L1.1 www.statslab.com.ac.uk/~sb2116

STATISTICS The "science" of processing quantitative evidence to make judgements or decisions. It can include

- Design of experiments
- Exploring data using graphics
- Quantifying evidence for or against a hypothesis
- Clear communication of evidence and uncertainty

Def A parametric family is a family of probability distribution indexed by a scalar or vector parameter θ in some parameter space Θ. Ex Family of $N(\mu, \sigma^2)$ distributions, $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R}^2$

In this course we will talk about formal inference of θ, and how to test hypothesis about θ (e.g. $H: \theta = 0$)

Probability review

Let Ω be a sample space of all possible outcomes of an experiment. An event is a measurable subset of Ω. Let $\mathcal{F}$ be the set of all events. A probability measure $\mathbb{P}: \mathcal{F} \to [0,1]$ satisfies

① $\mathbb{P}(\emptyset) = 0$ ② $\mathbb{P}(\Omega) = 1$ ③ $\mathbb{P}(\bigsqcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

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

Ex 2 coin tosses, $\Omega = \{HH, HT, TH, TT\}$

$X(\omega) = \#$ heads in ω

The distribution function of X is

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

We'll use f_X to denote either

① The probability mass f^n of X if X is discrete, $f_X(x) = \mathbb{P}(X = x)$

② The pdf of X if X is continuous, i.e. $F_X(x) = \int_{-\infty}^{x} f(t)\,dt$

The expectation of X is

$$\mathbb{E}X = \begin{cases} \sum x f_X(x) & \text{if } X \text{ discrete} \\ \int_{-\infty}^{\infty} x f_X(x)\,dx & \text{if } X \text{ continuous} \end{cases}$$

L1.2

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

Linearity of expectation $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$, $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$

We say that $X_1, \dots, X_n$ are independent if

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

If $X_1, \dots, X_n$ are independent with pdfs $(f_{X_1}, \dots, f_{X_n})$ or pmfs then the joint pdf or pmf for the random vector $\underline{X} = (X_1, \dots, X_n)$ is

$$f_{\underline{X}}(\underline{x}) = \prod f_{X_i}(x_i)$$

For independent $X_1, \dots, X_n$,

$$\mathrm{Var}(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) = a_1^2 \mathrm{Var}(X_1) + \cdots + a_n^2 \mathrm{Var}(X_n)$$

Standardisation Let $X_1, \dots, X_n$ be iid with $\mathbb{E}X_1 = \mu$, $\mathrm{Var}X_1 = \sigma^2$

We usually write $S_n = \sum_{i=1}^{n} X_i$. By linearity $\mathbb{E}S_n = n\mu$, $\mathrm{Var}\, S_n = n\sigma^2$

The empirical mean $\bar{X}_n = S_n/n$ has $\mathbb{E}\bar{X}_n = \mu$, $\mathrm{Var}\,\bar{X}_n = \sigma^2/n$

A standardised statistic $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$

has $\mathbb{E}Z_n = 0$, $\mathrm{Var}\, Z_n = 1$.

Convergence

Weak law of large numbers: $\forall \varepsilon > 0$, $\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \to 0$ as $n \to \infty$

Strong law of large numbers: $\mathbb{P}(\bar{X}_n \to \mu) = 1$

Cheese Lettuce Bacon: Z_n is approximately $N(0,1)$ when n is large

Maxima (or minima) of iid rvs If $X_1, \dots, X_n$ are iid, $Y = \max(X_1, \dots, X_n)$ then

$$F_Y(y) = \mathbb{P}(Y \le y) = \mathbb{P}(X_1 \le y, \dots, X_n \le y)$$
$$= (F_{X_1}(y))^n$$

Can derive pdf or pmf by differentiation or differencing.

Moment generating function $M_X(t) = \mathbb{E}(e^{tX})$ provided the expectation exists in some neighbourhood of 0.

We can use this to find moments, $\mathbb{E}(X^n) = M_X^{(n)}(0)$.

Also useful to find distro of a sum of rvs, because under broad conditions, $M_X = M_Y \Rightarrow F_X = F_Y$

Ex $X_1, \dots, X_n \overset{iid}{\sim}$ Poisson (μ) Distribution of S_n?

- $$M_{S_n}(t) = \mathbb{E}(e^{tS_n}) = \mathbb{E}(e^{t(X_1+\cdots+X_n)})$$
$$= \prod_{i=1}^{n} \mathbb{E}(e^{tX_i}) = (\mathbb{E}e^{tX_1})^n$$
$$= M_{X_1}(t)^n = e^{-n\mu(1-e^t)}$$

$\Rightarrow S_n \sim$ Poisson $(n\mu)$

Conditioning If X, Y have a joint pdf (or pmf) $f_{X,Y}$, then the marginal pdf of X is

- $$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$$
$$\left(f_X(x) = \sum_y f_{X,Y}(x,y) \right)$$

The conditional pdf of X given $Y=y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{if } f_Y(y) \neq 0$$

Conditional expectation $\mathbb{E}(X|Y=y) = \int x f_{X|Y}(x|y)\,dx$

$$\left(\sum_x x f_{X|Y}(x|y) \right)$$

- We usually write simply $\mathbb{E}(X|Y)$, a rv which is a function of Y. $\mathrm{Var}(X|Y) = \mathbb{E}[(X - \mathbb{E}(X|Y))^2 | Y)$

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

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

-

Probability review (cont.)

Change of variable Let X, Y be rvs, continuous. Let $(X,Y) \mapsto (U,V)$ be a diff. bijection. Then U, V have joint pdf

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

where $J = \begin{pmatrix} \partial x/\partial u & \partial x/\partial v \\ \partial y/\partial u & \partial y/\partial v \end{pmatrix}$.

Important distributions

① Binomial: $X \sim \text{Bin}(n,p)$, $n \in \mathbb{N}$, $p \in [0,1]$

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

X = # successes in n indep Bernoulli trials

$$\mathbb{E}X = np, \quad \text{Var}\, X = np(1-p)$$

② Poisson: $X \sim \text{Poi}(\lambda)$, $\lambda > 0$

$$\mathbb{P}(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

$\text{Poi}(\lambda)$ is the limit of $\text{Bin}(n,p)$ as $n \to \infty$, keeping $np = \lambda$

In a Poisson process, the # of events happening in an interval of length t has a $\text{Poi}(\lambda t)$ distribution and disjoint intervals are indep

③ Negative Binomial: $X \sim \text{NB}(k,p)$, $k \in \mathbb{N}$, $p \in [0,1]$

$$\mathbb{P}(X = x) = \binom{x-1}{k-1} (1-p)^{x-k} p^k, \quad x = k, k+1, \ldots$$

X = # trials until k^{th} success in a sequence of iid Bernoulli trials

$$\mathbb{E}X = \frac{k}{p}, \quad \text{Var}\, X = \frac{k(1-p)}{p^2}$$

④ Multinomial: n indep trials with k possible outcomes, with probabilities $p_1, \ldots, p_k$

$(N_1, \ldots, N_k) \sim \text{Multin}(n, \underline{p})$ if N_i = # trials with outcome i

$$\mathbb{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}$$

for $n_1, \ldots, n_k \in \{0, \ldots, n\}$ with $\sum n_i = n$

⑤ Normal: $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in (0, \infty)$

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

We call $N(0,1)$ the standard normal. Denote its pdf ϕ, distribution function Φ.

⑥ Uniform: $X \sim \text{Unif}(a,b)$, $a, b \in \mathbb{R}$, $a < b$

$$f_X(x) = \frac{1}{b-a}, \quad x \in [a,b]$$

$$\mathbb{E}X = \frac{a+b}{2}, \quad \text{Var}\, X = \frac{(b-a)^2}{12}$$

⑦ Gamma: $X \sim \Gamma(\alpha, \lambda)$, $\alpha > 0$, $\lambda > 0$

$$f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geq 0$$

$$\mathbb{E}X = \frac{\alpha}{\lambda}, \quad \text{Var}\, X = \frac{\alpha}{\lambda^2}, \quad M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha$$

If $X_i \sim \Gamma(\alpha_i, \lambda)$ are indep, then

- For any $t \in \mathbb{R}$, $tX_1 \sim \Gamma(\alpha_1, \frac{\lambda}{t})$
- $\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n \alpha_i, \lambda)$

Show using MGF (trivial)

⑧ Exponential: $X \sim \text{Exp}(\lambda) = \Gamma(1, \lambda)$

Memoryless, distribution of waiting times in a Poisson process with rate λ

⑨ Chi squared: $X \sim \chi^2_k$, $k \in \mathbb{N}$, degrees of freedom

$$= \Gamma\left(\frac{k}{2}, \frac{1}{2}\right)$$

If $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$, then $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$

⑩ Beta: $X \sim \text{Beta}(\alpha, \beta)$, $\alpha, \beta > 0$

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad \text{for } x \in [0,1]$$

$$\mathbb{E}X = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}\, X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

L2.3

Estimation

A Suppose $X_1, \dots, X_n$ are iid, each has pdf $f_X(x \mid \theta)$ with θ unknown.

An <u>estimator</u> of θ is a function of the data $X_1, \dots, X_n$ (a <u>statistic</u>)

$$\hat{\theta} = T(X_1, \dots, X_n).$$

As $\hat{\theta}$ is a function of rvs, it is itself a r.v. Its distribution is called the <u>sampling distribution</u> of θ.

<u>Ex</u> $X_1, \dots, X_n \sim N(\mu, 1)$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{1}{n})$

<u>Def</u> The bias of $\hat{\theta}$ is $\text{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta$

↳ expectation wrt joint pdf $\prod_{i=1}^{n} f_X(x_i \mid \theta)$

B

In the above example, $\mathbb{E}_\mu(\hat{\mu}) = \mu$.

So $\text{bias}(\hat{\mu}) = 0$, we say $\hat{\mu}$ is <u>unbiased for μ</u>.

<u>Def</u> The <u>mean squared error</u> of $\hat{\theta}$ is $\text{mse}(\hat{\theta}) = \mathbb{E}_\theta((\hat{\theta}-\theta)^2)$ (a function of θ) wh

$$\begin{aligned}
\mathbb{E}_\theta((\hat{\theta}-\theta)^2) &= \mathbb{E}_\theta((\hat{\theta} - \mathbb{E}_\theta\hat{\theta} + \mathbb{E}_\theta\hat{\theta} - \theta)^2) \\
&= \mathbb{E}_\theta((\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2) + \mathbb{E}_\theta((\mathbb{E}_\theta\hat{\theta} - \theta)^2) \\
&\quad + 2\,\mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta\hat{\theta})\,\mathbb{E}_\theta(\mathbb{E}_\theta\hat{\theta} - \theta) \\
&= \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2
\end{aligned}$$

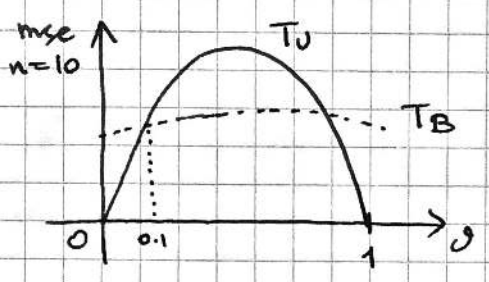
<u>NB</u> Sometimes a biased estimator can have lower MSE than an unbiased one. There is a tradeoff between bias and variance.

<u>Ex</u> Two different estimators of a binomial mean

Let $X \sim \text{Bin}(n, \theta)$, θ unknown

① $T_U = \frac{X}{n}$

② $T_B = \frac{1}{n+2}X + \frac{1}{n+2}$



$\mathbb{E}_\theta T_U = \frac{n\theta}{n} = \theta$, so unbiased

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

$\mathbb{E}_\theta T_B = \frac{n}{n+2}\theta + \frac{1}{n+2}$, so biased for $\theta \neq \frac{1}{2}$

$\text{Var}_\theta T_B = \frac{n\theta(1-\theta)}{(n+2)^2}$, lower variance ...

L2.4

T_B has a lower mse for much of the range of θ. If we know θ is likely to be close to $\frac{1}{2}$, it makes sense to use the biased estimat

In general, prior knowledge will change our choice of estimator.

Sufficiency

As usual, let $X = (X_1, \dots, X_n)$ be iid, from some distribution with parameter θ.

Def A statistic T is *sufficient for θ* if the conditional distribution of the data X given T does not depend on θ.

NB T and θ can be vectors, not necessarily of the same dimension. In a sense, a sufficient statistic summarises all information in X relevant for estimating θ.

Ex Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, so $f_X(x|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$

This depends on θ only through $= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$

$T(X) = \sum_{i=1}^n X_i$. Note $T(X) \sim \text{Bin}(n, \theta)$.

$$f_{X|T(X)=t}(x \mid T=t) = \frac{\mathbb{P}(X=x, T=t)}{\mathbb{P}(T=t)} = \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \left[\binom{n}{t}\right]^{-1} \qquad \text{zero if } T(x) \neq t$$

This does not depend on θ, hence $T(X) = \sum_{i=1}^n X_i$ is sufficient.

Thm (Factorisation criterion) T is sufficient for θ iff

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

for some functions g, h.

Ex (cont.) $f_X(x|\theta) = \underbrace{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}_{g(T(x),\theta)} \cdot \underbrace{1}_{h(x)}$

Hence $\sum_{i=1}^n X_i$ is sufficient (again).

Proof [Discrete case] Suppose $f_X(x|\theta) = g(T(x), \theta)\, h(x)$.

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

$$= \frac{h(x)}{\sum_{x', T(x')=t} h(x')},$$

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

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

$$\mathbb{P}_\theta(X=x) = \mathbb{P}_\theta(X=x,\ T(X)=T(x))$$
$$= \underbrace{\mathbb{P}_\theta(X=x \mid T(X)=T(x))}_{\text{indep of } \theta \text{ by sufficiency of } T, \text{ so write as } h(x)} \underbrace{\mathbb{P}_\theta(T(X)=T(x))}_{g(T(x),\theta)}$$

□

Ex $X_1, \dots, X_n \overset{iid}{\sim} \text{Unif}(0,\theta)$, $\theta > 0$

$$f_X(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i)$$
$$= \underbrace{\frac{1}{\theta^n} \mathbb{1}_{\{\max_i x_i \le \theta\}}}_{g(\max_i x_i, \theta)} \underbrace{\mathbb{1}_{\{\min_i x_i \ge 0\}}}_{h(x)}$$

Hence $T = \max_i X_i$ is sufficient for θ.

Remark Sufficient statistics are not unique. In fact, $T(X) = X$ is a sufficient statistic. Consider the equivalence relation $x \sim y$ if $T(x) = T(y)$. Every (sufficient) statistic partitions the sample space into equivalence classes of the form $\{x \in \mathcal{X}^n : T(x) = t\}$. What we want is for the partition to be as coarse as possible.

Def A sufficient statistic $T(X)$ is minimal sufficient if it is a function of any other sufficient statistic, i.e. for $S(X)$ sufficient,

$$S(x) = S(y) \Rightarrow T(x) = T(y).$$

Alternatively, the partition induced by T is coarser than that of S.

Thm Suppose $T(X)$ is a statistic such that $\frac{f_X(x|\theta)}{f_X(y|\theta)}$ does not depend on θ iff $T(x) = T(y)$.

Then $T(X)$ is minimal sufficient.

Proof (sketch) First we'll prove $T(X)$ is sufficient.

For each value of $T(X)$, let x_t be a representative of

$$\{x \in \mathcal{X}^n : T(x) = t\}$$

By hypothesis, $\frac{f_X(x'|\theta)}{f_X(x_{T(x')}|\theta)}$ does not depend on θ.

Then

$$f_X(x'|\theta) = \underbrace{f_X(x_{T(x')}|\theta)}_{\text{call } g(T(x'),\theta)} \underbrace{\frac{f_X(x'(\theta))}{f_X(x_{T(x')}|\theta)}}_{\text{call } h(x')}$$

(bar over $x'(\theta)$)

By the factorisation criterion, T is sufficient.

Now let S be another sufficient statistic. Then $\exists$ functions g_S, h_S such that $f_X(x|\theta) = g_S(S(x), \theta)\, h_S(x)$.

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

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

which does not depend on θ.

Hence by hypothesis $T(x) = T(y)$. So T is minimal. □

Ex $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\frac{f_X(x|\mu,\sigma^2)}{f_X(y|\mu,\sigma^2)} = \frac{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right)}{(2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2\right)}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2\right) + \frac{\mu}{\sigma^2}\left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i\right)\right\}$$

This is indep of μ, σ^2 iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ and $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$.

Therefore $T(X) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is minimal sufficient. (!)

NB Any bijection of a minimal sufficient statistic is minimal sufficient (only need injection). Therefore, setting

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \quad \text{get} \quad T(X) = \left(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2\right)$$

is also minimal sufficient in (μ, σ^2).

Sufficiency (cont.)

- * Minimal sufficient need not exist, but do under mild conditions
  * " are unique [If S, T both minimal sufficienct, then $S(x) = S(y) \iff T(x) = T(y)$]

Rao-Blackwell theorem

Let T be a sufficient statistic for a parameter θ. Let $\tilde{\theta}$ be an estimator with $\mathbb{E}\,\tilde{\theta}^2 < \infty$ for all θ.

Let $\hat{\theta} = \mathbb{E}(\tilde{\theta} \mid T)$. Then, for all θ,

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

- The inequality is strict unless $\tilde{\theta}$ is a function of T.

—o—

$$\mathbb{E}_\theta(\tilde{\theta}) = \int \tilde{\theta}(x) f_X(x \mid \theta)\, dx$$

Drop the subscripts LMAO

—o—

Proof Note $\mathbb{E}\,\hat{\theta} = \mathbb{E}(\mathbb{E}(\tilde{\theta} \mid T)) = \mathbb{E}\,\tilde{\theta}$, so $\hat{\theta}$ and $\tilde{\theta}$ have the same bias.

Also, $\operatorname{Var}\tilde{\theta} = \mathbb{E}(\underbrace{\operatorname{Var}(\tilde{\theta} \mid T)}_{>0 \text{ with +ve prob unless } \tilde{\theta} \text{ a f}^n \text{ of } T}) + \underbrace{\operatorname{Var}(\mathbb{E}(\tilde{\theta} \mid T))}_{\operatorname{Var}\hat{\theta}}$

- $\therefore$ mse $\tilde{\theta}$ = bias $(\tilde{\theta})^2 + \operatorname{var}\tilde{\theta}$
  $\geq$ bias $(\hat{\theta})^2 + \operatorname{var}\hat{\theta}$ = mse $\hat{\theta}$

with equality iff $\tilde{\theta}$ is a function of T. □

NB Where did we use that T is sufficient?

Sufficiency implies that the conditional distribution of X given T does not depend on θ, therefore $\hat{\theta} = \mathbb{E}(\tilde{\theta} \mid T)$ does not depend on θ, only on the data X. Hence it is a well defined estimator.

- Ex Suppose $X_1, \ldots, X_n \sim \text{Poi}(\lambda)$, let $\theta = e^{-\lambda} = \mathbb{P}(X_1 = 0)$

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

From factorisation criterion, we see that $T = \sum_{i=1}^n X_i$ is sufficient.

Define a simple estimator of θ, $\tilde{\theta} = \mathbb{1}_{\{X_1 = 0\}}$. This is unbiased.

Define $\hat{\tilde{\theta}} = \mathbb{E}(\tilde{\theta} \mid T) = \mathbb{P}(X_1 = 0 \mid \sum_{i=1}^n X_i = T)$

$$= \frac{\mathbb{P}(X_1 = 0, \sum_{i=1}^n X_i = T)}{\mathbb{P}(\sum_{i=1}^n X_i = T)} = \left(\frac{n-1}{n}\right)^t.$$

~ ugh? t vs T

So $\hat{\theta} = \left(\frac{n-1}{n}\right)^T$. By Rao-Blackwell, $\text{mse}\,\hat{\theta} \leq \text{mse}\,\tilde{\theta}$ $\forall\, \theta$.

Maximum Likelihood Estimation

Let $X_1, \ldots, X_n$ be iid rvs with joint pdf / pmf $f_X(x|\theta)$.

Def The **likelihood** is $\mathcal{L}(\theta) = f_X(x|\theta)$, regarded as a function of θ, given x. The **max likelihood estimator** (MLE) is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta).$$

It is equivalent, and also easier, to maximise the logarithm $\ell(\theta)$

$$\ell(\theta) = \log \mathcal{L}(\theta) = \log \prod_{i=1}^n f_{X_i}(x_i|\theta) = \sum \log f_{X_i}(x_i|\theta)$$

Ex $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p)

$$\ell(p) = \sum x_i \log p + (n - \textstyle\sum x_i) \log(1-p)$$

$$d\ell/dp = \textstyle\sum x_i / p - (n - \sum x_i)/(1-p)$$

This is zero at $\hat{p} = \sum x_i / n$.

Since $\sum X_i \sim \text{Bin}(n, p)$, $\mathbb{E}\hat{p} = p$ is unbiased.

So the MLE is cool.

Ex 2 $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$

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

$$\partial\ell/\partial\mu = -\frac{1}{\sigma^2}\sum_{i=1}^n (x_i - \mu)$$

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

ℓ maximised when $\partial\ell/\partial\mu = 0$, $\partial\ell/\partial\sigma^2 = 0$, this happens at

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \tfrac{1}{n} S_{XX})$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$.

Not hard to see $\hat{\mu}$ is unbiased, what about σ^2?

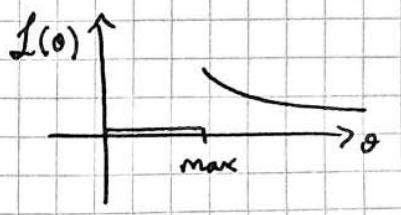
Later, we'll show that $\frac{S_{xx}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$

$\Rightarrow \frac{n}{\sigma^2} \mathbb{E}\hat{\sigma}^2 = n-1 \Rightarrow \mathbb{E}\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2$

So $\hat{\sigma}^2$ is biased, but asymptotically unbiased, $\mathbb{E}\hat{\sigma}^2 \xrightarrow[n\to\infty]{} \sigma^2$.

It will sometimes be convenient to use the unbiased estimator

$$\tilde{\sigma}^2 = \frac{S_{xx}}{n-1}.$$



Ex 3 $X_1, \dots, X_n \overset{iid}{\sim} \text{Unif}(0,\theta)$. Then

$$\mathcal{L}(\theta) = \frac{1}{\theta^n} \mathbb{1}\{\max_i x_i \le \theta\}.$$

This is zero for $\theta \le \max_i x_i$, and decreasing in θ for θ larger than $\max_i x_i$. So $\hat{\theta} = \max_i x_i$.

$F_{\hat{\theta}}(t) = \mathbb{P}(\hat{\theta} \le t) = \mathbb{P}(X_1 \le t, \dots, X_n \le t) = (t/\theta)^n$

Differentiating, the pdf of $\hat{\theta}$ is

$$f_{\hat{\theta}}(t) = \frac{n t^{n-1}}{\theta^n}$$

for $t \in [0,\theta]$.

$$\mathbb{E}\hat{\theta} = \int_0^\theta \frac{n t^n}{\theta^n} dt = \frac{n}{n+1}\theta$$

$\therefore \hat{\theta}$ is biased, but asymptotically unbiased.

Maximum Likelihood Estimator (cont.)

Properties of the MLE

① If T is sufficient for θ, then $\mathcal{L}(\theta) = g(T(x), \theta)\, h(x)$, thus the MLE is a function of T.

② Invariance If h is injective, the MLE of $\phi = h(\theta)$ is $\hat{\phi} = h(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ.

③ Asymptotic normality (II Principles of Statistics) Under regularity conditions $\sqrt{n}(\hat{\theta} - \theta)$ is approximately $N(0, \Sigma)$, where Σ is "the smallest variance attainable"

④ Sometimes the MLE has no closed form expression, but it can be found numerically.

Confidence Intervals

Def^n A $100\gamma\%$ confidence interval (CI) for θ is a random interval $(A(X), B(X))$ such that $P(A(X) < \theta < B(X)) = \gamma$ for all values of the parameter θ.

Remarks
- It is the endpoints which are random, not θ
- We can interpret this interval in terms of "repetitions of the experiment", i.e. let $X^{(1)}, X^{(2)}, \ldots$ be iid copies of X. Then $100\gamma\%$ of the intervals $(A(X^{(1)}), B(X^{(1)})), \ldots$ contain θ
- WRONG interpretation: there is a $100\gamma\%$ probability that the true parameter θ is in the interval $(A(x), B(x))$

Ex 1 Let X_1, X_2 be iid $\text{Unif}(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. What is a sensible confidence interval for θ?

$$\mathbb{P}(\min(X_1, X_2) < \theta < \max(X_1, X_2)) = \begin{matrix} \mathbb{P}(X_1 < \theta < X_2) \\ + \mathbb{P}(X_2 < \theta < X_1) \end{matrix} = \frac{1}{2}$$

Hence $(\min(X_1, X_2), \max(X_1, X_2))$ is a 50% confidence interval.

What happens if $|X_1 - X_2| \geq \frac{1}{2}$? e.g. $X_1 = 0.2$, $X_2 = 0.9$?

L5.2

We can be sure that $\theta \in (\min(X_1, X_2), \max(X_1, X_2))$.

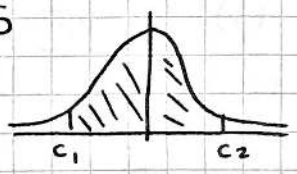
0.4 0.7
1 (11111) 1
0.2 0.9

Hence the confidence interval is not a sensible quantifier of uncertainty.

Ex 2 Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$. What is a 95% confidence interval for θ?

We know $\sqrt{n}(\bar{X} - \theta) \sim N(0,1)$, no matter what θ is.

Let c_1, c_2 be such that $\Phi(c_2) - \Phi(c_1) = 0.95$



Then $\mathbb{P}(c_1 < \sqrt{n}(\bar{X} - \theta) < c_2) = 0.95 \quad \forall \theta$.

$\therefore \mathbb{P}(\bar{X} - \frac{c_2}{\sqrt{n}} < \theta < \bar{X} - \frac{c_1}{\sqrt{n}}) = 0.95$

So, $(\bar{X} - \frac{c_2}{\sqrt{n}}, \bar{X} - \frac{c_1}{\sqrt{n}})$ is a 95% confidence interval.

We choose c_1, c_2 to make the interval as narrow as possible, this yields $c_2 = -c_1 \cong 1.96$.

Usual procedure ① Find a quantity $R(X, \theta)$ whose distribution does not depend on θ, e.g. $R(X, \theta) = \sqrt{n}(\bar{X} - \theta)$. This is what we call a pivot.

② Write out a statement $\mathbb{P}(c_1 \leq R(X, \theta) \leq c_2) = \gamma$, by choosing c_1, c_2 as appropriate quantiles of the distribution of $R(X, \theta)$.

③ Rearrange to put θ in the middle, i.e. $\mathbb{P}(\_ \leq \theta \leq \_)$

Remark The asymptotic normality of the MLE is useful to find confidence intervals, because we can choose $\sqrt{n}(\hat{\theta} - \theta)$ and apply the above reasoning (when n is large)

Ex 3 $X_1, \dots, X_n \overset{iid}{\sim} N(0, \sigma^2)$. What is a 100γ% CI for σ^2?

We know $\frac{X_i}{\sigma} \overset{iid}{\sim} N(0,1)$. Then let $R(X, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2 \sim \chi^2_n$

Pick c_1, c_2 such that $F_{\chi^2_n}(c_2) - F_{\chi^2_n}(c_1) = \gamma$.

Then $\mathbb{P}(c_1 \leq \frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2 \leq c_2) = \gamma$

$\therefore \mathbb{P}(\frac{1}{c_2} \sum_{i=1}^{n} X_i^2 \leq \sigma^2 \leq \frac{1}{c_1} \sum_{i=1}^{n} X_i^2) = \gamma$

so $(\frac{1}{c_2} \sum_{i=1}^{n} X_i^2, \frac{1}{c_1} \sum_{i=1}^{n} X_i^2)$ is a 100γ% CI for σ^2.

Ex 4 Let $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli (p). Find an approximate (for n large) 95% CI for p.

The MLE is $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. By CLT, $\hat{p}$ is approximately $N\left(p, \frac{1}{n}p(1-p)\right)$ for n large. Thus $\frac{\sqrt{n}(\hat{p}-p)}{\sqrt{p(1-p)}}$ is approximately $N(0,1)$.

So $\mathbb{P}\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$

The endpoints depend on p but we can plug in $\hat{p}$ for p so

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

is an approx 95% is an approx CI for p (oopsie).

NB $p(1-p) \leq \frac{1}{4}$ for $p \in [0,1]$, thus a conservative 95% CI for p is $\left(\hat{p} - \frac{1}{\sqrt{n}}, \hat{p} + \frac{1}{\sqrt{n}}\right)$.

Bayesian Inference <3

Thus far, we've worked in a Frequentist framework. We assume the parameter $\theta \in \Theta$ is fixed, observations X have pdf / pmf $f_X(\cdot \mid \theta)$. Inferences on θ through confidence intervals based on sampling X repeatedly. Statements hold no matter what θ is.

In a Bayesian framework, we must specify a prior distribution $\pi(\theta)$ on the parameter, which captures our "belief" on θ before seeing any data.

Inferences are based on the posterior distribution

$$\pi(\theta \mid x) = \frac{f_X(x \mid \theta)\, \pi(\theta)}{f_X(x)}, \qquad \text{(Bayes' rule)}$$

where $f_X(x) = \int_\Theta f_X(x \mid \phi)\, \pi(\phi)\, d\phi$. The posterior $\pi(\theta \mid x)$ represents our belief after having seen the data x.

NB · If θ, X are joint rvs with pdf $f_X(x \mid \theta)\pi(\theta)$, the posterior is simply the conditional pdf of θ given x.

· As $f_X(x)$ is just a normalisation constant, we usually write

$$\pi(\theta \mid x) \propto f_X(x \mid \theta)\, \pi(\theta).$$

Seeing this as a function of θ, we deduce what kind of distribution this is, and $f_X(x)$.

· By factorisation criterion, $\pi(\theta \mid x)$ only depends on x through sufficient statistics.

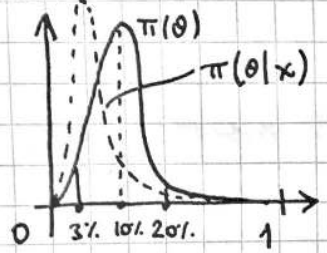
Ex 1 We want to estimate θ, mortality rate of new operation at Addenbrookes. In the UK, 10% of people who have the operation die, this varies from 3% to 20% in different hospitals.

Of the first 10 patients getting operation in Addenbrookes, none die.

$$X_1, X_2, \dots, X_{10} \overset{iid}{\sim} \text{Bernoulli}(\theta).$$

A reasonable prior for θ is a Beta(a, b) with $a = 3$, $b = 27$.

L6.2



$\pi(3\% \leq \theta \leq 20\%) = 0.9$

We observe $x_0 = x_1 = \cdots = x_9 = 0$.

Posterior $\pi(\theta|x) \propto \pi(\theta) f_X(x|\theta)$

$= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\, \theta^{a-1}(1-\theta)^{b-1}$

$= \theta^{a-1}(1-\theta)^{n+b-1}$,

which is a Beta$(a+\sum x_i, b+n-\sum x_i)$, or Beta$(3,37)$ in our example. This has mean 0.075 (contrast this with MLE $\hat{\theta} = 0$).

Inferential statement $\pi(\theta \geq 0.1 \mid x) \approx 0.24$. It is fair to read this as "the mortality rate has probability 0.24 of exceeding 10%", but this depends on the choice of prior.

When is it best to use frequentist vs Bayesian approach? Debate!

Be a frequentist when
- There's a lot of data
- Want to make "objective" statements, which hold uniformly over the true value of θ
- Ex High-energy physics. We want an objective test of wether the Higgs Boson exists

Be a Bayesian when
- Not a lot of data, complex data
- Have useful prior information
- We wish to make decisions under uncertainty
- Ex "Nowcasting and forecasting the potential domestic & international spread of 2019 nCov..."

Bayesian estimation

Let $L(\theta, a)$ be the loss incurred when we estimate a parameter to be a when its value is θ, e.g. $L(\theta,a) = (\theta-a)^2$ or $|\theta-a|$.

The posterior expected loss is $h(a) = \int_{\Theta} L(\theta,a)\,\pi(\theta|x)\,d\theta$.

Def^n A Bayes estimator under loss $L(\theta,a)$ minimises posterior expected loss, i.e. $\hat{\theta}_{Bayes} = \arg\min_{a \in \Theta} h(a)$.

Cases

① Quadratic loss, $h(a) = \int_\Theta (a-\theta)^2 \pi(\theta|x)\, d\theta$.

$h'(a) = 0$ if $a \underbrace{\int \pi(\theta|x)\, d\theta}_{=1} = \int \theta\, \pi(\theta|x)\, d\theta$.

Hence Bayes estimator one $\hat\theta$ is the posterior mean.

② Mean absolute error loss, $h(a) = \int_\Theta |a-\theta|\, \pi(\theta|x)\, d\theta$,

$h(a) = \int_{-\infty}^{a} (a-\theta)\pi(\theta|x)\, d\theta + \int_a^\infty (\theta - a)\pi(\theta|x)\, d\theta$,

$h'(a) = 0$ if $\int_{-\infty}^{a} \pi(\theta|x)\, d\theta = \int_a^\infty \pi(\theta|x)\, d\theta$.

Hence Bayes estimator $\hat\theta$ is the posterior median.

Ex 2 $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Let $\pi(\mu)$ be a $N(0, \tau^{-2})$.

$$\begin{aligned}\pi(\mu|x) &\propto f_x(x|\mu)\,\pi(\mu) \\ &\propto \exp\left\{-\tfrac{1}{2}\sum_{i=1}^n (x_i-\mu)^2\right\} \cdot \exp\left\{-\frac{\tau^2\mu^2}{2}\right\} \qquad \text{dropping factors indep of } \mu \\ &\propto \exp\left\{-\tfrac{1}{2}(n+\tau^2)\left\{\mu - \frac{\sum x_i}{n+\tau^2}\right\}^2\right\}\end{aligned}$$

Here $\pi(\mu|x)$ is $N\left(\frac{\sum x_i}{n+\tau^2}, \frac{1}{n+\tau^2}\right)$.

The Bayes estimator under quadratic or absolute error loss is $\frac{\sum x_i}{n+\tau^2}$.

(contrast with MLE $\hat\mu = \frac{1}{n}\sum x_i$).

Bayesian Inference (cont.)

Ex 3 $X_1, \dots, X_n \sim \text{Poi}(\lambda)$, let $\pi(\lambda)$ be an $\text{Exp}(1)$.

Posterior $\pi(\lambda | x) \propto f_X(x|\lambda) \cdot \pi(\lambda)$

$$\propto e^{-n\lambda} \lambda^{\sum x_i} e^{-\lambda}$$

$$= \lambda^{\sum x_i} e^{-(n+1)\lambda}$$

So $\pi(\lambda | x)$ is a Gamma $(\sum x_i + 1, n+1)$.

The Bayes estimator under quadratic loss is the mean $\hat{\lambda} = \dfrac{\sum x_i + 1}{n+1}$

Hypothesis Testing

Let $X = (X_1, \dots, X_n)$ be a vector of iid rvs taking values in $\mathcal{X}$, with pdf / pmf f. Let H_0, H_1 be two hypotheses about f. We shall call H_0 the null hypothesis and H_1 the alternative hypothesis.

On the basis of an observation $x \in \mathcal{X}^n$ we want to choose between them.

Ex 1 $X_1, \dots, X_n \overset{iid}{\sim}$ Bernoulli (θ)

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

Ex In the same setting, we may have

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

Def When a hypothesis completely determines f, we say it is simple (e.g. $H_0 : \theta = \frac{1}{2}$), otherwise we say it is composite (e.g. $H_1 : \theta \neq \frac{1}{2}$)

Ex 3 Say X_i's take values in $\mathbb{N}$

H_0: f is Poisson pmf with some unknown mean

H_1: f is not Poisson

This is called a goodness of fit test.

Ex 4 f is in some parametric family $\{ f(\cdot|\theta) : \theta \in \Theta \}$

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

$H_1 : \theta \in \Theta_1 \subseteq \Theta$ with $\Theta_0 \cap \Theta_1 = \emptyset$

Ex 5 $H_0 : f = f_0$, $H_1 : f = f_1$ where f_0, f_1 are pdfs, not necessarily in the same family

Def^n We say a **hypothesis test** for H_0 against H_1 with **critical region** $C \subseteq \mathcal{X}^n$ rejects H_0 if $x \in C$, and it fails to reject H_0 when $x \notin C$

NB You will often hear that if $x \notin C$, the test "accepts H_0", but this can be misleading

A hypothesis test can make 2 types of error

Type 1 We reject H_0 when H_0 is true

Type 2 We fail to reject H_0 when H_1 is true

From now on, assume H_0, H_1 are simple hypotheses.

Let $\alpha = \mathbb{P}(\text{Type 1 error}) = \mathbb{P}(X \in C \mid H_0)$,

$\beta = \mathbb{P}(\text{Type 2 error}) = \mathbb{P}(X \notin C \mid H_1)$.

Def^n We say the **size** of a test is α, the **power** to detect H_1 is $1-\beta$

We'd like to minimise α, β, but typically they cannot be made arbitrarily small. Usually a tradeoff (decreasing α increases β).

Typical procedure for choosing test (or C) Fix Type 1 error probability α that we are happy to accept, e.g. $\alpha = 0.05$. Find test of size α which minimises β.

Given 2 simple hypotheses $H_0: f = f_0$, $H_1: f = f_1$, the likelihood ratio given data x is

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

A likelihood ratio (LR) test is one with critical region

$$\{ x \in \mathcal{X}^n : \Lambda_x(H_0, H_1) > k \} \quad \text{for some } k \geq 0.$$

Thm (Neyman-Pearson lemma) Suppose $H_0: f = f_0$, $H_1: f = f_1$, where f_0, f_1 are continuous densities which are non-zero on the same set.

Then among all tests of size $\leq \alpha$, the test with the highest power is a LR test, i.e.

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

with k chosen such that $\alpha = \mathbb{P}(X \in C \mid H_0) = \int_C f_0(x)\,dx$.

NB We require continuous f_0, f_1 such that LR of size α exists for all α with $0 \leq \alpha \leq 1$.

Proof Let $\beta = \mathbb{P}(X \notin C \mid H_1) = \int_{\bar{C}} f_1(x)\,dx$, where $\bar{C}$ is the complement of C in $\mathcal{X}^n$.

Take another test with critical region D, define $\alpha^* = \int_D f_0(x)\,dx$, $\beta^* = \int_{\bar{D}} f_1(x)\,dx$. Suppose $\alpha^* \leq \alpha$. WTS $\beta^* \geq \beta$.

$$\begin{aligned}
\beta - \beta^* &= \int_{\bar{C}} f_1(x)\,dx - \int_{\bar{D}} f_1(x)\,dx \\
&= \Big(\int_{\bar{C}\cap D} + \cancel{\int_{\bar{C}\cap\bar{D}}} - \int_{\bar{D}\cap C} - \cancel{\int_{\bar{D}\cap\bar{D}}}\Big) f_1(x)\,dx \\
&= \Big(\underbrace{\int_{\bar{C}\cap D}}_{\subseteq \bar{C},\ \therefore \Lambda \leq k} - \underbrace{\int_{\bar{D}\cap C}}_{\subseteq C,\ \therefore \Lambda > k}\Big) f_0(x)\,\Lambda_x(H_0, H_1)\,dx \\
&\leq k \int_{\bar{C}\cap D} f_0(x)\,dx - k\int_{\bar{D}\cap C} f_0(x)\,dx \\
&= k\Big(\int_{\bar{C}\cap D} + \int_{C\cap D} - \int_{C\cap D} - \int_{\bar{D}\cap C}\Big) f_0(x)\,dx \\
&= k\Big(\int_D - \int_C\Big) f_0(x)\,dx \\
&= k(\alpha^* - \alpha) \\
&\leq 0.
\end{aligned}$$

□

Ex Let $X_0, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 known. Want to find most powerful test of size α for $H_0 : \mu = \mu_0$, against $H_1 : \mu = \mu_1$.

$$\begin{aligned}
\Lambda_x(H_0, H_1) &= \exp\Big(-\frac{1}{2\sigma^2}\sum (x_i - \mu_1)^2 + \frac{1}{2\sigma^2}\sum (x_i - \mu_0)^2\Big) \\
&= \exp\Big(\frac{(\mu_1 - \mu_0)}{\sigma^2} n\bar{x} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma^2}\Big)
\end{aligned}$$

For $\mu_0 < \mu_1$, Λ_x is increasing in $\bar{x}$, hence

$\Lambda_x > k \iff \bar{x} > c$ for some c.

The critical region of LR test is $\{x : \bar{x} > c\}$ for some c.

Let $Z = \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma}$. Then $Z \sim N(0,1)$ under H_0.

And $\bar{x} > c \iff z > c^*$.

Hence the LR test has optimal region $C = \{x : z > c^*\}$.

Lastly, we find c^* s.t.

$$\mathbb{P}(z > c^* \mid H_0) = \alpha \quad \text{i.e.} \quad c^* = z_\alpha = \Phi^{-1}(1-\alpha).$$

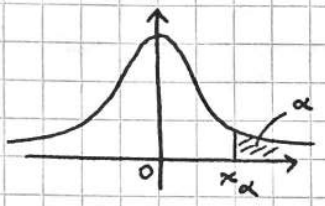
Example $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, σ^2 known

Test $H_0 : \mu = \mu_0$, $H_1 : \mu = \mu_1$, $\mu_0 < \mu_1$

$\Lambda_x(H_0, H_1)$ increasing function of $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$.

Hence $\Lambda_x(H_0, H_1) > k \iff Z > c$ for some c depending on k.

Under H_0, $Z \sim N(0,1)$. So a LR test of size α has critical region $\{ x \in \mathcal{X}^n : z > z_\alpha \}$ where z_α is upper α point of $N(0,1)$.



We say Z is a test statistic, the integral

$$\int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = p$$

is called the p-value, i.e. the probability that the test statistic is more extreme than the observed value. (under H_0?)

Let $\sigma^2 = 1$, $n = 10$, $\mu_0 = 0$, $\mu_1 = 1$. Then $z = 6.32$. ($\bar{X} = 2$)

Say we wish to make a test of size 0.05. Then $z_\alpha = 1.69$.

Hence the test rejects H_0, since $6.32 > 1.69$. In this case, the p-value is $p = 1.3 \times 10^{-10}$.

This tells us that we could have rejected a test of much smaller size, so this is strong evidence against H_0.

Prop The p-value has a Unif(0,1) distribution under H_0.

Proof Note $p = \int_z^\infty f_Z(u) \, du$.

$$\begin{aligned} \mathbb{P}(P \le t) &= \mathbb{E}\left[\mathbb{1}_{\{P \le t\}} \right] = \mathbb{E}\left[\mathbb{1}_{\{ z \ge F_Z^{-1}(1-t) \}} \right] \\ &= 1 - F_Z\left(F_Z^{-1}(1-t) \right) = t. \end{aligned}$$

□

Composite hypotheses

Suppose $X_1, \dots, X_n$ iid from parametric $f(x \mid \theta)$, $\theta \in \Theta$.

For a composite hypothesis such as $H : \theta > 0$, the probabilities of Type I and Type II error do not have a single value.

Def The power function of a test with critical region C is

$$W(\theta) = \mathbb{P}(X \in C \mid \theta) = \mathbb{P}(\text{reject } H_0 \mid \theta).$$

Consider $H_0: \theta \in \Theta_0$, $H_1: \theta \in \Theta_1$, $\Theta_0, \Theta_1 \in \mathcal{P}(\Theta)$.

The <u>size</u> of test is $\alpha = \sup_{\theta \in \Theta_0} W(\theta)$.

We say a test of H_0 against H_1 is <u>uniformly most powerful</u> (UMP) <u>of size α</u> if ① $\sup_{\theta \in \Theta_0} W(\theta) = \alpha$

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

<u>NB</u> A UMP test need not exist, but in many cases LR test is uniformly most powerful.

<u>Ex</u> $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, σ^2 known

<u>Test</u> $H_0: \mu \leq \mu_0$ vs $H_1: \mu \geq \mu_0$

Let C be the critical region $\{x \in \mathcal{X}^n : \sqrt{n}(\bar{X} - \mu_0)/\sigma > z_\alpha\}$ (i.e. the LR test for $H_0: \mu = \mu_0$, $H_1 = \mu = \mu_1$, $\mu_0 < \mu_1$).

The power function

$$W(\mu) = \mathbb{P}_\mu(\text{reject } H_0) = \mathbb{P}_\mu\left(\sqrt{n}(\bar{X} - \mu_0)/\sigma > z_\alpha\right)$$

$$= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} > z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right)$$

$$= 1 - \Phi\left(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right).$$

So $W(\mu)$ is increasing in μ. So

$\sup_{\mu \leq \mu_0} W(\mu) = W(\mu_0) = \alpha$ by previous example.

So ① is satisfied. Now, take C^* to be any other critical with size $\leq \alpha$. Then by Neyman-Pearson lemma, $W(\mu_1) \geq W^*(\mu_1)$ for any $\mu_1 > \mu_0$, i.e. any μ_1 in H_1. Hence ② is also satisfied and the LR test is UMP.

<u>NB</u> In many examples of the form $H_0: \theta \leq \theta_0$, $H_1: \theta > \theta_0$, we can apply the same reasoning.

(1) Determine LR test for $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$, $\theta_0 < \theta_1$

(2) Observe that critical region C does not depend on θ_1

(3) Hence by NP lemma, the test is UMP

Suppose $\Theta_0 \subseteq \Theta_1$. (nested hypotheses). The generalised likelihood ratio is

$$\Lambda_x(H_0, H_1) = \frac{\sup_{\theta \in \Theta_1} f(x|\theta)}{\sup_{\theta \in \Theta_0} f(x|\theta)} \geq 1.$$

Ex $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 known

Test $H_0: \theta = \theta_0$, $H_1: \theta \in \mathbb{R}$ ← meant μ's

The generalised likelihood ratio is

$$\Lambda_x(H_0, H_1) = \frac{(2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i - \bar{x})^2\right)}{(2\pi\sigma)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum(x_i - \mu_0)^2\right)}.$$

Note $2 \log \Lambda_x(H_0, H_1) = \frac{n}{\sigma^2}(\bar{X} - \mu_0)^2$.

Under H_0, this is a χ^2_1 v.r. The GLR test rejects when $\Lambda_x(H_0, H_1)$ is large.

Hence for a size α test, we reject H_0 when $\frac{n}{\sigma^2}(\bar{X} - \mu_0)^2 > \chi^2_1(\alpha)$, the upper α point of a χ^2_1 dist.

Note $Z^2 = \frac{(\bar{X} - \mu_0)^2 n}{\sigma^2}$. Thus this test rejects when Z is very large or very small (called a "2-tailed test").

Chi-squared tests

Recall example from last time: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, σ^2 known

- $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$

The generalised likelihood ratio Λ had

$$2 \log \Lambda = \frac{(\bar{x} - \mu_0)^2 n}{\sigma^2} \sim \chi^2_1$$

Thm Suppose $X_1, \dots, X_n$ are iid from a parametric model $f(\cdot|\theta)$.

Let $H_0: \theta \in \Theta_0$, $H_1: \theta \in \Theta_1$, with $\Theta_0 \subseteq \Theta_1$.

Let $\dim \Theta_0$ denote the dimension of Θ_0. Usually $\Theta \subseteq \mathbb{R}^k$ and Θ_0 is a linear subspace of $\mathbb{R}^k$, $\{\theta \in \mathbb{R}^k : A\theta = b\}$ for some matrix $A \in \mathbb{R}^{p \times k}$, $b \in \mathbb{R}^p$; then $\dim \Theta_0 = k - p$.

- Similarly define $\dim \Theta_1$.

Thm (Wilks') Suppose $\Theta_0 \subseteq \Theta_1$, $\dim \Theta_1 - \dim \Theta_0 = p$. Then under regularity conditions, if H_0 is true, i.e. $X_1, \dots, X_n$ are iid with pdf $f(\cdot|\theta)$ for $\theta \in \Theta_0$, then as $n \to \infty$, the distribution of $2 \log \Lambda$ converges to a χ^2_p distribution.

[More precisely, if F_n is the distribution of $2 \log \Lambda$, and F is that of a χ^2_p distribution, then $F_n \to F$ pointwise]

- If H_1 is true but H_0 is not, then $2 \log \Lambda$ will tend to be bigger. Hence a test which rejects H_0 when $2 \log \Lambda > \chi^{-1}_p(\alpha)$ has size approximately α, by the theorem, and some power to detect H_1.

Goodness-of-fit test

Birth month of everyone admitted to Cambridge in 2012

Month	Sep	Oct	Nov	Dec	⋯	Aug
n_i	470	515	470	457	⋯	399

Is this compatible with the proportions $\tilde{p}_1, \dots, \tilde{p}_{12}$ of UK births each month?

- NB $\tilde{p}$ can be determined from Census data, is not uniform

"Christmas effect" more births around September

L 9.2

Model Let N_i be the # births in month i $\quad k=12$

$(N_1, \dots, N_{12}) \sim$ Multinomial $(n, (p_1, \dots, p_{12}))$, n fixed

$H_0: p_i = \tilde{p}_i$, $H_1:$ p_i's unrestricted , $p \in [0,1]^{12}$, $\sum_{i=1}^{12} p_i = 1$

Here $\dim(\Theta_0) = 0$, $\dim(\Theta_1) = 11$

$$\Lambda = \frac{\sup_{p \in \Theta_1} \binom{n}{n_1 \cdots n_k} p_1^{n_1} \cdots p_k^{n_k}}{\sup_{p \in \Theta_0} \binom{n}{n_1 \cdots n_k} p_1^{n_1} \cdots p_k^{n_k}} = \frac{\hat{p}_1^{n_1} \cdots \hat{p}_k^{n_k}}{\tilde{p}_1^{n_1} \cdots \tilde{p}_k^{n_k}} \leftarrow \hat{p} \text{ is MLE}$$

The log likelihood is

$$\ell(p) = \text{const.} + \sum_{i=1}^{k} n_i \log p_i$$

Using the Lagrangian $\mathcal{L}(p, \lambda) = \sum_{i=1}^{k} n_i \log p_i - \lambda\left(\sum_{i=1}^{k} p_i - 1\right)$ we can minimise $\ell(p)$ subject to $\sum p_i = 1$.

We obtain $\hat{p}_i = \frac{n_i}{n}$ for $i = 1, \dots, k$.

Hence

$$2 \log \Lambda = 2 \log \left[\left(\frac{n_1}{n\tilde{p}_1}\right)^{n_1} \cdots \left(\frac{n_k}{n\tilde{p}_k}\right)^{n_k} \right]$$

$$= 2 \sum_{i=1}^{k} n_i \log \left(\frac{n_i}{n\tilde{p}_i}\right).$$

A chi-squared test of size approx α rejects the null H_0 when

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

In this example we have $2 \log \Lambda = 44.9$. Say $\alpha = 0.01$, we have p-value $\mathbb{P}(\chi^2_{k-1} > 44.9) \approx 3 \times 10^{-9}$. As this is < 0.01 we can reject H_0 at level 1%.

Modify H_0 to be $H_0: p_i = p_i(\theta)$ for some scalar parameter θ,

e.g. $p_1(\theta) = \frac{1}{12} + \theta$, $p_2(\theta) = \cdots = p_{12}(\theta) = \frac{1}{12} - \frac{\theta}{11}$, $\theta \in \mathbb{R}$ represents the Christmas effect.

We have still H_1: p_i's unrestricted.

In this case $2 \log \Lambda = \sum n_i \log \left(\frac{n_i}{n p_i(\hat{\theta})}\right)$ where $\hat{\theta}$ is the MLE of θ under H_0.

Pearson's χ^2 statistic Let $o_i = n_i$, the "observed # samples of type i".

Let e_i be the "expected # samples of type i" under H_0.

In our first example, $e_i = n \times \tilde{p}_i$.

In our second example, $e_i = n \times p_i(\hat{\theta})$.

We can write

$$\begin{aligned} 2\log\Lambda &= 2\sum o_i \log\left(\frac{o_i}{e_i}\right), \quad \text{set } \delta_i = o_i - e_i \\ &= 2\sum (e_i+\delta_i)\log\left(1+\frac{\delta_i}{e_i}\right) \\ &\approx 2\sum\left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{1}{2}\frac{\delta_i^2}{e_i}\right) \qquad [\text{Taylor}] \\ &= \sum \frac{\delta_i^2}{e_i} \qquad [\textstyle\sum \delta_i = \sum o_i - \sum e_i = n-n] \\ &= \sum \frac{(e_i-o_i)^2}{e_i} \end{aligned}$$

We call this the Pearson χ^2 statistic. This is referred to a χ^2_p distribution, where $p = \dim(\Theta_1) - \dim(\Theta_0) = k-1-\dim(\Theta_0)$

Ex Mendel crossed 556 SY peas with WG peas.

From progeny, N_1 were SY

N_2 wee SG

N_3 were WY

N_4 wee WG

He observed $(n_1, n_2, n_3, n_4) = (315, 108, 103, 31)$

Mendel's hypothesis $(p_1, \dots, p_4) = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right)$

Here, $2\log\Lambda = 0.618$, $\sum \frac{(o_i-e_i)^2}{e_i} = 0.604$.

The p-value is $\mathbb{P}(\chi^2_3 > 0.604) \approx 0.96$.

So no evidence against Mendel's hypothesis.

It wouldn't be rejected by a test of size 10%.

Actually, the high p-value (small test statistic) would suggest that the observed numbers are too close to the theoretical proportions.

Chi-squared tests (cont.)

Contingency tables

Ex Survey 500 people

	Prefer Kate	Prefer Meghan	No preference
Leave	56	52	42
Remain	50	83	67
No vote	18	51	81

N_{ij} : # of samples in row i column j

Observations are $(N_{ij}, i=1,\dots,r, j=1,\dots,c)$

Denote $N_{i+} = \sum_{j=1}^{c} N_{ij}$, $N_{+j} = \sum_{i=1}^{r} N_{ij}$

Model 1 $(N_{ij}, i=1,\dots,r, j=1,\dots,c) \sim \text{Multinomial}(n, (p_{ij}, \substack{i=1,\dots,r \\ j=1,\dots,c}))$

n fixed, in this example $n=500$

p_{ij} probability that a sample falls in cell (i,j)

Independence test H_0: Preferred royal indep of Brexit vote

i.e. $p_{ij} = p_{i+} p_{+j}$ where $p_{i+} = \sum_{j=1}^{c} p_{ij}$, $p_{+j} = \sum_{i=1}^{r} p_{ij}$

$(p_{1+},\dots,p_{r+}), (p_{+1},\dots,p_{+c})$ are unconstrained probability vectors

H_1: $(p_{ij} : i=1,\dots,r, j=1,\dots,c)$ is unconstrained probability vector

MLE of p under H_1 $\hat{p}_{ij} = \frac{N_{ij}}{n}$ (same as multinomial model! in previous lecture)

MLE of p under H_0

$$\mathcal{L}(p) = \frac{n!}{\prod_{i,j} N_{ij}!} \prod_{i,j} (p_{i+} p_{+j})^{N_{ij}}$$

Take logarithm, maximise subject to $\sum_{i=1}^{r} p_{i+} = \sum_{j=1}^{c} p_{+j} = 1$

Use Lagrangian method to obtain $\tilde{p}_{ij} = \tilde{p}_{i+} \tilde{p}_{+j} = \frac{N_{i+} N_{+j}}{n^2}$

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

In this case, $o_{ij} = N_{ij}$, "observed numbers"

$e_{ij} = \tilde{p}_{ij} n$ "expected number under H_0"

$= N_{i+} N_{+j} / n$

With these numbers we can compute $2 \log \Lambda$ or the Pearson χ^2 stat.

By Wilkes' theorem, this has distribution approx $\chi^2_{\dim \Theta_1 - \dim \Theta_0}$

Here $\dim \Theta_1 = rc - 1$ [1 equality condition $\sum_{ij} p_{ij} = 1$]

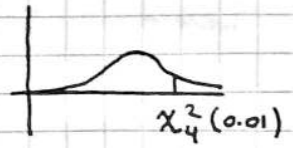
$\dim \Theta_0 = (r-1) + (c-1)$ [$(p_{1+}, \dots, p_{r+})$ has 1 equality constraint; $(p_{+1}, \dots, p_{+c})$ has 1 equality constraint]

Thus, the degrees of freedom are $rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$

In this example $\sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 36.20$

This is referred to a $\chi^2_{(3-1)(3-1)} = \chi^2_4$.

Observe that upper 1% point $\chi^2_4(0.01) = 13.28$



As $36.20 > 13.28$, a test of size 1% rejects H_0.

Strong evidence against independence.

Ex 50 patients are assigned to each of 3 treatments. Outcomes are recorded in a contingency table.

	Improvement	No Improvement	Worsening	
Placebo	18	17	15	50
½ dose	20	10	20	50
Full dose	25	13	12	50

↑ # people getting each treatment is fixed

Model 2 n_{i+} fixed for $i = 1, \dots, r$

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

independently for each row $i = 1, \dots, r$

In the example, $n_{1+} = n_{2+} = n_{3+} = 50$

Homogeneity test H_0: Treatment has no effect on the outcome

i.e. the probabilities of each outcome are the same for different treatments

$p_{1j} = p_{2j} = \dots = p_{rj} =: q_j$ for $j = 1, \dots, c$

for some vector of probabilities $(q_1, \dots, q_c)$

H_1: $(p_{i1}, \dots, p_{ic})$ is unconstrained probability vector for each row

MLE under H_1 $\mathcal{L}(p) = \prod_{i=1}^{r} \binom{n_{i+}}{N_{i1} \cdots N_{ic}} p_{i1}^{N_{i1}} \cdots p_{ic}^{N_{ic}}$

Take a log, maximise subject to $\sum_j p_{ij} = 1$ for each i

$\Rightarrow \hat{p}_{ij} = \frac{N_{ij}}{N_{i+}}$ [NB This amounts to maximising each factor separately]

MLE of p under H_0 $\mathcal{L}(q) = \prod_{i=1}^{r} \binom{n_{i+}}{N_{i1} \cdots N_{ic}} q_1^{N_{i1}} \cdots q_c^{N_{ic}}$

Maximise subject to $\sum_j q_j = 1$

$\Rightarrow \hat{q}_j = \frac{N_{+j}}{n}$ where $n = \sum_i n_{i+}$

$o_{ij} = N_{ij}$, $e_{ij} = n_{i+}\tilde{q}_j = n_{i+}N_{+j}/n$

NB e_{ij}, o_{ij} are identical to in the independence test

The only difference is that previously n_{i+} was random, whereas now it is fixed.

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

Pearson's $\chi^2 \sum_{i,j}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$

are identical in both tests.

The reference distribution is $\chi^2_{\dim\Theta_1 - \dim\Theta_0}$

Here, $\dim\Theta_1 = r(c-1)$ [r probability vectors of length c]

$\dim\Theta_0 = c-1$

So degrees of freedom are $(r-1)(c-1)$.

Conclude Homogeneity test has the same conclusions as the independence test despite the differences in their models.

Relationship between confidence sets and hypothesis tests

Suppose $(X_1, \ldots, X_n) = X$ have joint pdf / pmf $f_X(x|\theta)$ for some $\theta \in \Theta$.

Thm (i) Suppose for any $\theta_0 \in \Theta$ there is a size α test for $H_0: \theta = \theta_0$ with critical region $C(\theta_0)$. Then the set $I(X) = \{\theta : X \notin C(\theta)\}$ is a $100(1-\alpha)\%$ confidence set for θ,

i.e. we build a confidence set from all θ_0 such that $H_0: \theta = \theta_0$ is not rejected.

(ii) Suppose that $I(X)$ is a $100(1-\alpha)\%$ confidence set for θ.

Then $C(\theta_0) = \{X : \theta_0 \notin I(X)\}$ is the critical region of a size α test for $H_0: \theta = \theta_0$,

i.e. the test rejects H_0 when θ_0 is not in $I(X)$

Proof Note that in each case $\theta_0 \in I(X) \Leftrightarrow X \notin C(\theta_0)$.

Observe

$$\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = \mathbb{P}(X \in C(\theta_0) \mid H_0 \text{ is true})$$
$$= 1 - \mathbb{P}(\theta_0 \in I(X) \mid \theta = \theta_0).$$

For part (i), we assume LHS $= \alpha$, thus

$$\mathbb{P}(\theta_0 \in I(X) \mid \theta = \theta_0) = 1 - \alpha$$

which implies $I(X)$ is a $100(1-\alpha)\%$ confidence set.

For part (ii), we assume $\mathbb{P}(\theta_0 \in I(X) \mid \theta = \theta_0) = 1 - \alpha$ which implies $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true}) = 1 - (1-\alpha) = \alpha$ and so $C(\theta_0)$ is the critical region of a test of size α for $H_0: \theta = \theta_0$. □

Ex $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$. Suppose we want a 95% confidence set for μ. Recall that a size 0.05 test for $H_0: \mu = \mu_0$ rejects when the test statistic $Z = \sqrt{n}(\bar{X} - \mu_0)$ has $|Z| > 1.96$ [1.96 is the upper 2.5% point of a $N(0,1)$ distribution]

By part (i) of the theorem, a 95% conf. set for μ is given by

$$I(X) = \{\mu_0 : X \notin C(\mu_0)\} = \{\mu_0 : |Z| \leq 1.96\}$$
$$= \{\mu_0 : |(\bar{X} - \mu_0)\sqrt{n}| \leq 1.96\} = \left(\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}}\right).$$

This is the CI we found previously.

Orthogonal projections

Let V be a subspace of $\mathbb{R}^n$. Let $V^\perp = \{ w \in \mathbb{R}^n : w^T v = 0 \ \forall v \in V \}$, be its orthogonal complement. Recall any vector $x \in \mathbb{R}^n$ can be written uniquely as $x = v + w$ with $v \in V$, $w \in V^\perp$.

Def An orthogonal projection matrix $\Pi \in \mathbb{R}^{n \times n}$ onto V acts on any $x = v + w$, $v \in V$, $w \in V^\perp$ by $\Pi(x) = v$.

Properties ① The column space of Π is V

② $I - \Pi$ is an orthogonal projection onto $V^\perp$. Indeed, if $x = v + w$, $v \in V$, $w \in V^\perp$, then

$$(I - \Pi)x = v + w - v = w.$$

③ $\Pi^2 = \Pi = \Pi^T$

($\Pi^2 = \Pi$: idempotent; $\Pi = \Pi^T$: self-adjoint)

$\Pi^2 x = \Pi v = v$, so idempotency is clear from def^n

For self-adjoint-ness, let $x_1, x_2 \in \mathbb{R}^n$. Then

$$(\Pi x_1)^T ((I - \Pi) x_2) = 0$$

$$\therefore x_1^T (\Pi^T - \Pi^T \Pi) x_2 = 0$$

This holds for all $x_1, x_2 \in \mathbb{R}^n$, so $\Pi^T = \Pi^T \Pi$.

Since $\Pi^T \Pi$ is symmetric, so are Π^T and Π.

Conversely, any Π which satisfies $\Pi^2 = \Pi = \Pi^T$ is an orthogonal projection onto $\mathrm{col}(\Pi)$. Proof left as an exercise.

④ An orthonormal basis for V, and $V^\perp$ is made up of eigenvectors of Π with eigenvalues 1 and 0 respectively. So we can write $\Pi = UDU^T$ for U orthogonal, D diagonal.

Moreover, D has p 1s and $n-p$ 0s on its diagonal, where

$$p = \dim V.$$

⑤ $\mathrm{Tr}(\Pi) = \mathrm{Tr}(UDU^T) = \mathrm{Tr}(D) = p = \dim V$ ($= \mathrm{Rank}(\Pi)$)

Multivariate Normal Distributions

- If $X = (X_1, \dots, X_n)$ is a random vector, we define

$$\mathbb{E}X = (\mathbb{E}X_1, \dots, \mathbb{E}X_n).$$

Similarly for a random matrix X, $(\mathbb{E}X)_{ij} = \mathbb{E}(X_{ij})$.

For 2 col. vectors V, W define $\operatorname{cov}(V, W) = \mathbb{E}\left[(V - \mathbb{E}V)(W - \mathbb{E}W)^T\right]$

We'll also let $\operatorname{cov}(X) = \operatorname{cov}(X, X)$.

Linearity of expectation implies

① For any fixed $A \in \mathbb{R}^{m \times n}$, $\mathbb{E}(AX) = A\mathbb{E}(X)$.

② For any matrices A, B,

- $$\begin{aligned}\operatorname{cov}(AX, BX) &= \mathbb{E}\left[(AX - \mathbb{E}(AX))(BX - \mathbb{E}(BX))^T\right] \\ &= \mathbb{E}\left[A(X - \mathbb{E}X)(X - \mathbb{E}X)^T B^T\right] \\ &= A \operatorname{cov}(X) B^T.\end{aligned}$$

Def We say a random vector X is multivariate normal if $\forall t \in \mathbb{R}^n$, $t^T X$ has a normal distribution.

Lemma The MVN distribution is completely determined by

$$\mu = \mathbb{E}X \quad \text{and} \quad \Sigma = \operatorname{cov}(X).$$

Hence we write $X \sim N(\mu, \Sigma)$.

- Proof The multivariate moment generating function is

$$M_X(t) = \mathbb{E}\, e^{t^T X} \quad \text{for } t \in \mathbb{R}^n.$$

When this exists for all $t \in \mathbb{R}^n$, it fully determines the distribution of X. Now if $X \sim N(\mu, \Sigma)$, then $t^T X \sim N(t^T\mu, t^T\Sigma t)$, by the properties above.

Hence MGF is $M_X(t) = \mathbb{E}\, e^{t^T X} = M_{t^T X}(1) = \exp(t^T\mu + \frac{1}{2} t^T \Sigma t)$, which only depends on μ and Σ. □

-

Lemma Let $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}\right)$ where $X_i, \mu_i \in \mathbb{R}^{n_i}$, and $\Sigma_{ij} \in \mathbb{R}^{n_i \times n_j}$. Then

(i) $X_i \sim N(\mu_i, \Sigma_{ii})$,

(ii) $X_1 \perp X_2 \iff \Sigma_{12} = 0$.

Proof (i) WLOG $i = 1$. Note X_1 is multivariate normal, because $\forall t \in \mathbb{R}^{n_1}$, $t^T X_1 = (t \;\; 0)^T \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is normal by assumption.

And $\mathbb{E} X_1 = \mu_1$, $\operatorname{Cov} X_1 = \Sigma_{11}$.

(ii) ($\Rightarrow$) If X_1, X_2 are indep, then

$$\Sigma_{12} = \operatorname{cov}(X_1, X_2) = \mathbb{E}\left((X_1 - \mu_1)(X_2 - \mu_2)^T\right) = 0.$$

($\Leftarrow$) As the MVN dist^n is characterised by its mean and covariance, if $\Sigma_{12} = 0$, then $X_1 \perp X_2$. □

NB (ii) does not necessarily hold if X_1, X_2 are not jointly normal.

Lemma Let π be an orthogonal projection of rank p. Let $X \sim N(0, I_n)$. Then $\|\pi X\|_2^2 \sim \chi_p^2$.

Proof Let $\pi = UDU^T$ be an eigendecomposition of π.

$$\|\pi X\|_2^2 = \|UDU^T X\|_2^2 \underset{U \text{ orthog}}{=} \|DU^T X\|_2^2 \overset{(d)}{=} \|DX\|_2^2$$

The last identity follows from the fact that $U^T X \sim N(0, I_n)$.

Indeed, $U^T X$ is MVN (check $t^T U^T X = (Ut)^T X$ is normal), and $\mathbb{E} U^T X = U^T \mathbb{E} X = 0$, $\operatorname{Cov}(U^T X) = U^T \operatorname{Cov} X \, U = U^T U = I_n$.

We note $\|DX\|_2^2 = \sum_{i=1}^{p} X_i^2 \sim \chi_p^2$.

Indeed, the X_i are iid $N(0,1)$ variables. □

Normal random sample

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. We don't assume σ^2 is known.

We want to estimate μ.

Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$

Claim (i) $\bar{X} \sim N(\mu, \sigma^2/n)$ [this was shown previously]

(ii) $S_{XX}/\sigma^2 \sim \chi_{n-1}^2$ (iii) $\bar{X} \perp S_{XX}$

Proof Let Π be the orthogonal projection onto span$\{\underline{1}\}$, i.e. $\Pi_{ij} = \frac{1}{n}$.

Observe if $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$, then $\Pi X = \begin{pmatrix} \bar{X} \\ \vdots \\ \bar{X} \end{pmatrix} = \bar{X}\underline{1}$.

Also, $S_{xx} = \|(I-\Pi)X\|_2^2$.

So $\bar{X}$ only depends on ΠX, and S_{xx} only depends on $(I-\Pi)X$.

Note $\begin{pmatrix} \Pi X \\ (I-\Pi)X \end{pmatrix} \sim N\left(\begin{pmatrix} \Pi\mu = \mu\underline{1} \\ (I-\Pi)\underline{1}\mu \\ \text{zero} \end{pmatrix}, \begin{pmatrix} \Pi^2 = \Pi & \text{zero} \\ (I-\Pi)\Pi & \Pi(I-\Pi) \\ & (I-\Pi)^2 = (I-\Pi) \end{pmatrix} \sigma^2 \right)$

As off-diagonal blocks of covariance are zero, by Lemma, $\Pi X \perp (I-\Pi)X$.

$X \sim N\left(\begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}, I\sigma^2\right)$

Hence $\bar{X} \perp S_{xx}$, giving (iii).

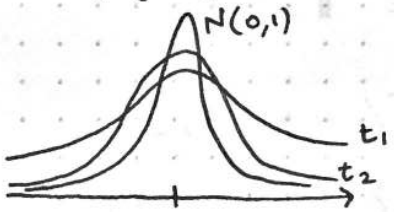
Furthermore, $\frac{S_{xx}}{\sigma^2} = \left\| (I-\Pi)\frac{X}{\sigma} \right\|_2^2 \sim \chi^2_p$

by Lemma, where $p = \text{rank}(I-\Pi) = n-1$. □

Student's t-distribution

Def^n If $Z \sim N(0,1)$, $Y \sim \chi^2_k$, $Z \perp Y$, then $T = \frac{Z}{\sqrt{Y/k}}$ has a t-distribution with k degrees of freedom, $T \sim t_k$.

The pdf of T is $f_T(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma(k)} \frac{1}{\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-k+\frac{1}{2}} \quad \forall\, t \in \mathbb{R}$.



Indeed, $\mathbb{P}(T \le z) \to \mathbb{P}(Z \le z)$ as $k \to \infty$ for every $z \in \mathbb{R}$.

Application to normal samples

Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then by Theorem,

$$\frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{S_{xx}/(n-1)\sigma^2}} \sim t_{n-1} \Rightarrow \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{S_{xx}/(n-1)}} \sim t_{n-1}$$

which does not depend on σ^2. This lets us define a $100(1-\alpha)\%$ confidence interval for μ when σ^2 is unknown.

$$\mathbb{P}\left(-t_{n-1}\left(\tfrac{\alpha}{2}\right) \le \frac{\sqrt{n}(\bar{X}-\mu)}{\tilde{\sigma}} \le t_{n-1}\left(\tfrac{\alpha}{2}\right) \right) = 1-\alpha$$

($\tilde{\sigma}$ ← def^n)

$$\Rightarrow \mathbb{P}\left(\bar{X} - t_{n-1}\left(\tfrac{\alpha}{2}\right)\frac{\tilde{\sigma}}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1}\left(\tfrac{\alpha}{2}\right)\frac{\tilde{\sigma}}{\sqrt{n}} \right) = 1-\alpha.$$

Hence $\bar{X} \pm t_{n-1}(\frac{\alpha}{2}) \frac{\tilde{\sigma}}{\sqrt{n}}$ is a $100(1-\alpha)\%$ CI for μ.

NB When σ^2 is known, we found a CI $\bar{X} \pm z(\frac{\alpha}{2}) \frac{\sigma}{\sqrt{n}}$. So in the new procedure, we replace σ by an estimator $\tilde{\sigma}$ and the normal quantile by a t_{n-1} quantile. This makes a big difference when n is small.

Linear models

- Observations or responses $Y_1, \dots, Y_n$ modelled as

$$Y_i = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i \qquad \text{for } i = 1, \dots, n$$

- $\beta_1, \dots, \beta_p$ are coefficients, the parameters of the model
- $x_{i1}, \dots, x_{ip}$ are fixed covariates or predictors or inputs
- $\varepsilon_1, \dots, \varepsilon_n$ are noise rvs

Ex Y_i # covid-19 cases in region i detected Jan 23–29

x_{i1} # daily flights to i from Wuhan

x_{i2} # tests conducted in i, Jan 23–29

- Matrix formulation Let $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$

and $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$, the design matrix. We will assume that X has full column rank, i.e. no covariate is a linear combination of the others. The model is $\underbrace{Y}_{\text{random}} = \underbrace{X\beta}_{\text{fixed}} + \underbrace{\varepsilon}_{\text{random}}$.

Moment assumptions for ε

① $\mathbb{E}\varepsilon = 0$

② $\mathrm{Cov}(\varepsilon) = \sigma^2 I_n$ for some $\sigma^2 \in (0, \infty)$ * homoskedasticity *

- i.e. the noise variables are uncorrelated with the same variance

Ordinary least-squares (OLS) estimator

Note $\mathbb{E}Y = \mathbb{E}(X\beta + \varepsilon) = X\beta$.

A reasonable way to estimate β is to minimise the residual sum of squares $S(\beta) = \sum_i (Y_i - (X\beta)_i)^2 = \|Y - X\beta\|_2^2$

(Y: observed responses; $X\beta$: mean of responses in model with param β)

The OLS estimator is the minimisor of $S(\beta)$ over $\mathbb{R}^p$, written $\hat{\beta}$.

- $\frac{\partial S(\beta)}{\partial \beta} = 2X^T(Y - X\beta)$ ← sign? so $2X^T(Y - X\hat{\beta}) = 0$

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

NB X^TX is invertible because X has rank p

Lemma $\mathbb{E}\hat{\beta} = \beta$ [$\hat{\beta}$ unbiased for β], $\operatorname{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

Proof $\mathbb{E}\hat{\beta} = \mathbb{E}((X^TX)^{-1}X^TY) = \mathbb{E}((X^TX)^{-1}X^T(X\beta+\varepsilon))$

$= \beta + \underbrace{(X^TX)^{-1}X^T\mathbb{E}(\varepsilon)}_{\text{zero}}$

$\operatorname{cov}\hat{\beta} = \operatorname{cov}(\underbrace{\beta}_{\text{const.}} + (X^TX)^{-1}X^T\varepsilon) = \operatorname{cov}((X^TX)^{-1}X^T\varepsilon)$

$= (X^TX)^{-1}X^T \operatorname{cov}(\varepsilon)\, X (X^TX)^{-1}$

$= \sigma^2 (X^TX)^{-1}$. □

Theorem (Gauss-Markov) Let $\hat{\beta}$ be the OLS estimator. Let $\beta^* = AY$ be a linear unbiased estimator of β. Then for all $t \in \mathbb{R}^p$,

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

The prediction of the linear model at any input t has the smallest variance with the OLS estimator. We say $\hat{\beta}$ is the Best Linear Unbiased Estimator (BLUE) of β.

Proof $\mathbb{E}(AY) = \mathbb{E}(A(X\beta+\varepsilon)) = AX\beta = \beta$ as β^* unbiased

Define $B = A - (X^TX)^{-1}X^T$, observe $BX = 0$.

Then $\operatorname{cov}(\beta^*) = \mathbb{E}[(\beta^*-\beta)(\beta^*-\beta)^T]$

$= \mathbb{E}[(\beta^* - \hat{\beta} + \hat{\beta} - \beta)(\underbrace{\beta^*-\hat{\beta}}_{BY} + \hat{\beta} - \beta)^T]$

$= \operatorname{cov}(\hat{\beta}) + \operatorname{cov}(BY)$

$+ \underbrace{\mathbb{E}[BY(\hat{\beta}-\beta)^T]}_{M} + \underbrace{\mathbb{E}[(\hat{\beta}-\beta)(BY)^T]}_{M^T}$.

$\operatorname{cov}(BY) = \operatorname{cov}(\underbrace{BX\beta}_{\text{zero}} + B\varepsilon) = \operatorname{cov}(B\varepsilon) = \sigma^2 BB^T$

$M = \mathbb{E}[B\varepsilon((X^TX)^{-1}X^T(X\beta+\varepsilon) - \beta)^T]$

$= \mathbb{E}[B\varepsilon((X^TX)^{-1}X^T\varepsilon)^T] = B\operatorname{cov}(\varepsilon)X(X^TX)^{-1}$

$= \sigma^2 \underbrace{BX}_{\text{zero}}(X^TX)^{-1} = 0$.

Hence $\operatorname{cov}(\beta^*) = \operatorname{cov}(\hat{\beta}) + \sigma^2 BB^T$.

Now for any $t \in \mathbb{R}^p$,

$$\operatorname{var}(t^T\beta^*) = t^T\operatorname{cov}(\beta^*)t = \operatorname{var}(t\hat{\beta}) + \overbrace{\sigma^2 t^T BB^T t}^{\geq 0}.$$

□

Def The fitted values in OLS estimator are the predictions of the response, $\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{P} Y$.

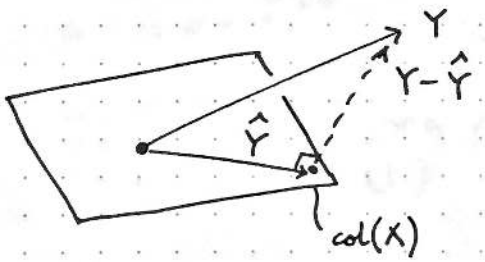
We call P the hat matrix. The residuals are $Y - \hat{Y} = (I - P)Y$.

Lemma P is an orthogonal projection matrix onto $\mathrm{col}(X)$.

By inspection, P is idempotent, symmetric. So an orthog projection.

P acts as the identity on $\mathrm{col}(X)$, maps onto $\mathrm{col}(X)$. □

$\mathbb{R}^n$:



Normal Linear model

From now on, we'll assume that $\varepsilon \sim N(0, \sigma^2 I)$.

This will make statistical inference much easier.

Normal linear model

We shall assume $\varepsilon \sim N(0, \sigma^2 I_p)$ [ε_i independent]

The log likelihood of this model is

$$\ell(\beta, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{\log 2\pi}{2} - \frac{\log \sigma^2}{2} - \frac{1}{2\sigma^2} (Y_i - (X\beta)_i)^2 \right)$$

$$= \text{constants} - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} S(\beta) \quad \leftarrow \|Y - X\beta\|_2^2$$

For any value of σ^2, $\ell(\beta, \sigma^2)$ is maximised as a function of β at the minimisor of $S(\beta)$.

So the MLE for β is $\hat{\beta} = (X^T X)^{-1} X^T Y$.

MLE of σ^2?

$$\frac{\partial \ell(\sigma^2, \beta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} S(\beta)$$

So at the MLE $(\hat{\sigma}^2, \hat{\beta})$,

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} S(\hat{\beta}) = 0 \implies \hat{\sigma}^2 = \frac{1}{n} S(\hat{\beta}) = \frac{\|Y - X\hat{\beta}\|_2^2}{n}$$

$$= \frac{\|(I-P)Y\|_2^2}{n}$$

Where $P = X(X^T X)^{-1} X^T$, the "hat matrix".

Theorem In the normal linear model,

(i) The fitted values PY, and the residuals $(I-P)Y$ are indpt

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

Proof The vector $\begin{pmatrix} PY \\ (I-P)Y \end{pmatrix} \in \mathbb{R}^{2n}$ is an affine function of ε, and hence is multivariate normal [check that for $t \in \mathbb{R}^{2n}$, $t^T \begin{pmatrix} PY \\ (I-P)Y \end{pmatrix}$ norm^L]

Furthermore, $\operatorname{cov}\begin{pmatrix} PY \\ (I-P)Y \end{pmatrix} = \operatorname{cov}\begin{pmatrix} P\varepsilon \\ (I-P)\varepsilon \end{pmatrix} = \sigma^2 \begin{pmatrix} P P^T & P(I-P)^T \\ (I-P)P^T & (I-P)(I-P)^T \end{pmatrix}$

Hence $PY = \hat{Y}$ is $\perp$ of $(I-P)Y$.

Note that $\hat{\sigma}^2$ is a function of $(I-P)Y$. Also, $\hat{\beta}$ is a function of PY. Indeed, $\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T PY$.

Therefore, $\hat{\beta} \perp \hat{\sigma}^2$.

Then $\hat{\beta}$ is a linear function of Y, therefore it is MVN.

Using the mean and covariance found in the last lecture, conclude $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

Finally, $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\|(I-P)Y\|_2^2 = \frac{1}{\sigma^2}\|(I-P)(X\beta+\varepsilon)\|_2^2$

$= \frac{1}{\sigma^2}\|(I-P)\varepsilon\|_2^2 = \|(I-P)\frac{\varepsilon}{\sigma}\|_2^2$

$\sim \chi^2_{n-p} \cdot 0.$ (rank $n-p$; $\frac{\varepsilon}{\sigma} \sim N(0, I_p)$) □

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

$\therefore \mathbb{E}[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2$, biased for small n

We'll define an unbiased estimator of σ^2, $\tilde{\sigma}^2 = \frac{n}{n-p}\hat{\sigma}^2$

$= \frac{\|(I-P)Y\|_2^2}{n-p}$

The F distribution

Def If $U \sim \chi^2_m$, $V \sim \chi^2_n$, $U \perp V$, then $X = \frac{U/m}{V/n}$ has an F-distribution with m & n degrees of freedom, $X \sim F_{m,n}$.

NB If $X \sim F_{1,n}$, then $\sqrt{X} \sim t_n$.

As a consequence of the theorem, the following is a pivot:

$$\frac{\frac{1}{p}(\hat{\beta}-\beta)^T(X^T*X)(\hat{\beta}-\beta)}{\tilde{\sigma}^2} = \frac{\frac{1}{p\sigma^2}(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta)}{\tilde{\sigma}^2/\sigma^2}$$

- Numerator (function of $\hat{\beta}$) $\perp$ denominator (function of $\tilde{\sigma}^2$)
- The denominator $\tilde{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}/n-p$
- $(\hat{\beta}-\beta) \sim N(0, (X^TX)^{-1}\sigma^2)$ Let $X^TX = UDU^T$, eigendecomp. define $(X^TX)^{1/2} = UD^{1/2}U^T$. Then $(\hat{\beta}-\beta)^T(X^TX)^{1/2} \sim N(0, \sigma^2 I_p)$.

Hence $\frac{1}{p\sigma^2}(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta)$

$= \frac{1}{p\sigma^2}\left[(\hat{\beta}-\beta)^T(X^TX)^{1/2}\right]\left[(X^TX)^{1/2}(\hat{\beta}-\beta)\right] \sim \chi^2_p/p$

Putting these together, the pivot has $F_{p,n-p}$ distribution.

Confidence sets for β

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

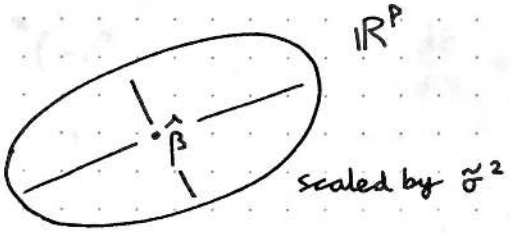
($F_{p,n-p}(\alpha)$: upper α point)

The set of β which satisfy the inequality

$$C = \{ b \in \mathbb{R}^p : (\hat{\beta} - b)^T X^T X (\hat{\beta} - b) \leq F_{p,n-p}(\alpha)\, \tilde{\sigma}^2 p \}$$

is an ellipsoid.

$\Rightarrow \mathbb{P}_\beta(\beta \in C) = 1 - \alpha$ hence C is $100(1-\alpha)\%$ confidence set for β. ☺



Inference for a single coeff. β_j

By the same argument, $(\hat{\beta}_j - \beta_j)^2 / [(X^TX)^{-1}]_{jj} \tilde{\sigma}^2 \sim F_{1,n-p}$

Or, taking the square root,

$$\frac{\hat{\beta}_j - \beta_j}{[(X^TX)^{-1}]_{jj}^{1/2} \tilde{\sigma}} \sim t_{n-p}$$

Hence, $\hat{\beta}_j \pm t_{n-p}(\frac{\alpha}{2})\, \tilde{\sigma} [(X^TX^{+1})^{-1}]_{jj}^{1/2}$ is an exact $100(1-\alpha)\%$ confidence interval for β_j.

Hypothesis Testing in the Normal Linear Model

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

Suppose $\underset{n\times p}{X} = [\underset{n\times p_0}{X_0}\ \underset{n\times(p-p_0)}{X_1}]$, and $\beta = \begin{pmatrix} \beta_0 \to \text{in } \mathbb{R}^{p_0} \\ \beta_1 \to \text{in } \mathbb{R}^{p-p_0} \end{pmatrix}$,

where $\text{rank}(X) = p$ and $\text{rank}(X_0) = p_0$.

Want to test $H_0 : \beta_1 = 0$ [predictors in X_1 are not associated to response]

$$H_1 : \beta_1 \neq 0$$

Under H_0, the linear model is $Y = X_0\beta_0 + \varepsilon$, so the MLEs are

$$\hat{\hat{\beta}}_0 = (X_0^T X_0)^{-1} X_0^T Y, \quad \hat{\hat{\beta}}_1 = 0, \quad \hat{\hat{\sigma}}^2 = \frac{\|Y - X_0\hat{\hat{\beta}}_0\|_2^2}{n} = \frac{1}{n}\|(I-P_0)Y\|_2^2 =: \frac{RSS_0}{n}$$

where $P_0 = X_0(X_0^T X_0)^{-1} X_0^T$ is the orthogonal projection onto $\text{col}(X_0)$.

The generalised likelihood ratio is

$$\Lambda_Y(H_0, H_1) = \frac{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}\|Y - X\hat{\beta}\|_2^2\right)}{(2\pi\hat{\hat{\sigma}}^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\hat{\hat{\sigma}}^2}\|Y - X_0\hat{\hat{\beta}}_0\|_2^2\right)} = \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{n/2}$$

$$= \left(\frac{RSS_0}{RSS}\right)^{n/2} = \left(1 + \frac{RSS_0 - RSS}{RSS}\right)^{n/2}$$

So generalised LR test rejects when $\Lambda_Y(H_0, H_1)$ is large.

A version of Wilks' theorem says that if p, p_0 fixed, under H_0, as $n \to \infty$, $2\log\Lambda_Y(H_0, H_1)$ is approximately $\chi^2_{p-p_0}$.

$\Lambda_Y(H_0, H_1)$ is monotone in the statistic

$$F =: \frac{(RSS_0 - RSS)/(p-p_0)}{RSS/(n-p)}$$

and we can find the exact distribution of F under H_0.

Thm Under H_0, $F \sim F_{p-p_0, n-p}$

Proof It's enough to show $RSS \perp RSS_0 - RSS$, $RSS \sim \chi^2_{n-p}$ (scaled by σ^2), $RSS_0 - RSS \sim \chi^2_{p-p_0}$ (scaled by σ^2)

Note $RSS_0 - RSS = \|(I-P_0)Y\|_2^2 - \|(I-P)Y\|_2^2$

$$= Y^T(I-P_0)Y - Y^T(I-P)Y$$

$$= Y^T(P-P_0)Y = \|(P-P_0)Y\|_2^2$$

where in the last identity we used that $P - P_0$ is idempotent symmetric.

Symetry is clear from that of P, P_0.

Also, $(P-P_0)(P-P_0) = P^2 - P_0P - PP_0 + P_0^2 = P + P_0 - P_0 - P_0 = P - P_0$. (# epic)

Thus $P - P_0$ is an orthogonal projection matrix and

$$\text{rank}(P-P_0) = \text{tr}(P-P_0) = \text{tr}\,P - \text{tr}\,P_0 = \text{rank}\,P - \text{rank}\,P_0$$
$$= \text{rank}(X) - \text{rank}(X_0) = p - p_0$$

Under H_0, $Y = X_0\beta_0 + \varepsilon$ (dies)

$$RSS_0 - RSS = \|(P-P_0)Y\|_2^2 = \|(P-P_0)(X_0\beta_0+\varepsilon)\|_2^2$$
$$= \|(P-P_0)\varepsilon\|_2^2$$

$$RSS = \|(I-P)Y\|_2^2 = \|(I-P)(X_0\beta_0+\varepsilon)\|_2^2$$
$$= \|(I-P)\varepsilon\|_2^2$$

Thus by a lemma in L13, $RSS_0 - RSS \sim \chi^2_{p-p_0}\sigma^2$

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

Finally, note $\begin{pmatrix}(P-P_0)\varepsilon\\(I-P)\varepsilon\end{pmatrix} \sim N\left(0, \sigma^2\begin{pmatrix}(P-P_0)(P-P_0)^T & (P-P_0)(I-P)^T\\(I-P)(P-P_0)^T & (I-P)(I-P)^T\end{pmatrix}\right)$

where $(I-P)(P-P_0) = (P-P_0)(I-P) = 0$ because they project onto orthogonal subspaces of $\mathbb{R}^n$.

So $(P-P_0)\varepsilon \perp (I-P)\varepsilon$ and the result follows. □

So the generalised LRT of size α rejects H_0 when

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

How to interpret F?

RSS is the squared error of the full model

$RSS_0 - RSS$ is the reduction in squared error when we include X_1

The statistic $R^2 = \frac{RSS_0 - RSS}{RSS}$ is "the proportion of the variance in the responses explained by the predictors in X_1"

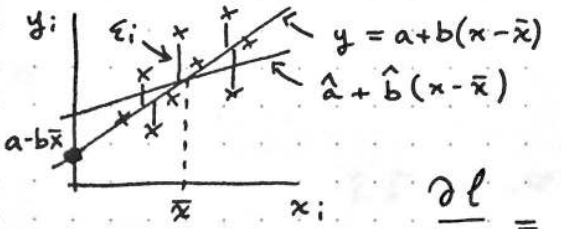
The F statistic is a version of R^2 scaled by degrees of freedom

Simple linear regression

$Y_i = a + b(x_i - \bar{x}) + \varepsilon_i$ where $\bar{x} = \frac{1}{n}\sum x_i$, $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

2 parameters a, b and also σ^2

Ex Sheet 3 How to define the design matrix X for this model.



The log-likelihood is

$$\ell(a, b, \sigma^2) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\quad \right)^2 \right.$$

$Y_i - a - b(x_i - \bar{x})$

$$\frac{\partial \ell}{\partial a} = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - a - b(x_i - \bar{x})\right)$$

$$= \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a)$$

this is zero at $\hat{a} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

$$\frac{\partial \ell}{\partial b} = \sum_{i=1}^{n} \frac{1}{2\sigma^2}\left(Y_i - a - b(x_i - \bar{x})\right)(x_i - \bar{x})$$

Plug in $\hat{a}$ for a, set to zero, solve for b

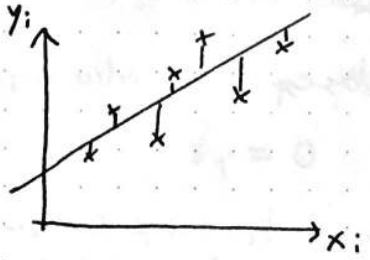
$$\hat{b} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Applications of the general normal linear model

● ① Simple linear model $Y_i = a + b(x_i - \bar{x}) + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Last time derived the MLEs $\hat{a} = \bar{Y}$

$$\hat{b} = \frac{\sum (Y_i - \bar{Y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$



We wish to test $H_0: b = 0$ vs $H_1: b \neq 0$.

Under H_0, the MLE is $\hat{\hat{a}} = \bar{Y}$, $\hat{\hat{b}} = 0$.

The generalised LR test rejects when F is large,

where $F = \dfrac{(RSS_0 - RSS)/p - p_0}{RSS/n - p}$.

● In this case $RSS_0 = \sum_i (Y_i - \bar{Y})^2$, $RSS = \sum_i \left(Y_i - \bar{Y} - \hat{b}(x_i - \bar{x})\right)^2$

$$R^2 = \frac{RSS_0 - RSS}{RSS_0} = \frac{\left[\sum (Y_i - \bar{Y})(X_i - \bar{X})\right]^2}{\sum (Y_i - \bar{Y})^2 \sum (X_i - \bar{X})^2}$$

← uhidk? (oh ok he made a mistake)

this is the square of Pearson's correlation coefficient.

The F statistic $F = R^2 \frac{n-p}{p-p_0} = R^2 \cdot \frac{n-2}{2-1}$ has, under H_0, an $F_{p-p_0, n-p} = F_{1, n-2}$ distribution.

② Analysis of Variance (ANOVA)

Example Patients in a clinical trial are randomly assigned to groups

● 1,.., I each of size J (balanced design)

Y_{ij} response of patient j in group i

$Y_{ij} = \alpha + \mu_i + \varepsilon_{ij}$, $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

We'll assume (WLOG) that group 1 is a "control" group, let $\mu_1 = 0$.

How to interpret parameters:

- α is the mean response in the control group,
- μ_i is the difference in mean response between group i and the control group.

● How to write this as a linear model $Y = X\beta + \varepsilon$?

$$Y = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1J} \\ Y_{21} \\ \vdots \\ Y_{IJ} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 & 0 \\ 1 & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 1 & 0 & \cdots & \cdots & 0 & 0 \\ 1 & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_I \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1J} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{IJ} \end{pmatrix}$$

The matrix has one column with all 1s, the subsequent columns are indicators (0,1) for groups $2, 3, \ldots, I$.

Equality of means test $H_0: \mu_2 = \mu_3 = \cdots = \mu_I = 0$ vs $H_1: \mu_i \neq 0$ for $i \neq 1$

(experimental treatments have no effect)

Let $X_0 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, $X = (X_0 \; X_1)$, $P = X(X^TX)^{-1}X^T$,

$P_0 = X_0(X_0^TX_0)^{-1}X_0^T$

$$PY = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_1 \\ \bar{Y}_2 \\ \vdots \\ \bar{Y}_I \end{pmatrix}, \quad \bar{Y}_i = \sum_j \frac{Y_{ij}}{J} \quad \text{"the average response for } i\text{"}$$

$$P_0Y = \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}, \quad \bar{Y} = \sum_i \frac{\bar{Y}_i}{I} = \sum_{i,j} \frac{Y_{ij}}{IJ}$$

$$F = \frac{(RSS_0 - RSS)/(p - p_0)}{RSS/(n-p)} = \frac{\frac{1}{p-p_0}\| (P_0 - P)Y \|_2^2}{\frac{1}{n-p}\| (I-P)Y \|_2^2}$$

$n = IJ$, $p = I$, $p_0 = 1$

$$F = \frac{I(J-1)}{I-1} \cdot \frac{\sum_{i,j} (\bar{Y} - \bar{Y}_i)^2}{\sum_{i,j} (Y_{ij} - \bar{Y}_i)^2}$$

numerator → "between group" variance; denominator → "within group" variance

The statistic has, under H_0, an $F_{p-p_0, n-p} = F_{I-1, I(J-1)}$ dist^n

Choice of control group If we chose another control group, $col(X)$, $col(X_0)$ would be the same. Thus, P, P_0 would be the same, the fitted values PY, P_0Y would be the same (under H_0, H_1), and the F statistic would be the same.

The interpretation of the parameters α, μ changes.

③ Paired observations Consider an ANOVA model with $I = 2$, but the people in the trial come in pairs and one member of each pair is assigned to each group, (at random).

Y_{ij} = response for person in j^{th} pair, group i

$$Y_{ij} = \alpha + \mu_{X_i} + \gamma_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

$\mu_1 = 0$, $\gamma_1 = 0$

We care about the ~~estimates~~ parameter μ_2, not about $\gamma_2, \ldots, \gamma_J$.

We can test $H_0: \mu_2 = 0$ vs $H_1: \mu_2 \neq 0$ using the F statistic