X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}\end{aligned}$$

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

$$\mathrm{Var}(t^\top \hat{\beta}) \le \mathrm{Var}(t^\top \beta^*)$$

We say $\hat{\beta}$ is "Best Linear Unbiased Estimator" (BLUE).

Remark. Think of $t \in \mathbb{R}^p$ as the value of the predictors for a new sample. Then $t^\top\hat{\beta}$, $t^\top\beta^*$ are estimators of the mean response. These are both unbiased, so the mse is the variance of $t^\top\hat{\beta}$, $t^\top\beta^*$. Theorem says variance is "best" using the least squares estimator.

Proof.

$$\mathrm{Var}(t^\top\beta^*) - \mathrm{Var}(t^\top\hat{\beta}) = t^\top(\mathrm{Var}\,\beta^* - \mathrm{Var}\,\hat{\beta})t \ge 0$$

This holds for all $t \in \mathbb{R}^p$ if and only if the matrix $\operatorname{Var}\beta^* - \operatorname{Var}\hat{\beta}$ is positive semi-definite. Recall $\beta^* = CY$, $\hat{\beta} = (X^\top X)^{-1}X^\top Y$. Let $A = C - (X^\top X)^{-1}X^\top$. Note:

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

(since β^* and $\hat{\beta}$ are unbiased). But also note

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

for all $\beta \in \mathbb{R}^p$, so we must have $AX = 0$. Then

$$\begin{aligned}
\operatorname{Var}\beta^* &= \operatorname{Var}((A + (X^\top X)^{-1}X^\top)Y) \\
&= (A + (X^\top X)^{-1}X^\top)\operatorname{Var}Y(A + (X^\top X)^{-1}X^\top)^\top \\
&= \sigma^2(AA^\top + (X^\top X)^{-1} + \cancel{AX(X^\top X)^{-1}} + \cancel{(X^\top X)^{-1}X^\top A^\top}) \\
&= \sigma^2 AA^\top + \operatorname{Var}(\hat{\beta}) \\
\implies \operatorname{Var}\beta^* - \operatorname{Var}\hat{\beta} &= \sigma^2 AA^\top
\end{aligned}$$

and this is positive definite, as desired. □

Fitted values and residuals: fitted values

$$\hat{Y} = X\hat{B} = \underbrace{X(X^\top X)^{-1}X^\top}_{P \text{ “hat matrix”}} Y$$

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

Proposition. P is the orthogonal projection onto $\operatorname{col}(X)$.

Proof. P is clearly symmetric. Also,

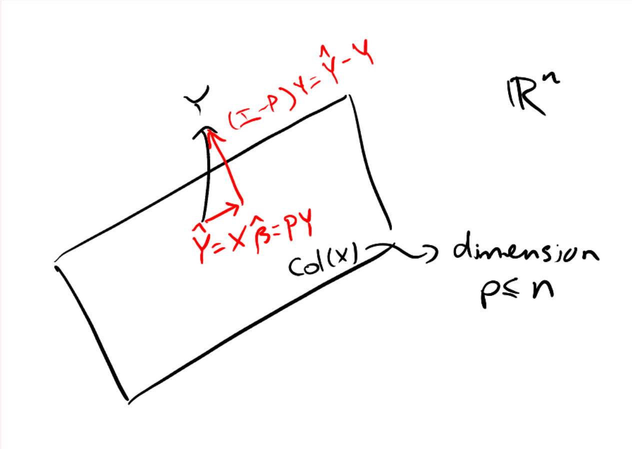
$$P^2 = X(X^\top X)^{-1}\cancel{X^\top X(X^\top X)^{-1}}X^\top = P$$

Therefore P is an orthogonal projection onto $\operatorname{col}(P)$. We need to show $\operatorname{col}(P) = \operatorname{col}(X)$. For any a, $Pa = X[(X^\top X)^{-1}X^\top a] \in \operatorname{col}(X)$. Also, if $b = Xc$ is a vector in $\operatorname{col}(X)$, then

$$b = Xc = X(X^\top X)^{-1}X^\top Xc = Pb \in \operatorname{col}(P)$$

□

Corollary. Fitted values are projections of Y onto $\mathrm{col}(X)$. Residuals are projections of Y onto $\mathrm{col}(X)^\perp$.



Normal assumptions

We assume in addition to $\mathbb{E}\varepsilon = 0$, $\mathrm{Var}\,\varepsilon = \sigma^2 I$, that ε is MVN, i.e.

$$\varepsilon \sim \mathrm{N}(0, \sigma^2 I_n)$$

σ^2 is usually unknown, so the parameters in the model are (β, σ^2). We'll see that mle of β is the least squares estimator $\hat{\beta}$.

Start of lecture 15

Normal linear model

Take $Y = XB + \varepsilon$, $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$. MLE: 2 parameters: $\beta \in \mathbb{R}^p$, $\sigma^2 \in \mathbb{R}_+$. Log-likelihood:

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

For any $\sigma^2 > 0$, we can see that $l(\beta, \sigma^2)$ is maximised as a function of β at the minimiser of $\|Y - XB\|^2$, i.e. the least squares estimator $\hat{\beta}$. Now find:

$$\arg\max_{\sigma^2 \geq 0} l(\hat{\beta}, \sigma^2)$$

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

As $\sigma^2 \mapsto l(\hat{\beta}, \sigma^2)$ is concave, there is unique maximiser where $\frac{\partial l(\hat{\beta}, \sigma^2)}{\partial \sigma^2} = 0$

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

Theorem. (1) $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(X^\top X)^{-1})$

(2) $\frac{\hat{\sigma}^2}{\sigma^2} n \sim \chi^2_{n-p}$

(3) $\hat{\beta}$, $\hat{\sigma}^2$ are independent(!)

Proof. $\hat{\beta}$ is linear in Y, hence MVN. We already know $\mathbb{E}\hat{\beta} = \beta$, $\operatorname{Var}\hat{\beta} = \sigma^2(X^\top X)^{-1}$. This proves (1). For (2) note

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

$\mathrm{rank}(I-P) = \mathrm{Tr}(I-P) = n-p$. ($X \in \mathbb{R}^{n-p}$ has full rank).

For (3), note $\hat{\sigma}^2$ is a function of $(I-P)\varepsilon$. We'll show that $\hat{\beta}$ is a function of $P\varepsilon$, which implies $\hat{\sigma}^2 \perp\!\!\!\perp \hat{\beta}$ since $P\varepsilon \perp\!\!\!\perp (I-P)\varepsilon$.

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

since $X^\top P = X^\top$. □

Corollary. $\hat{\sigma}^2$ is biased

$$\mathbb{E}\frac{\hat{\sigma}^2 n}{\sigma^2} = n-p \implies \mathbb{E}\hat{\sigma}^2 = \left(\frac{n-p}{n}\right)\sigma^2$$

Student's t-distribution

If $U \sim \mathrm{N}(0,1)$, $V \sim \chi^2_n$, $U \perp\!\!\!\perp V$ then we say $T = \frac{U}{\sqrt{V/n}}$ has a t_n distribution.

The F distribution

If $V \sim \chi^2_n$, $W \sim \chi^2_n$, $V \perp\!\!\!\perp W$ then we say

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

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

Confidence sets for β

Suppose we want a $100(1-\alpha)\%$ confidence interval for one of the coefficients (WLOG take β_1). Note:

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

because $\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2 (X^\top X)^{-1}_{11})$. Also,

$$\frac{\hat{\sigma}^2}{\sigma^2} n \sim \chi^2_{n-p}$$

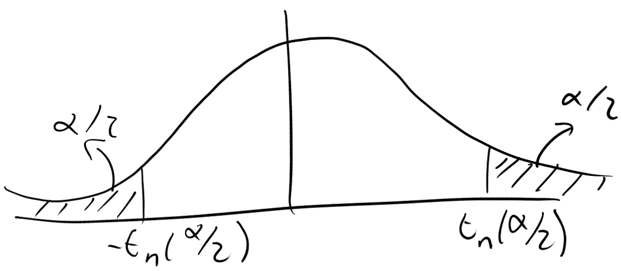
and these two statistics are independent.

$$\implies \frac{\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\cancel{\sigma^2} (X^\top X)^{-1}_{11}}}}{\sqrt{\frac{\hat{\sigma}^2}{\cancel{\sigma^2}} \frac{n}{n-p}}} \sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p}$$

Now this only depends on β_1 and *not* on σ^2, so we can use this as a pivot.

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

We use that t_n distribution is symmetric around 0.



Rearranging the inequalities, we get

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

We conclude that

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

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

Remark. This is *not* asymptotic.

By the duality between tests of significance and confidence intervals, we can find a size α test for H_0: $\beta_1 = \beta^*$ vs H_1: $\beta_1 \neq \beta^*$. Simply reject H_0 if β^* is not contained in the $100 \cdot (1-\alpha)\%$ confidence interval for β_1.

Confidence ellipsoids for β

Note $\hat{\beta} - \beta \sim \mathrm{N}(0, \sigma^2(X^\top X)^{-1})$. As X has full rank, $X^\top X$ is positive definite. So it has eigendecomposition

$$(X^\top X) = UDU^\top$$

where $D_{ii} > 0$ for $i = 1, \dots, p$. Define

$$(X^\top X)^\alpha = UD^\alpha U^\top$$

$$D^\alpha = \begin{pmatrix} D_{11}^\alpha & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & D_{pp}^\alpha \end{pmatrix}$$

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

Hence

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

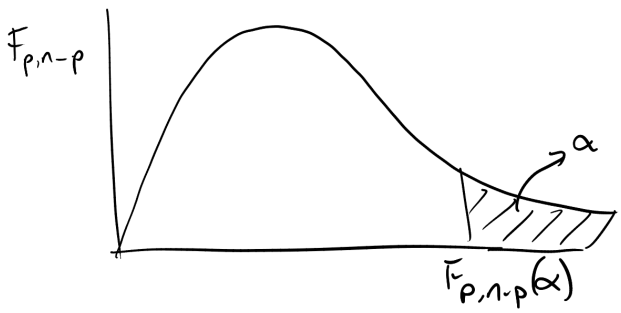
This is a function of $\hat{\beta}$, so it's independent of

$$\frac{\hat{\sigma}^2 n}{\sigma^2} \sim \chi_{n-p}^2$$

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

This only depends on β, *not* on σ^2, so it can be used as a pivot. For all β, σ^2:

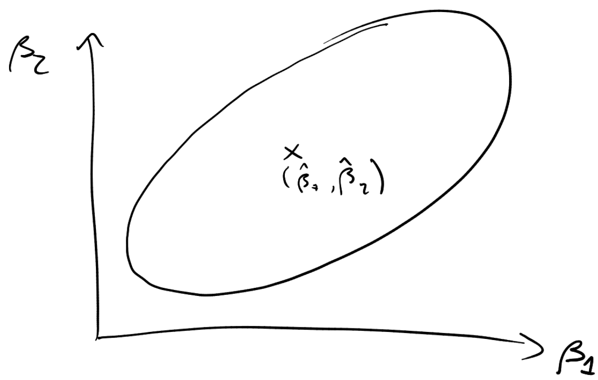
$$\mathbb{P}_{\sigma^2,\beta}\left(\frac{\|X(\hat{\beta} - \beta\|^2/p}{\hat{\sigma}^2 n/(n-p)} \leq F_{p,n-p}(\alpha)\right) = 1 - \alpha$$



So, we can say that the set

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

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



Principal axes are given by eigenvectors of $(X^\top X)$.

In the next section we'll talk about hypothesis tests for H_0: $\beta_1 = \cdots = \beta_p = 0$, H_1: $\beta \in \mathbb{R}^p$.

Start of lecture 16

The F-test

$Y = X\beta + \varepsilon$, $\varepsilon \sim \mathrm{N}(0, \sigma^2 I)$. H_0: $\beta_1 = \beta_2 = \cdots = \beta_{p_0} = 0$. H_1: $\beta \in \mathbb{R}^p$. Let $X = (x_0, x_1)$ (X_0 is $n \times p_0$ and X_1 is $n \times (p - p_0)$)

$$\beta = \begin{pmatrix} \beta^0 \\ \beta^1 \end{pmatrix} \qquad \beta^0 = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p_0} \end{pmatrix} \qquad \beta^1 = \begin{pmatrix} \beta_{p_0+1} \\ \vdots \\ \beta_p \end{pmatrix}$$

Null: $\beta^0 = 0$. This is a normal linear model:

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

Write $P = X(X^\top X)^{-1}X^\top$, $P_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$. As X, P have full rank, so do X_1, P_1. Recall that the maximum log-likelihood in a linear model is

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

The generalised log likelihood ratio statistic is

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

This is a monotone increasing function in

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

(The cancel takes place because the columns of $P - P_1$ are in $\mathrm{col}(X)$). This is monotone increasing in

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

"F statistic".

Lemma. $P - P_1$ is an orthogonal projection with rank p_0.

Proof. $P - P_1$ is symmetric as both P and P_1 are

$$(P-P_1)(P-P_1) = P + P_1 - 2\underbrace{PP_1}_{=P_2} = P - P_1$$

$$\begin{aligned}\mathrm{rank}(P-P_1) &= \mathrm{Tr}(P-P_1)\\ &= \mathrm{Tr}(P) - \mathrm{Tr}(P_1)\\ &= p - (p-p_0)\\ &= p_0\end{aligned}$$

□

To recap the generalised LRT rejects H_0 when F is large. What is the null distribution of F? Under H_0:

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

Therefore, under H_0:

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

with numerator $\sim \left(\frac{\chi^2_{p_0}}{p_0}\right)$ and denominator $\sim \left(\frac{\chi^2_{n-p}}{n-p}\right)$. Furthermore,

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

is MVN with $\mathrm{Cov}((P - P_1)\varepsilon, (I - P)\varepsilon) = \sigma^2(P - P - 1)(I - P) = 0$. Hence $(P - P_1)\varepsilon \perp \perp (I - P)\varepsilon$. Hence numerator $\perp\!\!\!\perp$ denominator in F. We conclude that

$$F \sim F_{p_0,n-p},$$

so the test rejects H_0 with size α if

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

Last time we derived a size α test for H_0: $\beta_1 = 0$ using the $100 \cdot (1 - \alpha)\%$ confidence interval for β_1. That test rejects H_0 when

$$|\beta_1| > t_{n-p}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2 n (X^\top X)^{-1}_{11}}{n - p}}$$

Lemma. This test is equivalent to the F-test with $p_0 = 1$.

Proof. Exercise. □

Categorical predictors

Example. $Y_i \in \mathbb{R}$: clinical response, $z_i \in \{\text{control}, \text{treatment 1}, \text{treatment 2}\}$.

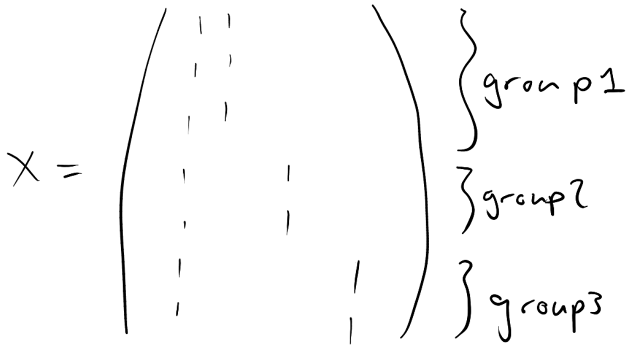
Let

$$x_{i,j} = \mathbb{1}_{\{z_i = j\}} = \mathbb{1}_{\{\text{subject } i \text{ was in group } j\}}$$

$x_i \in \mathbb{R}^3$ this is numerical.

$$Y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

Problem:



This has rank $3 < 4$. Corner point constraint: call one of the groups the "baseline" and remove it from the linear model. Interpretation of β_j depends on baseline. β_j is effect of being in group j *relative* to baseline. β_j is effect of being in group j *relative* to baseline. However, $\mathrm{col}(X)$ and matrix P are insensitive of choice of baseline, and therefore so are the fitted values

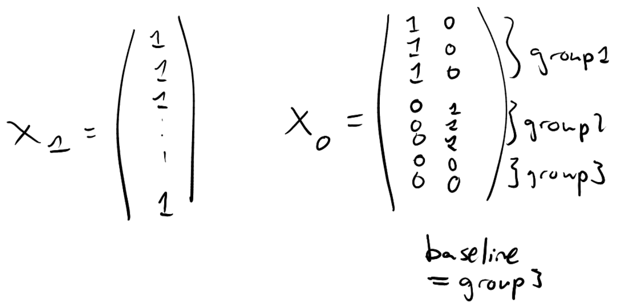
$$\hat{Y} = PY.$$

This can be extended to a model with more than 1 categorical predictor, for example group and gender.

ANOVA: Analysis of Variance. The F-test for

- H_0: $\beta_j = 0$ for a categorical predictor $\alpha \neq 0$.
- H_1: $\begin{pmatrix} \alpha_1 \\ \beta \end{pmatrix} \in \mathbb{R}^3$.

In this case, we can write the F statistic in a simpler way.



P_1 projection onto constant vectors.

$$P_1 = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

P = projection onto vectors which are constant for each group

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

$$P_1 Y = \begin{pmatrix} \overline{Y} \\ \overline{Y} \\ \vdots \\ \overline{Y} \end{pmatrix} \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$Py = \begin{pmatrix} \overline{Y}_1 \\ \overline{Y}_1 \\ \vdots \\ \overline{Y}_2 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_3 \\ \overline{Y}_3 \end{pmatrix} \qquad \overline{Y}_j = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}_{\{z_i=j\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{z_i=j\}}} = \text{average response for group } j$$

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

Assume all groups of size N ($n = 3N$). Numerator is variance between groups, denominator is variance within groups.