Lecture 1. Introduction and probability review

What is "Statistics"?

There are many definitions: I will use

"A set of principle and procedures for gaining and processing quantitative evidence in order to help us make judgements and decisions"

It can include

- Design of experiments and studies
- Exploring data using graphics
- Informal interpretation of data
- Formal statistical analysis
- Clear communication of conclusions and uncertainty

It is NOT just data analysis!

In this course we shall focus on formal *statistical inference*: we assume

- we have data generated from some unknown probability model
- we aim to use the data to learn about certain properties of the underlying probability model

Idea of parametric inference

- Let X be a random variable (r.v.) taking values in $\mathcal{X}$
- Assume distribution of X belongs to a family of distributions indexed by a scalar or vector parameter θ, taking values in some parameter space Θ
- Call this a *parametric family:*

For example, we could have

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

BIG ASSUMPTION

For some results (bias, mean squared error, linear model) we do not need to specify the precise parametric family.

But generally we assume that we know which family of distributions is involved, but that the value of θ is unknown.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) with the same distribution as X, so that $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a simple random sample (our data).

We use the observed $\mathbf{X} = \mathbf{x}$ to make inferences about θ, such as,

(a) giving an estimate $\hat{\theta}(\mathbf{x})$ of the true value of θ (point estimation);

(b) giving an interval estimate $(\hat{\theta}_1(\mathbf{x}), (\hat{\theta}_2(\mathbf{x}))$ for θ;

(c) testing a hypothesis about θ, eg testing the hypothesis $H : \theta = 0$ means determining whether or not the data provide evidence against H.

We shall be dealing with these aspects of *statistical inference*.

Other tasks (not covered in this course) include

- Checking and selecting probability models
- Producing predictive distributions for future random variables
- Classifying units into pre-determined groups ('supervised learning')
- Finding clusters ('unsupervised learning')

Statistical inference is needed to answer questions such as:

- What are the voting intentions before an election? *[Market research, opinion polls, surveys]*
- What is the effect of obesity on life expectancy? *[Epidemiology]*
- What is the average benefit of a new cancer therapy? *Clinical trials*
- What proportion of temperature change is due to man? *Environmental statistics*
- What is the benefit of speed cameras? *Traffic studies*
- What portfolio maximises expected return? *Financial and actuarial applications*
- How confident are we the Higgs Boson exists? *Science*
- What are possible benefits and harms of genetically-modified plants? *Agricultural experiments*
- What proportion of the UK economy involves prostitution and illegal drugs? *Official statistics*
- What is the chance Liverpool will best Arsenal next week? *Sport*

Probability review

Let Ω be the *sample space* of all possible outcomes of an experiment or some other data-gathering process.

E.g when flipping two coins, $\Omega = \{HH, HT, TH, TT\}$.

'Nice' (measurable) subsets of Ω are called *events*, and $\mathcal{F}$ is the set of all events - when Ω is countable, $\mathcal{F}$ is just the power set (set of all subsets) of Ω.

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

- $\mathbb{P}(\phi) = 0$
- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$, whenever $\{A_n\}$ is a disjoint sequence of events.

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

Thus for the two coins, we might set

$X(HH) = 2, XX(HT) = 1, X(TH) = 1, X(TT) = 0,$

so X is simply the number of heads.

Our data are modelled by a vector $\mathbf{X} = (X_1, \ldots, X_n)$ of random variables – each observation is a random variable.

The *distribution function* of a r.v. X is $F_X(x) = \mathbb{P}(X \leq x)$, for all $x \in \mathbb{R}$. So F_X is

- non-decreasing,
- $0 \leq F_X(x) \leq 1$ for all x,
- $F_X(x) \to 1$ as $x \to \infty$,
- $F_X(x) \to 0$ as $x \to -\infty$.

A *discrete* random variable takes values only in some countable (or finite) set $\mathcal{X}$, and has a *probability mass function* (pmf) $f_X(x) = \mathbb{P}(X = x)$.

- $f_X(x)$ is zero unless x is in $\mathcal{X}$.
- $f_X(x) \geq 0$ for all x,
- $\sum_{x \in \mathcal{X}} f_X(x) = 1$
- $\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x)$ for a set A.

We say X has a continuous (or, more precisely, absolutely continuous) distribution if it has a *probability density function* (pdf) f_X such that

- $\mathbb{P}(X \in A) = \int_A f_X(t)dt$ for "nice" sets A.

Thus

- $\int_{-\infty}^{\infty} f_X(t)dt = 1$
- $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$

[Notation note: There will be inconsistent use of a subscript in mass, density and distributions functions to denote the r.v. Also f will sometimes be p.]

Expectation and variance

If X is discrete, the *expectation* of X is

$$\mathbb{E}(X) = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x)$$

(exists when $\sum |x|\mathbb{P}(X = x) < \infty$).

If X is continuous, then

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

(exists when $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$).

$\mathbb{E}(X)$ is also called the expected value or the mean of X.

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

$$\mathbb{E}(g(X)) = \begin{cases} \sum_{x \in \mathcal{X}} g(x)\mathbb{P}(X = x) & \text{if } X \text{ is discrete} \\ \\ \int g(x) f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

The *variance* of X is $var(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}\big(X^2\big) - \big(\mathbb{E}(X)\big)^2$.

Independence

The random variables $X_1, \ldots, X_n$ are *independent* if for all $x_1, \ldots, x_n$,

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

If the independent random variables $X_1, \ldots, X_n$ have pdf's or pmf's $f_{X_1}, \ldots, f_{X_n}$, then the random vector $\mathbf{X} = (X_1, \ldots, X_n)$ has pdf or pmf

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_i f_{X_i}(x_i).$$

Random variables that are independent and that all have the same distribution (and hence the same mean and variance) are called *independent and identically distributed (iid) random variables*.

Maxima of iid random variables

Let $X_1, \ldots, X_n$ be iid r.v.'s, and $Y = \max(X_1, \ldots, X_n)$.

Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\max(X_1, \ldots, X_n) \leq y)$$

$$= \mathbb{P}(X_1 \leq y, \ldots, X_n \leq y) = \mathbb{P}(X_i \leq y)^n = [F_X(y)]^n$$

The density for Y can then be obtained by differentiation (if continuous), or differencing (if discrete).

Can do similar analysis for minima of iid r.v.'s.

Sums and linear transformations of random variables

For any random variables,

$$\begin{aligned}
\mathbb{E}(X_1 + \cdots + X_n) &= \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) \\
\mathbb{E}(a_1 X_1 + b_1) &= a_1 \mathbb{E}(X_1) + b_1 \\
\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) &= a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n) \\
\text{var}(a_1 X_1 + b_1) &= a_1^2 \text{var}(X_1)
\end{aligned}$$

For independent random variables,

$$\mathbb{E}(X_1 \times \ldots \times X_n) = \mathbb{E}(X_1) \times \ldots \times \mathbb{E}(X_n),$$

$$\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n),$$

and

$$\text{var}(a_1 X_1 + \cdots + a_n X_n) = a_1^2 \text{var}(X_1) + \cdots + a_n^2 \text{var}(X_n).$$

Standardised statistics

Suppose $X_1, \ldots, X_n$ are iid with $\mathbb{E}(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2$.

Write their sum as

$$S_n = \sum_{i=1}^{n} X_i$$

From preceding slide, $\mathbb{E}(S_n) = n\mu$ and $\text{var}(S_n) = n\sigma^2$.

Let $\bar{X}_n = S_n/n$ be the sample mean.

Then $\mathbb{E}(\bar{X}_n) = \mu$ and $\text{var}(\bar{X}_n) = \sigma^2/n$.

Let

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}.$$

Then $\mathbb{E}(Z_n) = 0$ and $\text{var}(Z_n) = 1$.

Z_n is known as a *standardised statistic*.

Moment generating functions

The *moment generating function* for a r.v. X is

$$M_X(t) = \mathbb{E}(e^{tX}) = \begin{cases} \sum_{x \in \mathcal{X}} e^{tx} \mathbb{P}(X = x) & \text{if } X \text{ is discrete} \\ \int e^{tx} f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

provided M exists for t in a neighbourhood of 0.

Can use this to obtain moments of X, since

$$\mathbb{E}(X^n) = M_X^{(n)}(0),$$

i.e. nth derivative of M evaluated at $t = 0$.

Under broad conditions, $M_X(t) = M_Y(t)$ implies $F_X = F_Y$.

Mgf's are useful for proving distributions of sums of r.v.'s since, if $X_1, ..., X_n$ are iid, $M_{S_n}(t) = M_X^n(t)$.

Example: sum of Poissons

If $X_i \sim \text{Poisson}(\mu)$, then

$$M_{X_i}(t) = \mathbb{E}(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} e^{-\mu} \mu^x / x! = e^{-\mu(1-e^t)} \sum_{x=0}^{\infty} e^{-\mu e^t} (\mu e^t)^x / x! = e^{-\mu(1-e^t)}.$$

And so $M_{S_n}(t) = e^{-n\mu(1-e^t)}$, which we immediately recognise as the mgf of a Poisson$(n\mu)$ distribution.

So sum of iid Poissons is Poisson. □

Convergence

The Weak Law of Large Numbers (WLLN) states that for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

The Strong Law of Large Numbers (SLLN) says that

$$\mathbb{P}\left(\bar{X}_n \to \mu\right) = 1.$$

The Central Limit Theorem tells us that

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \text{ is approximately } \mathrm{N}(0,1) \text{ for large } n \text{ .}$$

Conditioning

Let X and Y be discrete random variables with joint pmf

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

Then the *marginal pmf* of Y is

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

The *conditional pmf* of X given $Y=y$ is

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

if $p_Y(y) \neq 0$ (and is defined to be zero if $p_Y(y)=0$)).

Conditioning

In the continuous case, suppose that X and Y have joint pdf $f_{X,Y}(x,y)$, so that for example

$$\mathbb{P}(X \leq x_1, Y \leq y_1) = \int_{-\infty}^{y_1} \int_{-\infty}^{x_1} f_{X,Y}(x,y) dx dy.$$

Then the *marginal pdf* of Y is

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

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

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

if $f_Y(y) \neq 0$ (and is defined to be zero if $f_Y(y) = 0$).

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

$$\mathbb{E}(X \mid Y=y) = \begin{cases} \sum x f_{X|Y}(x \mid y) & \text{pmf} \\ \\ \int x f_{X|Y}(x \mid y) dx & \text{pdf.} \end{cases}$$

Thus $\mathbb{E}(X \mid Y=y)$ is a function of y, and $\mathbb{E}(X \mid Y)$ is a function of Y and hence a r.v..

The conditional expectation formula says

$$\mathbb{E}[X] = \mathbb{E}\left[\mathbb{E}(X \mid Y)\right].$$

Proof [discrete case]:

$$\mathbb{E}\left[\mathbb{E}(X \mid Y)\right] = \sum_{\mathcal{Y}} y \left[\sum_{\mathcal{X}} x \; f_{X|Y}(x \mid y)\right] f_Y(y) = \sum_{\mathcal{X}} \sum_{\mathcal{Y}} x \; y \; f_{X,Y}(x,y)$$

$$= \sum_{\mathcal{X}} x \left[\sum_{\mathcal{Y}} y \; f_{Y|X}(y \mid x)\right] f_X(x) = \sum_{\mathcal{X}} x \; f_X(x). \square$$

The conditional variance of X given $Y = y$ is defined by

$$\text{var}(X \mid Y=y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X \mid Y=y)\big)^2 \mid Y = y\Big],$$

and this is equal to $\mathbb{E}(X^2 \mid Y=y) - \big(\mathbb{E}(X \mid Y=y)\big)^2$.
We also have the conditional variance formula:

$$\text{var}(X) = \mathbb{E}[\text{var}(X \mid Y)] + \text{var}[\mathbb{E}(X \mid Y)].$$

Proof:

$$\begin{aligned}
\text{var}(X) &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \\
&= \mathbb{E}\Big[\mathbb{E}(X^2 \mid Y)\Big] - \Big[\mathbb{E}\Big[\mathbb{E}(X \mid Y)\Big]\Big]^2 \\
&= \mathbb{E}\Big[\mathbb{E}(X^2 \mid Y) - \big[\mathbb{E}(X \mid Y)\big]^2\Big] + \mathbb{E}\Big[\big[\mathbb{E}(X \mid Y)\big]^2\Big] - \Big[\mathbb{E}\Big[\mathbb{E}(X \mid Y)\Big]\Big]^2 \\
&= \mathbb{E}\Big[\text{var}(X \mid Y)\Big] - \text{var}\Big[\mathbb{E}(X \mid Y)\Big].
\end{aligned}$$

Some important discrete distributions: Binomial

X has a **binomial** distribution with parameters n and p ($n \in \mathbb{N}$, $0 \leq p \leq 1$), $X \sim \text{Bin}(n, p)$, if

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

(zero otherwise).

We have $\mathbb{E}(X) = np$, $\text{var}(X) = np(1-p)$.

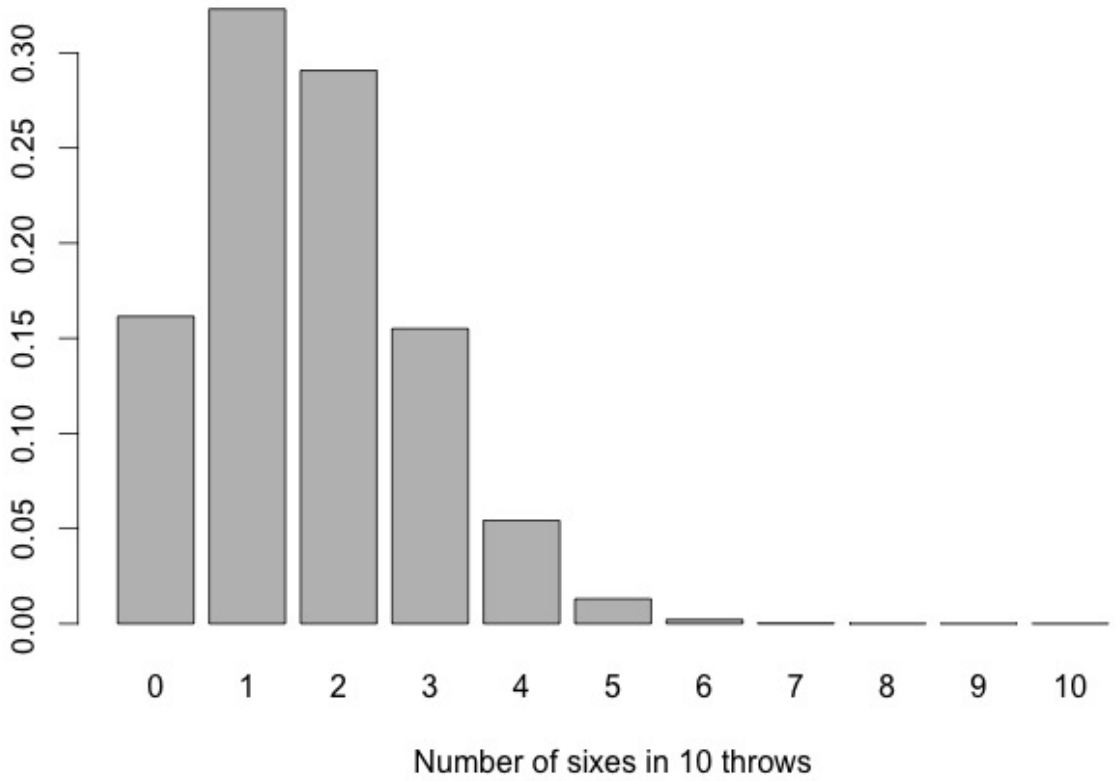
This is the distribution of the number of successes out of n independent Bernoulli trials, each of which has success probability p.

Example: throwing dice

let X = number of sixes when throw 10 fair dice, so $X \sim \text{Bin}(10, \frac{1}{6})$

R code:

```
barplot(  dbinom(0:10, 10, 1/6), names.arg=0:10,
                xlab="Number of sixes in 10 throws" )
```



Some important discrete distributions: Poisson

X has a **Poisson** distribution with parameter μ ($\mu > 0$), $X \sim \text{Poisson}(\mu)$, if

$$\mathbb{P}(X = x) = e^{-\mu}\mu^x/x!, \text{ for } x \in \{0, 1, 2, \ldots\},$$

(zero otherwise).

Then $\mathbb{E}(X) = \mu$ and $\text{var}(X) = \mu$.

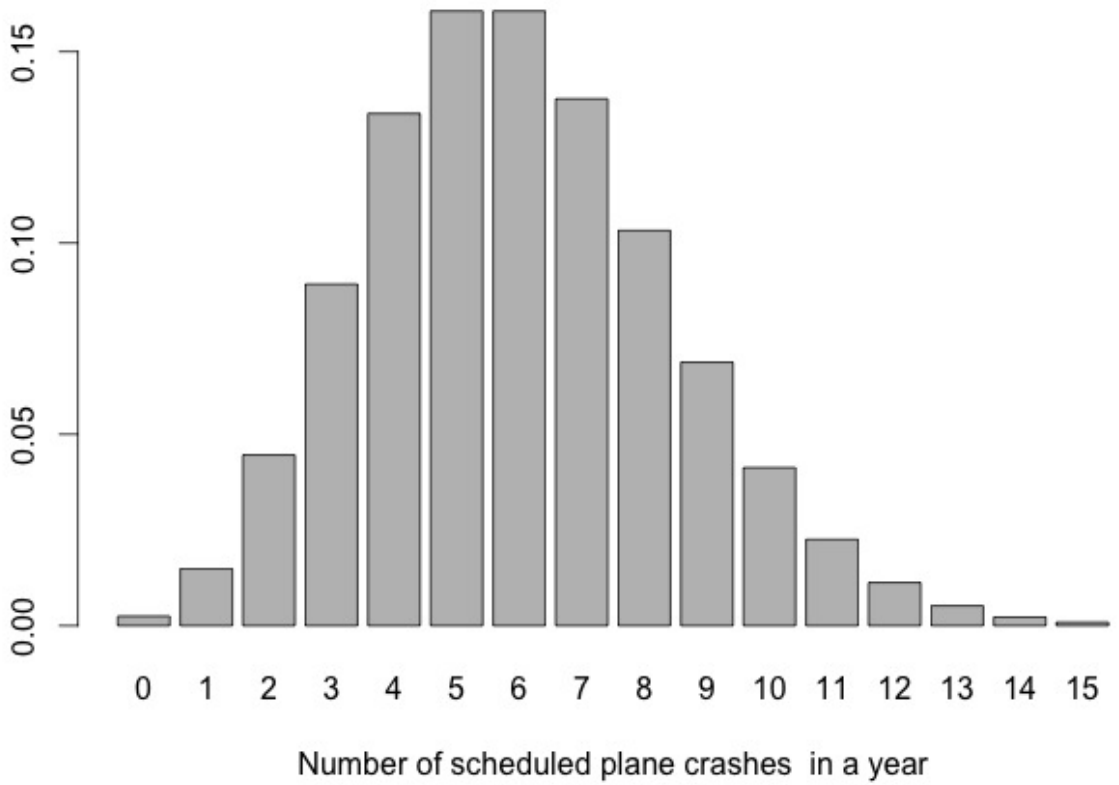
In a *Poisson process* the number of events $X(t)$ in an interval of length t is Poisson(μt), where μ is the rate per unit time.

The Poisson(μ) is the limit of the Bin(n,p) distribution as $n \to \infty$, $p \to 0$, $\mu = np$.

Example: plane crashes. Assume scheduled plane crashes occur as a Poisson process with a rate of 1 every 2 months. How many (X) will occur in a year (12 months)?

Number in two months is Poisson(1), and so $X \sim$ Poisson(6).

```
barplot(  dpois(0:15, 6), names.arg=0:15,
              xlab="Number of scheduled plane crashes  in a year" )
```



Some important discrete distributions: Negative Binomial

X has a **negative binomial** distribution with parameters k and p ($k \in \mathbb{N}$, $0 \leq p \leq 1$), if

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

(zero otherwise). Then $\mathbb{E}(X) = k/p$, $\text{var}(X) = k(1-p)/p^2$. This is the distribution of the number of trials up to and including the kth success, in a sequence of independent Bernoulli trials each with success probability p.

The negative binomial distribution with $k = 1$ is called a **geometric** distribution with parameter p.

The r.v $Y = X - k$ has

$$\mathbb{P}(Y = y) = \binom{y+k-1}{k-1}(1-p)^y p^k, \text{ for } y = 0, 1, \ldots.$$

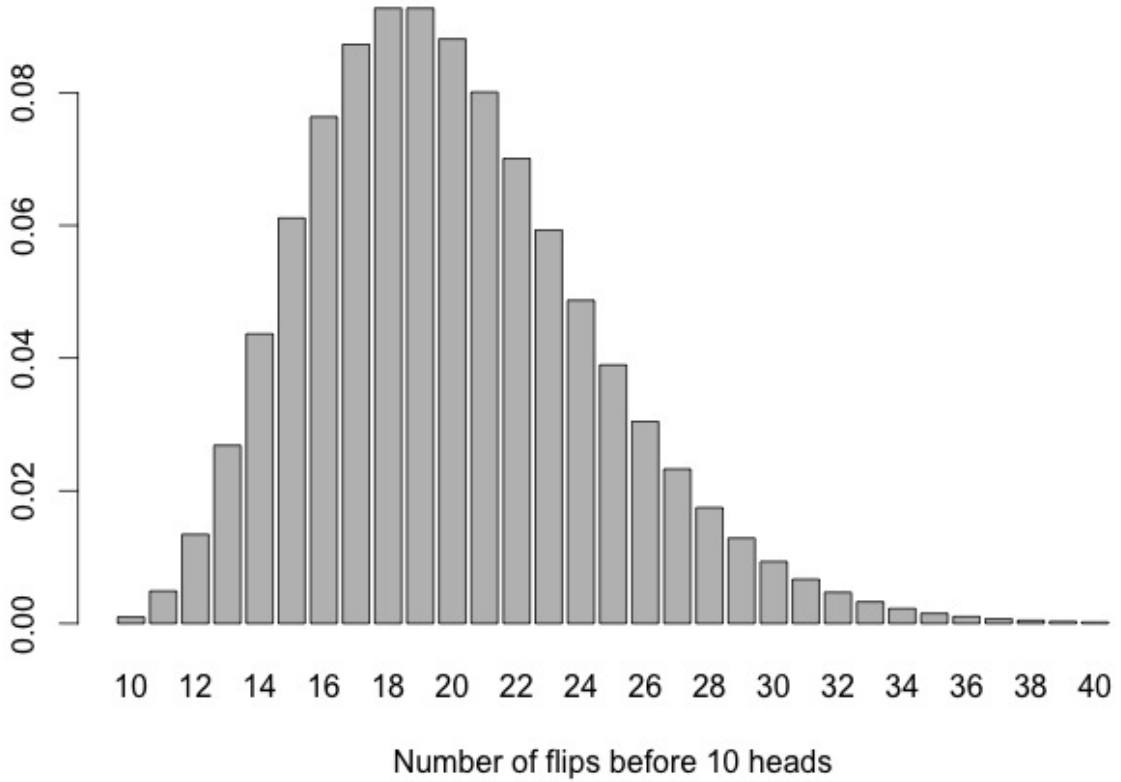
This is the distribution of the number of failures before the kth success in a sequence of independent Bernoulli trials each with success probability p. It is *also* sometimes called the negative binomial distribution: be careful!

Example: How many times do I have to flip a coin before I get 10 heads?

This is first (X) definition of the Negative Binomial since it includes all the flips.

R uses second definition (Y) of Negative Binomial, so need to add in the 10 heads:

```
barplot(  dnbinom(0:30, 10, 1/2), names.arg=0:30 + 10,
          xlab="Number of flips before 10 heads" )
```



Some important discrete distributions: Multinomial

Suppose we have a sequence of n independent trials where at each trial there are k possible outcomes, and that at each trial the probability of outcome i is p_i.

Let N_i be the number of times outcome i occurs in the n trials and consider $N_1, \ldots, N_k$. They are discrete random variables, taking values in $\{0, 1, \ldots, n\}$.

This **multinomial** distribution with parameters n and $p_1, \ldots, p_k$, $n \in \mathbb{N}$, $p_i \geq 0$ for all i and $\sum_i p_i = 1$ has joint pmf

$$\mathbb{P}(N_1 = n_1, \ldots, N_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}, \qquad \text{if } \sum_i n_i = n,$$

and is zero otherwise.

The rv's $N_1, \ldots, N_k$ are not independent, since $\sum_i N_i = n$.

The marginal distribution of N_i is Binomial(n, p_i).

Example: I throw 6 dice: what is the probability that I get one of each face 1,2,3,4,5,6? Can calculate to be $\frac{6!}{1!\ldots1!} \left(\frac{1}{6}\right)^6 = 0.015$

```
dmultinom( x=c(1,1,1,1,1,1), size=6, prob=rep(1/6,6))
```

Some important continuous distributions: Normal

X has a **normal** (Gaussian) distribution with mean μ and variance σ^2 ($\mu \in \mathbb{R}$, $\sigma^2 > 0$), $X \sim N(\mu, \sigma^2)$, if it has pdf

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

We have $\mathbb{E}(X) = \mu$, $\text{var}(X) = \sigma^2$.

If $\mu = 0$ and $\sigma^2 = 1$, then X has a **standard normal** distribution, $X \sim N(0, 1)$. We write ϕ for the standard normal pdf, and Φ for the standard normal distribution function, so that

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right), \qquad \Phi(x) = \int_{-\infty}^{x} \phi(t)dt.$$

The upper $100\alpha\%$ point of the standard normal distribution is z_α where

$$\mathbb{P}(Z > z_\alpha) = \alpha, \text{ where } Z \sim N(0, 1).$$

Values of Φ are tabulated in normal tables, as are percentage points z_α.

Some important continuous distributions: Uniform

X has a **uniform** distribution on $[a, b]$, $X \sim U[a, b]$ $(-\infty < a < b < \infty)$, if it has pdf

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

Then $\mathbb{E}(X) = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a)^2}{12}$.

Some important continuous distributions: Gamma

X has a **Gamma** (α, λ) distribution ($\alpha > 0$, $\lambda > 0$) if it has pdf

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

where $\Gamma(\alpha)$ is the gamma function defined by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ for $\alpha > 0$.
We have $\mathbb{E}(X) = \frac{\alpha}{\lambda}$ and $\text{var}(X) = \frac{\alpha}{\lambda^2}$.
The moment generating function $M_X(t)$ is

$$M_X(t) = \mathbb{E}\left(e^{Xt}\right) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha, \qquad \text{for } t < \lambda.$$

Note the following two results for the gamma function:
(i) $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$,
(ii) if $n \in \mathbb{N}$ then $\Gamma(n) = (n-1)!$.

Some important continuous distributions: Exponential

X has an **exponential** distribution with parameter λ ($\lambda > 0$) if $X \sim \text{Gamma}(1, \lambda)$, so that X has pdf

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

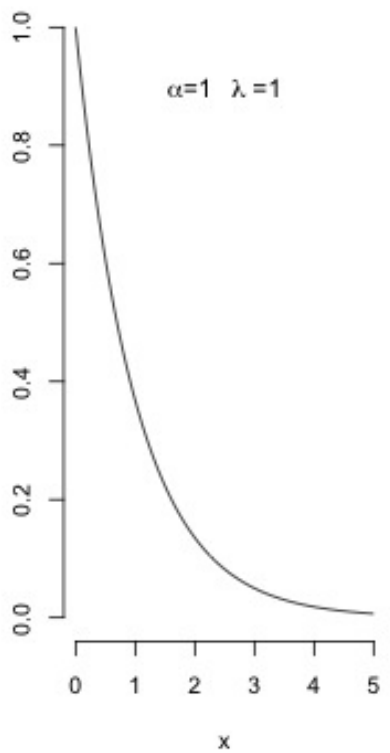
Then $\mathbb{E}(X) = \frac{1}{\lambda}$ and $\text{var}(X) = \frac{1}{\lambda^2}$.

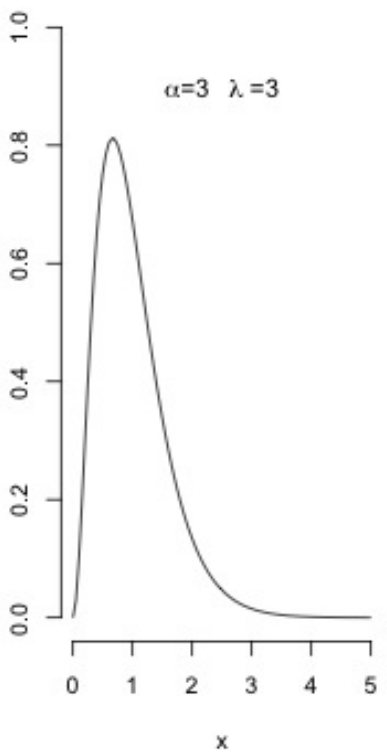
Note that if $X_1, \ldots, X_n$ are iid Exponential(λ) r.v's then $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \lambda)$.

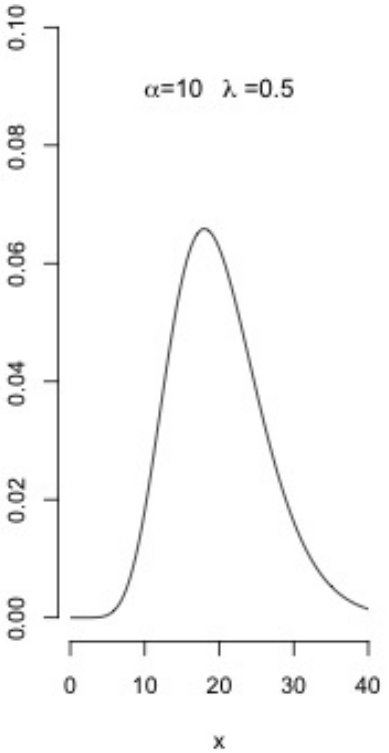
Proof: mgf of X_i is $\left(\frac{\lambda}{\lambda - t}\right)$, and so mgf of $\sum_{i=1}^n X_i$ is $\left(\frac{\lambda}{\lambda - t}\right)^n$, which we recognise as the mgf of a Gamma(n, λ).$\square$

Some Gamma distributions:

```
a<-c(1, 3, 10);   b<-c(1, 3, 0.5)
for(i in 1:3){
  y= dgamma(x, a[i],b[i])
  plot(x,y,.......) }
```







Some important continuous distributions: Chi-squared

If $Z_1, \ldots, Z_k$ are iid $N(0,1)$ r.v.'s, then $X = \sum_{i=1}^k Z_i^2$ has a **chi-squared** distribution on k degrees of freedom, $X \sim \chi_k^2$.

Since $\mathbb{E}(Z_i^2) = 1$ and $\mathbb{E}(Z_i^4) = 3$, we find that $\mathbb{E}(X) = k$ and $\text{var}(X) = 2k$.

Further, the moment generating function of Z_i^2 is

$$M_{Z_i^2}(t) = \mathbb{E}\left(e^{Z_i^2 t}\right) = \int_{-\infty}^{\infty} e^{z^2 t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = (1-2t)^{-1/2} \text{ for } t < 1/2$$

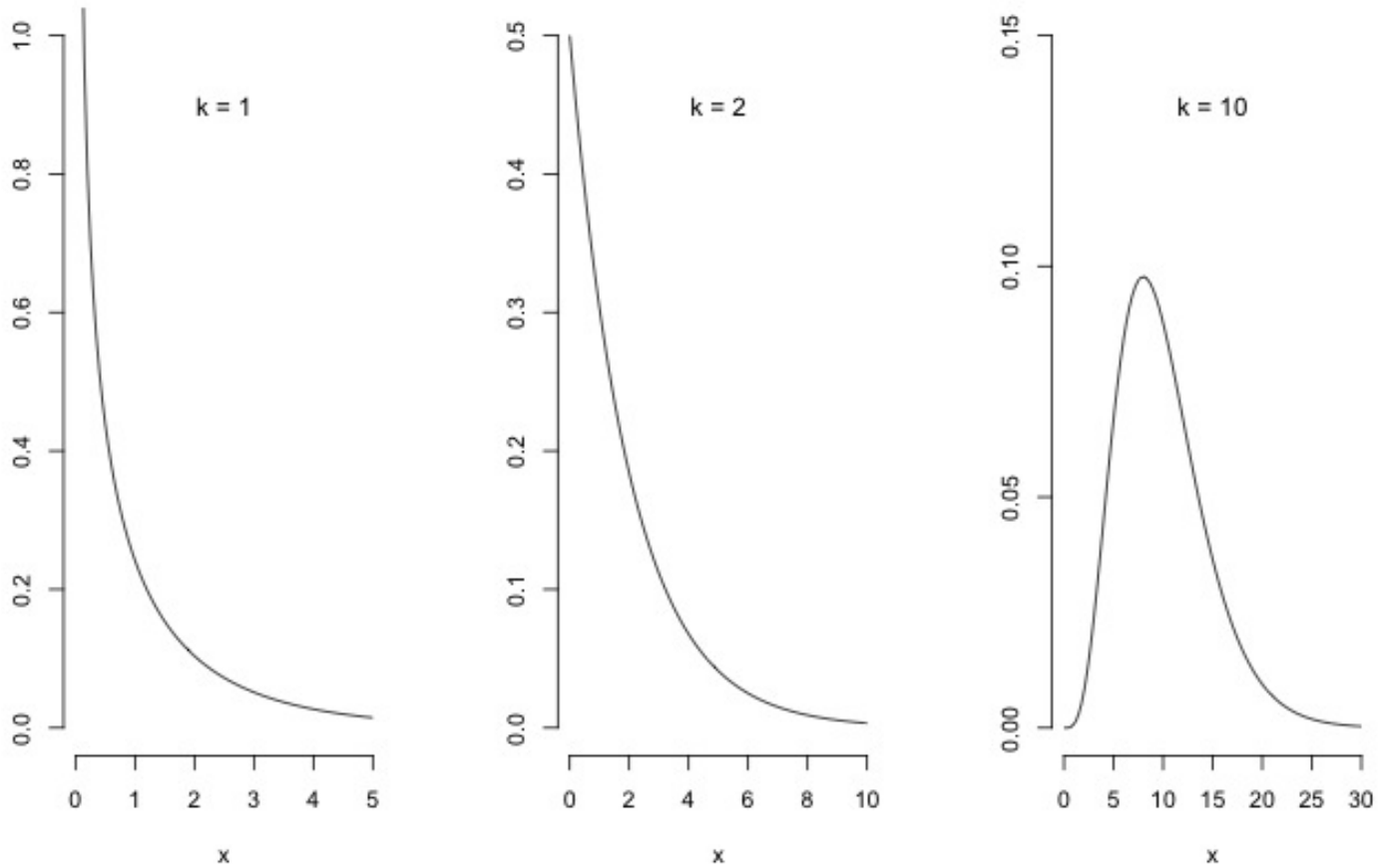
(check), so that the mgf of $X = \sum_{i=1}^k Z_i^2$ is $M_X(t) = (M_{Z^2}(t))^k = (1-2t)^{-k/2}$ for $t < 1/2$.

We recognise this as the mgf of a $\text{Gamma}(k/2, 1/2)$, so that X has pdf

$$f_X(x) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{k/2-1} e^{-x/2}, \qquad x > 0.$$

Some chi-squared distributions: $k = 1,2,10$:

```
k<-c(1,2,10)
for(i in 1:3){
y=dchisq(x, k[i])
  plot(x,y,.......) }
```



Note:

1. We have seen that if $X \sim \chi^2_k$ then $X \sim \mathrm{Gamma}(k/2, 1/2)$.
2. If $Y \sim \mathrm{Gamma}(n, \lambda)$ then $2\lambda Y \sim \chi^2_{2n}$ (prove via mgf's or density transformation formula).
3. If $X \sim \chi^2_m$, $Y \sim \chi^2_n$ and X and Y are independent, then $X + Y \sim \chi^2_{m+n}$ (prove via mgf's). This is called the additive property of χ^2.
4. We denote the upper $100\alpha\%$ point of χ^2_k by $\chi^2_k(\alpha)$, so that, if $X \sim \chi^2_k$ then $\mathbb{P}(X > \chi^2_k(\alpha)) = \alpha$. These are tabulated. The above connections between gamma and χ^2 means that sometimes we can use χ^2-tables to find percentage points for gamma distributions.

Some important continuous distributions: Beta

X has a **Beta(**α, β) distribution ($\alpha > 0$, $\beta > 0$) if it has pdf

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < x < 1,$$

where $B(\alpha, \beta)$ is the beta function defined by

$$B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta).$$

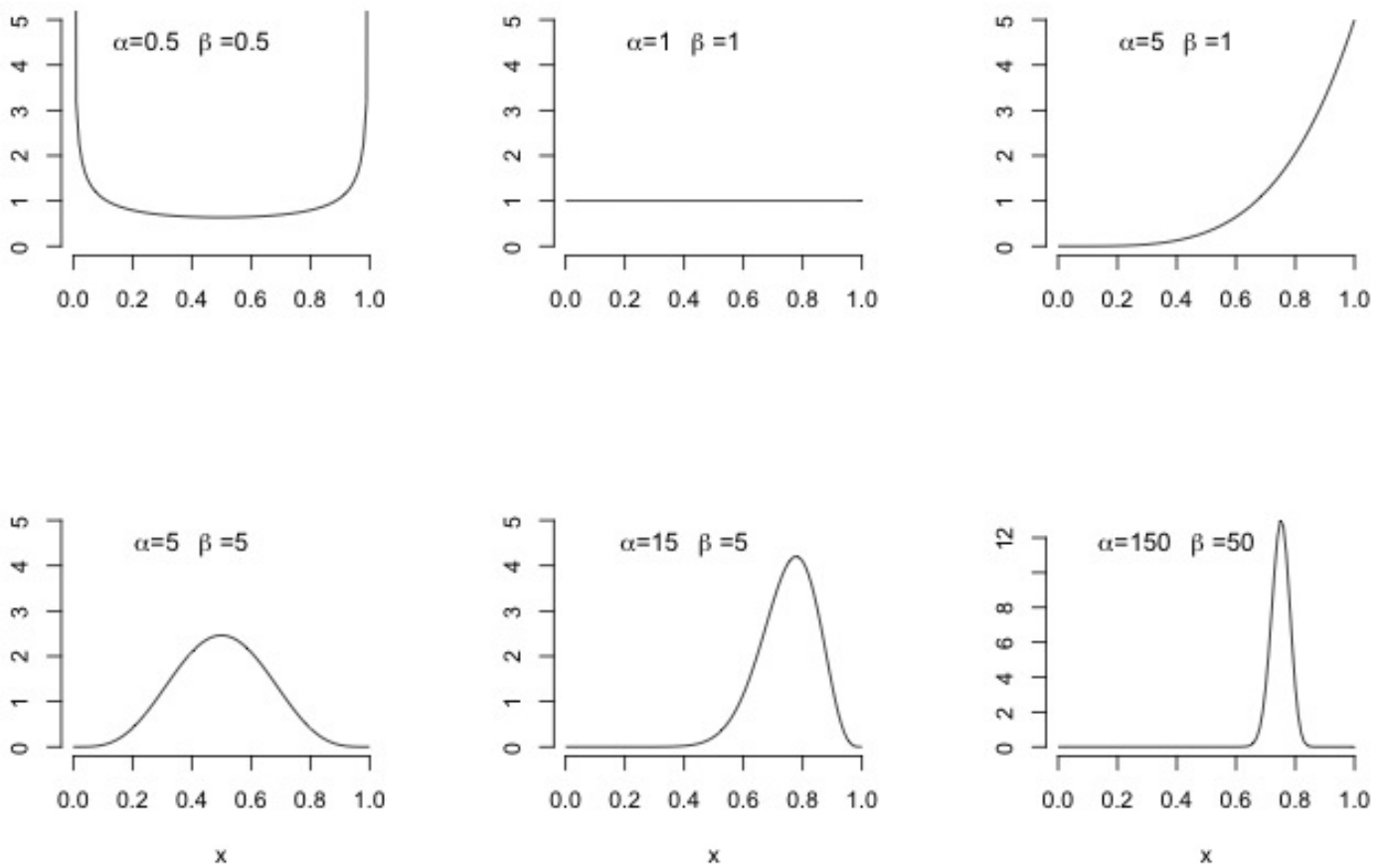
Then $\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta}$ and $\text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

The mode is $(\alpha - 1)/(\alpha + \beta - 2)$.

Note that Beta(1,1)$\sim U[0, 1]$.

Some beta distributions :

```
k<-c(1,2,10)
for(i in 1:3){
y=dbeta(x, a[i],b[i])
  plot(x,y,.......) }
```



Lecture 2. Estimation, bias, and mean squared error

Estimators

- Suppose that $X_1, \ldots, X_n$ are iid, each with pdf/pmf $f_X(x \mid \theta)$, θ unknown.
- We aim to estimate θ by a **statistic**, ie by a function T of the data.
- If $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_n)$ then our estimate is $\hat{\theta} = T(\mathbf{x})$ (does not involve θ).
- Then $T(\mathbf{X})$ is our **estimator** of θ, and is a rv since it inherits random fluctuations from those of $\mathbf{X}$.
- The distribution of $T = T(\mathbf{X})$ is called its **sampling distribution**.

Example

Let $X_1, \ldots, X_n$ be iid $N(\mu, 1)$.

A possible estimator for μ is $T(\mathbf{X}) = \frac{1}{n} \sum X_i$.

For any particular observed sample $\mathbf{x}$, our estimate is $T(\mathbf{x}) = \frac{1}{n} \sum x_i$.

We have $T(\mathbf{X}) \sim N(\mu, 1/n)$. $\square$

Bias

If $\hat{\theta} = T(\mathbf{X})$ is an estimator of θ, then the *bias* of $\hat{\theta}$ is the difference between its expectation and the 'true' value: i.e.

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

An estimator $T(\mathbf{X})$ is **unbiased** for θ if $\mathbb{E}_\theta T(\mathbf{X}) = \theta$ for all θ, otherwise it is **biased**.

In the above example, $\mathbb{E}_\mu(T) = \mu$ so T is unbiased for μ.

[Notation note: when a parameter subscript is used with an expectation or variance, it refers to the parameter that is being conditioned on. i.e. the expectation or variance will be a function of the subscript]

Mean squared error

Recall that an estimator T is a function of the data, and hence is a random quantity. Roughly, we prefer estimators whose sampling distributions "cluster more closely" around the true value of θ, whatever that value might be.

Definition 2.1

The **mean squared error** (mse) of an estimator $\hat{\theta}$ is $\mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right]$.

For an unbiased estimator, the mse is just the variance.

In general

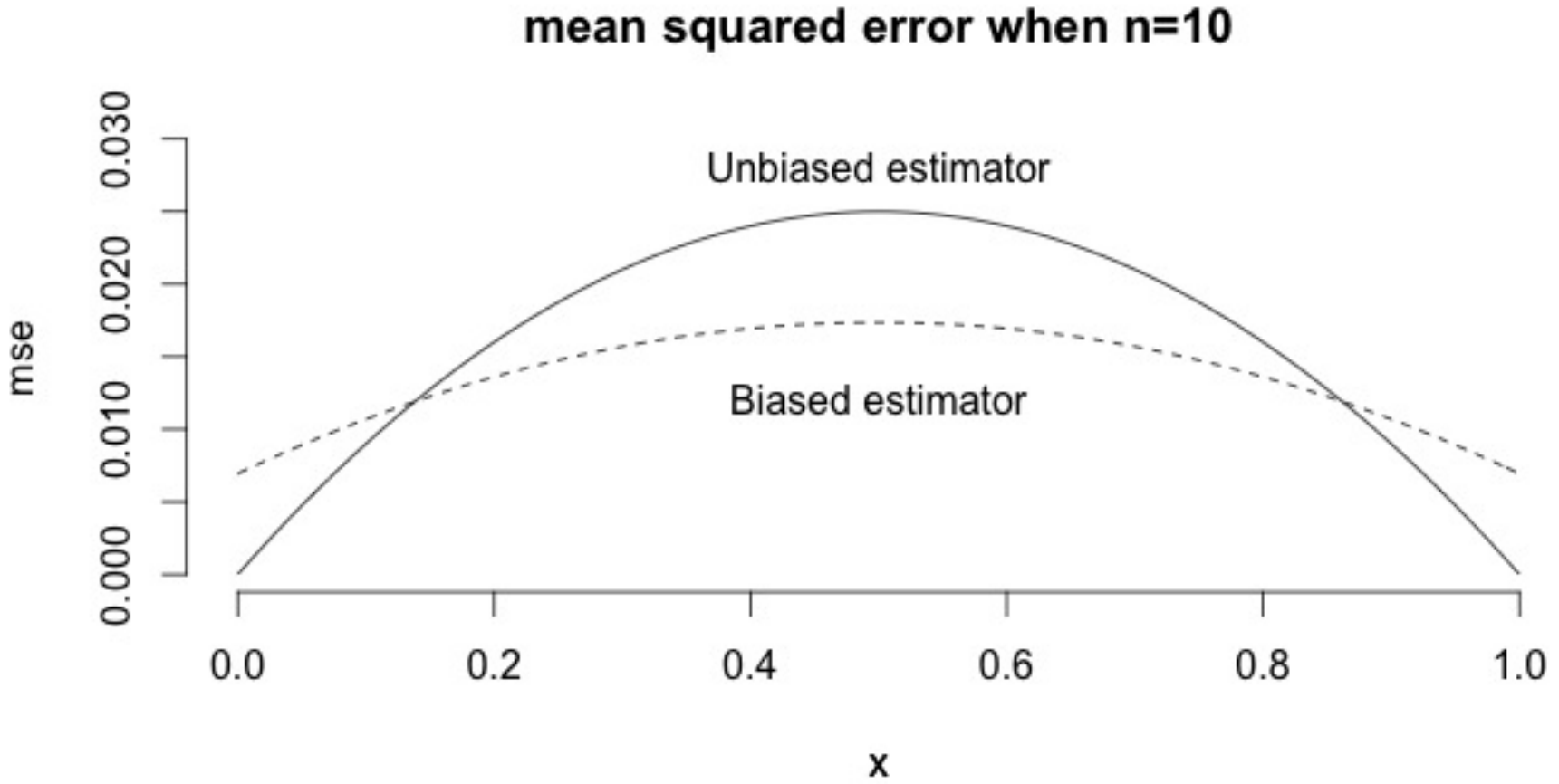
$$\begin{aligned}\mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right] &= \mathbb{E}_\theta\left[(\hat{\theta}-\mathbb{E}_\theta\hat{\theta}+\mathbb{E}_\theta\hat{\theta}-\theta)^2\right] \\ &= \mathbb{E}_\theta\left[(\hat{\theta}-\mathbb{E}_\theta\hat{\theta})^2\right]+\left[\mathbb{E}_\theta(\hat{\theta})-\theta\right]^2+2\left[\mathbb{E}_\theta(\hat{\theta})-\theta\right]\mathbb{E}_\theta\left[\hat{\theta}-\mathbb{E}_\theta\hat{\theta}\right] \\ &= \mathrm{var}_\theta(\hat{\theta})+\mathrm{bias}^2(\hat{\theta}),\end{aligned}$$

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

[NB: sometimes it can be preferable to have a biased estimator with a low variance - this is sometimes known as the 'bias-variance tradeoff'.]

Example: Alternative estimators for Binomial mean

- Suppose $X \sim \text{Binomial}(n, \theta)$, and we want to estimate θ.
- The standard estimator is $T_U = X/n$, which is Unbiassed since $\mathbb{E}_\theta(T_U) = n\theta/n = \theta$.
- T_U has variance $var_\theta(T_U) = var_\theta(X)/n^2 = \theta(1-\theta)/n$.
- Consider an alternative estimator $T_B = \frac{X+1}{n+2} = w\frac{X}{n} + (1-w)\frac{1}{2}$, where $w = n/(n+2)$.
- (Note: T_B is a weighted average of X/n and $\frac{1}{2}$.)
- e.g. if X is 8 successes out of 10 trials, we would estimate the underlying success probability as $T(8) = 9/12 = 0.75$, rather than 0.8.
- Then $\mathbb{E}_\theta(T_B) - \theta = \frac{n\theta+1}{n+2} - \theta = (1-w)\left(\frac{1}{2} - \theta\right)$, and so it is biased.
- $var_\theta(T_B) = \frac{var_\theta(X)}{(n+2)^2} = w^2\theta(1-\theta)/n$.
- Now $mse(T_U) = var_\theta(T_U) + \text{bias}^2(T_U) = \theta(1-\theta)/n$.
- $mse(T_B) = var_\theta(T_B) + \text{bias}^2(T_B) = w^2\theta(1-\theta)/n + (1-w)^2\left(\frac{1}{2} - \theta\right)^2$



So the biased estimator has smaller MSE in much of the range of θ

T_B may be preferable if we do not think θ is near 0 or 1.

So our *prior judgement* about θ might affect our choice of estimator.

Will see more of this when we come to Bayesian methods,.

Why unbiasedness is not necessarily so great

Suppose $X \sim \text{Poisson}(\lambda)$, and for some reason (which escapes me for the moment), you want to estimate $\theta = [\mathbb{P}(X = 0)]^2 = e^{-2\lambda}$.

Then any unbiassed estimator $T(X)$ must satisfy $\mathbb{E}_\theta(T(X)) = \theta$, or equivalently

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

The only function T that can satisfy this equation is $T(X) = (-1)^X$ [coefficients of polynomial must match].

Thus the only unbiassed estimator estimates $e^{-2\lambda}$ to be 1 if X is even, -1 if X is odd.

This is not sensible.

Lecture 3. Sufficiency

Sufficient statistics

The concept of sufficiency addresses the question

"Is there a statistic $T(\mathbf{X})$ that in some sense contains all the information about θ that is in the sample?"

Example 3.1

$X_1, \ldots, X_n$ iid Bernoulli(θ), so that $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = \theta$ for some $0 < \theta < 1$.

So $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$.

This depends on the data only through $T(\mathbf{x}) = \sum x_i$, the total number of ones.

Note that $T(\mathbf{X}) \sim \text{Bin}(n, \theta)$.

If $T(\mathbf{x}) = t$, then

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

ie the conditional distribution of $\mathbf{X}$ given $T = t$ does not depend on θ.

Thus if we know T, then additional knowledge of $\mathbf{x}$ (knowing the exact sequence of 0's and 1's) does not give extra information about θ. □

Definition 3.1

A statistic T is **sufficient** for θ if the conditional distribution of $\mathbf{X}$ given T does not depend on θ.

Note that T and/or θ may be vectors. In practice, the following theorem is used to find sufficient statistics.

Theorem 3.2

(The Factorisation criterion) T is sufficient for θ iff $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$ for suitable functions g and h.

Proof (Discrete case only)

Suppose $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$.

If $T(\mathbf{x}) = t$ then

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

which does not depend on θ, so T is sufficient.

Now suppose that T is sufficient so that the conditional distribution of $\mathbf{X} \mid T = t$ does not depend on θ. Then

$$\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t(\mathbf{x})) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x} \mid T = t)\mathbb{P}_\theta(T = t).$$

The first factor does not depend on θ by assumption; call it $h(\mathbf{x})$. Let the second factor be $g(t, \theta)$, and so we have the required factorisation. □

Example 3.1 continued

For Bernoulli trials, $f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$.

Take $g(t,\theta) = \theta^t(1-\theta)^{n-t}$ and $h(\mathbf{x}) = 1$ to see that $T(\mathbf{X}) = \sum X_i$ is sufficient for θ. □

Example 3.2

Let $X_1, \ldots, X_n$ be iid $U[0,\theta]$.

Write $1_{[A]}$ for the indicator function of A.

We have

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} 1_{[0 \le x_i \le \theta]} = \frac{1}{\theta^n} 1_{[\max_i x_i \le \theta]} 1_{[\min_i x_i \ge 0]}.$$

Then $T(\mathbf{X}) = \max_i X_i$ is sufficient for θ. □

Minimal sufficient statistics

Sufficient statistics are not unique. If T is sufficient for θ, then so is any (1-1) function of T.

$\mathbf{X}$ itself is always sufficient for θ; take $\mathbf{T}(\mathbf{X}) = \mathbf{X}$, $g(\mathbf{t}, \theta) = f_{\mathbf{X}}(\mathbf{t} \mid \theta)$ and $h(\mathbf{x}) = 1$. But this is not much use.

The sample space $\mathcal{X}^n$ is partitioned by T into sets $\{\mathbf{x} \in \mathcal{X}^n : T(\mathbf{x}) = t\}$.

If T is sufficient, then this data reduction does not lose any information on θ.

We seek a sufficient statistic that achieves the maximum-possible reduction.

Definition 3.3

A sufficient statistic $T(\mathbf{X})$ is *minimal sufficient* if it is a function of every other sufficient statistic:
i.e. if $T'(\mathbf{X})$ is also sufficient, then $T'(\mathbf{X}) = T'(\mathbf{Y}) \rightarrow T(\mathbf{X}) = T(\mathbf{Y})$
i.e. the partition for T is coarser than that for T'.

Minimal sufficient statistics can be found using the following theorem.

Theorem 3.4

Suppose $T = T(\mathbf{X})$ is a statistic such that $f_{\mathbf{X}}(\mathbf{x};\theta)/f_{\mathbf{X}}(\mathbf{y};\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then T is minimal sufficient for θ.

Sketch of proof : Non-examinable

First, we aim to use the Factorisation Criterion to show sufficiency. Define an equivalence relation $\sim$ on $\mathcal{X}^n$ by setting $\mathbf{x} \sim \mathbf{y}$ when $T(\mathbf{x}) = T(\mathbf{y})$. (Check that this is indeed an equivalence relation.) Let $\mathcal{U} = \{T(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^n\}$, and for each u in $\mathcal{U}$, choose a representative $\mathbf{x}_u$ from the equivalence class $\{\mathbf{x} : T(\mathbf{x}) = u\}$. Let $\mathbf{x}$ be in $\mathcal{X}^n$ and suppose that $T(\mathbf{x}) = t$. Then $\mathbf{x}$ is in the equivalence class $\{\mathbf{x}' : T(\mathbf{x}') = t\}$, which has representative $\mathbf{x}_t$, and this representative may also be written $\mathbf{x}_{T(\mathbf{x})}$. We have $\mathbf{x} \sim \mathbf{x}_t$, so that $T(\mathbf{x}) = T(\mathbf{x}_t)$, ie $T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$. Hence, by hypothesis, the ratio $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_{T(\mathbf{x})};\theta)}$ does not depend on θ, so let this be $h(\mathbf{x})$. Let $g(t,\theta) = f_{\mathbf{X}}(\mathbf{x}_t,\theta)$. Then

$$f_{\mathbf{X}}(\mathbf{x};\theta) = f_{\mathbf{X}}(\mathbf{x}_{T(\mathbf{x})};\theta)\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{x}_{T(\mathbf{x})};\theta)} = g(T(\mathbf{x}),\theta)h(\mathbf{x}),$$

and so $T = T(\mathbf{X})$ is sufficient for θ by the Factorisation Criterion.

Next we aim to show that $T(\mathbf{X})$ is a function of every other sufficient statistic.

Suppose that $S(\mathbf{X})$ is also sufficient for θ, so that, by the Factorisation Criterion, there exist functions g_S and h_S (we call them g_S and h_S to show that they belong to S and to distinguish them from g and h above) such that

$$f_{\mathbf{X}}(\mathbf{x};\theta) = g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x}).$$

Suppose that $S(\mathbf{x}) = S(\mathbf{y})$. Then

$$\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)} = \frac{g_S(S(\mathbf{x}),\theta)h_S(\mathbf{x})}{g_S(S(\mathbf{y}),\theta)h_S(\mathbf{y})} = \frac{h_S(\mathbf{x})}{h_S(\mathbf{y})},$$

because $S(\mathbf{x}) = S(\mathbf{y})$. This means that the ratio $\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{\mathbf{X}}(\mathbf{y};\theta)}$ does not depend on θ, and this implies that $T(\mathbf{x}) = T(\mathbf{y})$ by hypothesis. So we have shown that $S(\mathbf{x}) = S(\mathbf{y})$ implies that $T(\mathbf{x}) = T(\mathbf{y})$, i.e T is a function of S. Hence T is minimal sufficient. $\square$

Example 3.3

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$.

Then

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

This is constant as a function of (μ, σ^2) iff $\sum_i x_i^2 = \sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$.
So $T(\mathbf{X}) = \left(\sum_i X_i^2, \sum_i X_i\right)$ is minimal sufficient for (μ, σ^2). □

1-1 functions of minimal sufficient statistics are also minimal sufficient.
So $\mathbf{T}'(\mathbf{X}) == (\bar{X}, \sum (X_i - \bar{X})^2)$ is also sufficient for (μ, σ^2), where $\bar{X} = \sum_i X_i / n$.
We write S_{XX} for $\sum (X_i - \bar{X})^2$.

Notes

- Example 3.3 has a vector T sufficient for a vector θ. Dimensions do not have to the same: e.g. for $N(\mu, \mu^2)$, $T(\mathbf{X}) = \left(\sum_i X_i^2, \sum_i X_i\right)$ is minimal sufficient for μ [check]
- If the range of X depends on θ, then "$f_{\mathbf{X}}(\mathbf{x};\theta)/f_{\mathbf{X}}(\mathbf{y};\theta)$ is constant in θ" means "$f_{\mathbf{X}}(\mathbf{x};\theta) = c(\mathbf{x},\mathbf{y})\, f_{\mathbf{X}}(\mathbf{y};\theta)$"

The Rao–Blackwell Theorem

The Rao–Blackwell theorem gives a way to improve estimators in the mse sense.

Theorem 3.5

(The Rao–Blackwell theorem)
Let T be a sufficient statistic for θ and let $\tilde{\theta}$ be an estimator for θ with $\mathbb{E}(\tilde{\theta}^2) < \infty$ for all θ. Let $\hat{\theta} = \mathbb{E}\big[\tilde{\theta} \,|\, T\big]$. Then for all θ,

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

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

Proof By the conditional expectation formula we have $\mathbb{E}\hat{\theta} = \mathbb{E}\big[\mathbb{E}(\tilde{\theta} \,|\, T)\big] = \mathbb{E}\tilde{\theta}$, so $\hat{\theta}$ and $\tilde{\theta}$ have the same bias. By the conditional variance formula,

$$\mathrm{var}(\tilde{\theta}) = \mathbb{E}\big[\mathrm{var}(\tilde{\theta} \,|\, T)\big] + \mathrm{var}\big[\mathbb{E}(\tilde{\theta} \,|\, T)\big] = \mathbb{E}\big[\mathrm{var}(\tilde{\theta} \,|\, T)\big] + \mathrm{var}(\hat{\theta}).$$

Hence $\mathrm{var}(\tilde{\theta}) \geq \mathrm{var}(\hat{\theta})$, and so $\mathrm{mse}(\tilde{\theta}) \geq \mathrm{mse}(\hat{\theta})$, with equality only if $\mathrm{var}(\tilde{\theta} \,|\, T) = 0$. □

Notes

(i) Since T is sufficient for θ, the conditional distribution of $\mathbf{X}$ given $T = t$ does not depend on θ. Hence $\hat{\theta} = \mathbb{E}\big[\tilde{\theta}(\mathbf{X}) \,|\, T\big]$ does not depend on θ, and so is a bona fide estimator.

(ii) The theorem says that given any estimator, we can find one that is a function of a sufficient statistic that is at least as good in terms of mean squared error of estimation.

(iii) If $\tilde{\theta}$ is unbiased, then so is $\hat{\theta}$.

(iv) If $\tilde{\theta}$ is already a function of T, then $\hat{\theta} = \tilde{\theta}$.

Example 3.4

Suppose $X_1, \ldots, X_n$ are iid Poisson(λ), and let $\theta = e^{-\lambda}$ ($= \mathbb{P}(X_1 = 0)$).

Then $p_{\mathbf{X}}(\mathbf{x}|\lambda) = \left(e^{-n\lambda}\lambda^{\sum x_i}\right) / \prod x_i!$, so that $p_{\mathbf{X}}(\mathbf{x}|\theta) = \left(\theta^n(-\log\theta)^{\sum x_i}\right) / \prod x_i!$.

We see that $T = \sum X_i$ is sufficient for θ, and $\sum X_i \sim$ Poisson($n\lambda$).

An easy estimator of θ is $\tilde{\theta} = 1_{[X_1=0]}$ (unbiased) [i.e. if do not observe any events in first observation period, assume the event is impossible!]

Then

$$\begin{aligned}\mathbb{E}\left[\tilde{\theta}\,|\,T=t\right] &= \mathbb{P}\left(X_1=0 \mid \sum_1^n X_i = t\right) \\ &= \frac{\mathbb{P}(X_1=0)\mathbb{P}\left(\sum_2^n X_i = t\right)}{\mathbb{P}\left(\sum_1^n X_i = t\right)} \left(\frac{n-1}{n}\right)^t \text{ (check).}\end{aligned}$$

So $\hat{\theta} = (1 - \frac{1}{n})^{\sum X_i}$. $\square$

[Common sense check: $\hat{\theta} = (1 - \frac{1}{n})^{n\overline{X}} \approx e^{-\overline{X}} = e^{-\hat{\lambda}}$]

Example 3.5

Let $X_1, \ldots, X_n$ be iid $U[0, \theta]$, and suppose that we want to estimate θ. From Example 3.2, $T = \max X_i$ is sufficient for θ. Let $\tilde{\theta} = 2X_1$, an unbiased estimator for θ [check].

Then

$$\begin{aligned}
\mathbb{E}\big[\tilde{\theta} \,|\, T = t\big] &= 2\mathbb{E}\big[X_1 \mid \max X_i = t\big] \\
&= 2\Big(\mathbb{E}\big[X_1 \mid \max X_i = t, X_1 = \max X_i\big]\mathbb{P}(X_1 = \max X_i) \\
&\quad + \mathbb{E}\big[X_1 \mid \max X_i = t, X_1 \neq \max X_i\big]\mathbb{P}(X_1 \neq \max X_i)\Big) \\
&= 2\Big(t \times \frac{1}{n} + \frac{t}{2}\frac{n-1}{n}\Big) = \frac{n+1}{n}t,
\end{aligned}$$

so that $\hat{\theta} = \frac{n+1}{n} \max X_i$. $\square$

In Lecture 4 we show directly that this is unbiased.

N.B. Why is $\mathbb{E}\big[X_1 \mid \max X_i = t, X_1 \neq \max X_i\big] = t/2$?

Because
$f_{X_1}(x_1 \mid X_1 < t) = \frac{f_{X_1}(x_1, X_1 < t)}{\mathbb{P}(X_1 < t)} = \frac{f_{X_1}(x_1)1_{[0 \leq X_1 < t]}}{t/\theta} = \frac{1/\theta \times 1_{[0 \leq X_1 < t]}}{t/\theta} = \frac{1}{t}1_{[0 \leq X_1 < t]}$, and so $X_1 \mid X_1 < t \sim U[0, t]$.

Lecture 4. Maximum Likelihood Estimation

Likelihood

Maximum likelihood estimation is one of the most important and widely used methods for finding estimators. Let $X_1, \ldots, X_n$ be rv's with joint pdf/pmf $f_{\mathbf{X}}(\mathbf{x} \mid \theta)$. We observe $\mathbf{X} = \mathbf{x}$.

Definition 4.1

The **likelihood** of θ is $\text{like}(\theta) = f_{\mathbf{X}}(\mathbf{x} \mid \theta)$, regarded as a function of θ. The **maximum likelihood estimator** (mle) of θ is the value of θ that maximises $\text{like}(\theta)$.

It is often easier to maximise the **log-likelihood**.

If $X_1, \ldots, X_n$ are iid, each with pdf/pmf $f_X(x \mid \theta)$, then

$$\begin{aligned} \text{like}(\theta) &= \prod_{i=1}^{n} f_X(x_i \mid \theta) \\ \text{loglike}(\theta) &= \sum_{i=1}^{n} \log f_X(x_i \mid \theta). \end{aligned}$$

Example 4.1

Let $X_1, \ldots, X_n$ be iid Bernoulli(p).

Then $l(p) = \text{loglike}(p) = (\sum x_i) \log p + (n - \sum x_i) \log(1 - p)$.

Thus

$$dl/dp = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{(1 - p)}.$$

This is zero when $p = \sum x_i/n$, and the mle of p is $\hat{p} = \sum x_i/n$.

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

Example 4.2

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. Then

$$l(\mu, \sigma^2) = \text{loglike}(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum(x_i - \mu)^2.$$

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

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum(x_i - \mu), \quad \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i - \mu)^2,$$

so the solution of the simultaneous equations is $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, S_{xx}/n)$.

(writing $\bar{x}$ for $\frac{1}{n}\sum x_i$ and S_{xx} for $\sum(x_i - \bar{x})^2$)

Hence the maximum likelihood estimators are $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, S_{XX}/n)$.

We know $\hat{\mu} \sim N(\mu, \sigma^2/n)$ so $\hat{\mu}$ is unbiased.

We shall see later that $\frac{S_{XX}}{\sigma^2} = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$, and so $\mathbb{E}(\hat{\sigma}^2) = \frac{(n-1)\sigma^2}{n}$, ie $\hat{\sigma}^2$ is biased.

However $\mathbb{E}(\hat{\sigma}^2) \to \sigma^2$ as $n \to \infty$, so $\hat{\sigma}^2$ is asymptotically unbiased.

[So sample variance estimator denominator: $n-1$ is *unbiased*, n is *mle*.]

Example 4.3

Let $X_1, \ldots, X_n$ be iid $U[0, \theta]$. Then

$$\text{like}(\theta) = \frac{1}{\theta^n} 1_{[\max x_i \leq \theta]} 1_{[\min x_i \geq 0]}.$$

For $\theta \geq \max x_i$, $\text{like}(\theta) = \frac{1}{\theta^n} > 0$ and is decreasing as θ increases, while for $\theta < \max x_i$, $\text{like}(\theta) = 0$.

Hence the value $\hat{\theta} = \max x_i$ maximises the likelihood.

Is $\hat{\theta}$ unbiased? First we need to find the distribution of $\hat{\theta}$. For $0 \leq t \leq \theta$, the distribution function of $\hat{\theta}$ is

$$F_{\hat{\theta}}(t) = \mathbb{P}(\hat{\theta} \leq t) = \mathbb{P}(X_i \leq t, \text{ all } i) = (\mathbb{P}(X_i \leq t))^n = \left(\frac{t}{\theta}\right)^n,$$

where we have used independence at the second step.

Differentiating with respect to t, we find the pdf $f_{\hat{\theta}}(t) = \frac{nt^{n-1}}{\theta^n}$, $0 \leq t \leq \theta$. Hence

$$\mathbb{E}(\hat{\theta}) = \int_0^\theta t \frac{nt^{n-1}}{\theta^n} dt = \frac{n\theta}{n+1},$$

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

Properties of mle's

(i) If T is sufficient for θ, then the likelihood is $g(T(\mathbf{x}), \theta)h(\mathbf{x})$, which depends on θ only through $T(\mathbf{x})$.

To maximise this as a function of θ, we only need to maximise g, and so the mle $\hat{\theta}$ is a *function of the sufficient statistic*.

(ii) If $\phi = h(\theta)$ where h is injective $(1-1)$, then the mle of ϕ is $\hat{\phi} = h(\hat{\theta})$. This is called the invariance property of mle's. IMPORTANT.

(iii) It can be shown that, under regularity conditions, that $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically multivariate normal with mean 0 and 'smallest attainable variance' (see Part II Principles of Statistics).

(iv) Often there is no closed form for the mle, and then we need to find $\hat{\theta}$ numerically.

Example 4.4

Smarties come in k equally frequent colours, but suppose we do not know k.

[Assume there is a vast bucket of Smarties, and so the proportion of each stays constant as you sample. Alternatively, assume you sample with replacement, although this is rather unhygienic]

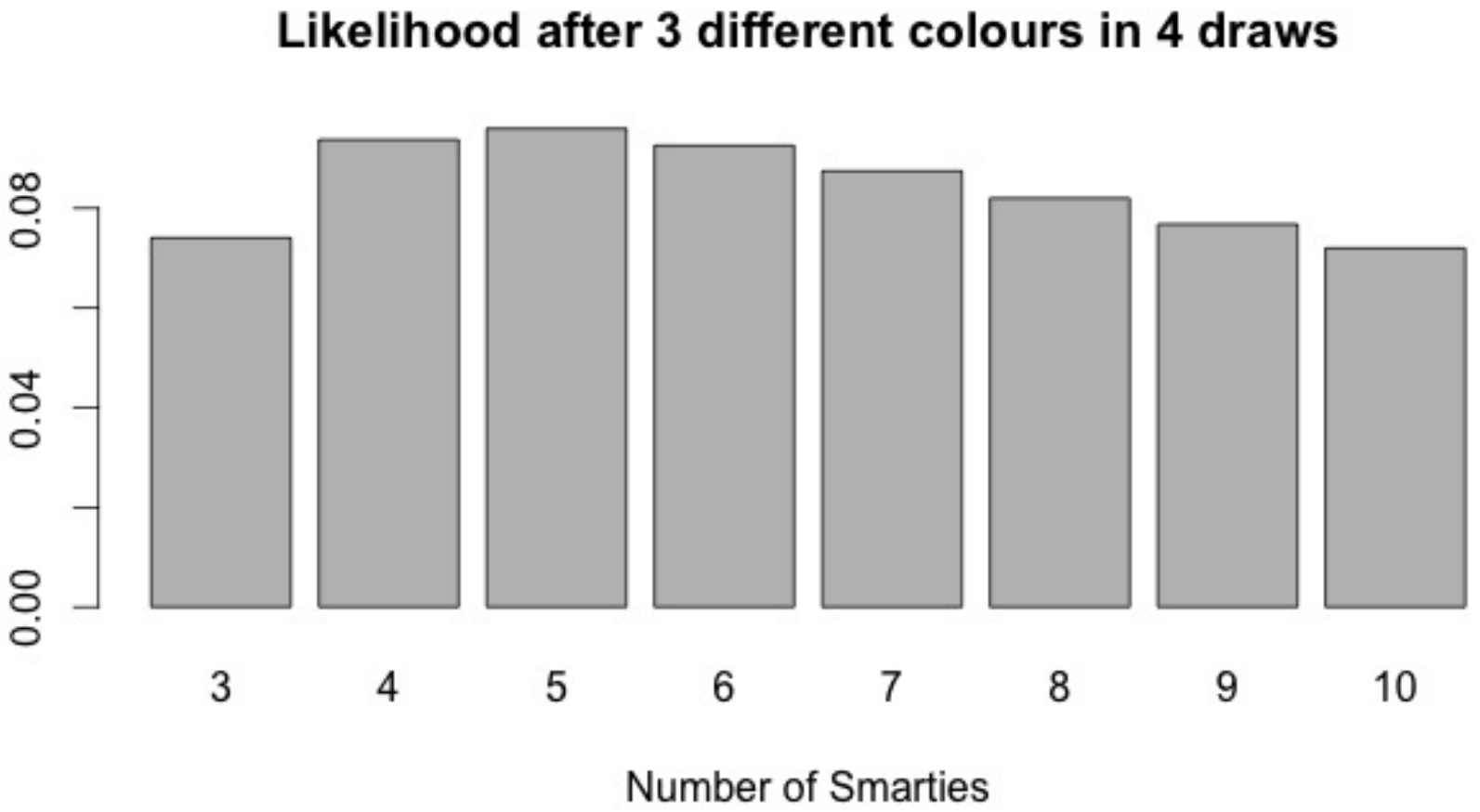
Our first four Smarties are Red, Purple, Red, Yellow.

The likelihood for k is (considered sequentially)

$$\begin{aligned}
\text{like}(k) &= \mathbb{P}_k(\text{1st is a new colour})\,\mathbb{P}_k(\text{2nd is a new colour}) \\
&\quad \mathbb{P}_k(\text{3rd matches 1st})\,\mathbb{P}_k(\text{4th is a new colour}) \\
&= 1 \times \frac{k-1}{k} \times \frac{1}{k} \times \frac{k-2}{k} \\
&= \frac{(k-1)(k-2)}{k^3}
\end{aligned}$$

(Alternatively, can think of Multinomial likelihood $\propto \frac{1}{k^4}$, but with $\binom{k}{3}$ ways of choosing those 3 colours.)

Can calculate this likelihood for different values of k:
$\text{like}(3) = 2/27, \text{like}(4) = 3/32, \text{like}(5) = 12/25, \text{like}(6) = 5/54$, maximised at $\hat{k} = 5$.



Fairly flat! Not a lot of information.

Lecture 5. Confidence Intervals

We now consider interval estimation for θ.

Definition 5.1

A $100\gamma\%$ $(0 < \gamma < 1)$ **confidence interval** (CI) for θ is a random interval $\big(A(\mathbf{X}), B(\mathbf{X})\big)$ such that $\mathbb{P}\big(A(\mathbf{X}) < \theta < B(\mathbf{X})\big) = \gamma$, no matter what the true value of θ may be.

Notice that it is the endpoints of the interval that are random quantities (not θ).

We can interpret this in terms of repeat sampling: if we calculate $\big(A(\mathbf{x}), B(\mathbf{x})\big)$ for a large number of samples $\mathbf{x}$, then approximately $100\gamma\%$ of them will cover the true value of θ.

IMPORTANT: having observed some data $\mathbf{x}$ and calculated a 95% interval $\big(A(\mathbf{x}), B(\mathbf{x})\big)$ we *cannot* say there is now a 95% probability that θ lies in this interval.

Example 5.2

Suppose $X_1, \ldots, X_n$ are iid $N(\theta, 1)$. Find a 95% confidence interval for θ.

- We know $\bar{X} \sim N(\theta, \frac{1}{n}\sigma^2)$, so that $\sqrt{n}(\bar{X} - \theta) \sim N(0, 1)$, no matter what θ is.
- Let z_1, z_2 be such that $\Phi(z_2) - \Phi(z_1) = 0.95$, where Φ is the standard normal distribution function.
- We have $\mathbb{P}\big[z_1 < \sqrt{n}(\bar{X} - \theta) < z_2\big] = 0.95$, which can be rearranged to give

$$\mathbb{P}\big[\bar{X} - \frac{z_2}{\sqrt{n}} < \theta < \bar{X} - \frac{z_1}{\sqrt{n}}\big] = 0.95.$$

  so that

$$(\bar{X} - \frac{z_2}{\sqrt{n}}, \bar{X} - \frac{z_1}{\sqrt{n}})$$

  is a 95% confidence interval for θ.
- There are many possible choices for z_1 and z_2. Since the $N(0, 1)$ density is symmetric, the shortest such interval is obtained by $z_2 = z_{0.025} = -z_1$ (where recall that z_α is the upper 100α% point of $N(0, 1)$).
- From tables, $z_{0.025} = 1.96$ so a 95% confidence interval is $(\bar{X} - \frac{1.96}{\sqrt{n}}, \bar{X} + \frac{1.96}{\sqrt{n}})$. $\square$

The above example illustrates a common procedure for findings CIs.

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

   In Example 5.2, $R(\mathbf{X}, \theta) = \sqrt{n}(\bar{X} - \theta)$.
2. Write down a probability statement of the form $\mathbb{P}_\theta\left(c_1 < R(\mathbf{X}, \theta) < c_2\right) = \gamma$.
3. Rearrange the inequalities inside $\mathbb{P}(...)$ to find the interval.

Notes:

- Usually c_1, c_2 are percentage points from a known standardised distribution, often equitailed so that use, say, 2.5% and 97.5% points for a 95% CI. Could use 0% and 95%, but interval would generally be wider.
- Can have confidence intervals for vector parameters
- If $\big(A(\mathbf{x}), B(\mathbf{x})\big)$ is a $100\gamma\%$ CI for θ, and $T(\theta)$ is a monotone increasing function of θ, then $\big(T(A(\mathbf{x})), T(B(\mathbf{x}))\big)$ is a $100\gamma\%$ CI for $T(\theta)$.

  If T is monotone decreasing, then $\big(T(B(\mathbf{x})), T(A(\mathbf{x}))\big)$ is a $100\gamma\%$ CI for $T(\theta)$.

Example 5.3

Suppose $X_1, \ldots, X_{50}$ are iid $N(0, \sigma^2)$. Find a 99% confidence interval for σ^2.

- Thus $X_i/\sigma \sim \mathrm{N}(0,1)$. So, from the Probability review, $\frac{1}{\sigma^2}\sum_{i=1}^n X_i^2 \sim \chi^2_{50}$.
- So $R(\mathbf{X}, \sigma^2) = \sum_{i=1}^n X_i^2/\sigma^2$ is a pivot.
- Recall that $\chi^2_n(\alpha)$ is the upper $100\alpha\%$ point of χ^2_n, i.e. $\mathbb{P}(\chi^2_n \le \chi^2_n(\alpha)) = 1-\alpha$.
- From χ^2-tables, we can find c_1, c_2 such that $F_{\chi^2_{50}}(c_2) - F_{\chi^2_{50}}(c_1) = 0.99$.
- An equi-tailed region is given by $c_1 = \chi^2_{50}(0.995) = 27.99$ and $c_2 = \chi^2_{50}(0.005) = 79.49$.
- In R,
  `qchisq(0.005,50) = 27.99075,  qchisq(0.995,50) = 79.48998`
- Then $\mathbb{P}_{\sigma^2}\big(c_1 < \frac{\sum X_i^2}{\sigma^2} < c_2\big) = 0.99$, and so $\mathbb{P}_{\sigma^2}\big(\frac{\sum X_i^2}{c_2} < \sigma^2 < \frac{\sum X_i^2}{c_1}\big) = 0.99$ which gives a confidence interval $\big(\frac{\sum X_i^2}{79.49}, \frac{\sum X_i^2}{27.99}\big)$.
- Further, a 99% confidence interval for σ is then $\big(\sqrt{\frac{\sum X_i^2}{79.49}}, \sqrt{\frac{\sum X_i^2}{27.99}}\big)$. $\square$

Example 5.4

Suppose $X_1, \ldots, X_n$ are iid Bernoulli(p). Find an approximate confidence interval for p.

- The mle of p is $\hat{p} = \sum X_i/n$.
- By the Central Limit Theorem, $\hat{p}$ is approximately $N(p, p(1-p)/n)$ for large n.
- So $\sqrt{n}(\hat{p} - p)/\sqrt{p(1-p)}$ is approximately $N(0,1)$ for large n.
- So we have

$$\mathbb{P}\left(\hat{p} - z_{(1-\gamma)/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{(1-\gamma)/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx \gamma.$$

- But p is unknown, so we approximate it by $\hat{p}$, to get an approximate $100\gamma\%$ confidence interval for p when n is large:

$$\left(\hat{p} - z_{(1-\gamma)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{(1-\gamma)/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

□

NB. There are many possible approximate confidence intervals for a Bernoulli/Binomial parameter.

Example 5.5

Suppose an opinion poll says 20% are going to vote UKIP, based on a random sample of 1,000 people. What might the true proportion be?

- We assume we have an observation of $x = 200$ from a Binomial(n, p) distribution with $n = 1,000$.
- Then $\hat{p} = x/n = 0.2$ is an unbiased estimate, also the mle.
- Now $\text{var}\left(\frac{X}{n}\right) = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.2 \times 0.8}{1000} = 0.00016$.
- So a 95% CI is $\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.20 \pm 1.96 \times 0.013 = (0.175, 0.225)$, or around 17% to 23%.
- Special case of common procedure for an unbiased estimator T: 95% CI $\approx T \pm 2\sqrt{\text{var}\, T} = T \pm 2\,\text{SE}$, where SE = 'standard error' $= \sqrt{\text{var}\, T}$
- NB: Since $p(1-p) \leq 1/4$ for all $0 \leq p \leq 1$, then a conservative 95% interval (i.e. might be a bit wide) is $\hat{p} \pm 1.96\sqrt{\frac{1}{4n}} \approx \hat{p} \pm \sqrt{\frac{1}{n}}$.
- So whatever proportion is reported, it will be 'accurate' to $\pm 1/\sqrt{n}$.
- Opinion polls almost invariably use $n = 1000$, so they are assured of $\pm 3\%$ 'accuracy'

(Slightly contrived) confidence interval problem*

Example 5.6

Suppose X_1 and X_2 are iid from Uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. What is a sensible 50% CI for θ?

- Consider the probability of getting one observation each side of θ,

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

  So $(\min(X_1, X_2), \max(X_1, X_2))$ is a 50% CI for θ.
- But suppose $\mid X_1 - X_2 \mid \geq \frac{1}{2}$, e.g. $x_1 = 0.2, x_2 = 0.9$. Then we *know* that, in this particular case, θ *must* lie in $(\min(X_1, X_2), \max(X_1, X_2))$.
- So guaranteed sampling properties does not necessarily mean a sensible conclusion in all cases.

Lecture 6. Bayesian estimation

The parameter as a random variable

- So far we have seen the *frequentist* approach to statistical inference
- i.e. inferential statements about θ are interpreted in terms of repeat sampling.
- In contrast, the Bayesian approach treats θ *as a random variable* taking values in Θ.
- The investigator's information and beliefs about the possible values for θ, before any observation of data, are summarised by a **prior distribution** $\pi(\theta)$.
- When data $\mathbf{X} = \mathbf{x}$ are observed, the extra information about θ is combined with the prior to obtain the **posterior distribution** $\pi(\theta \,|\, \mathbf{x})$ for θ given $\mathbf{X} = \mathbf{x}$.
- There has been a long-running argument between proponents of these different approaches to statistical inference
- Recently things have settled down, and Bayesian methods are seen to be appropriate in huge numbers of application where one seeks to assess a probability about a 'state of the world'.
- Examples are spam filters, text and speech recognition, machine learning, bioinformatics, health economics and (some) clinical trials.

Prior and posterior distributions

- By Bayes' theorem,

$$\pi(\theta \,|\, \mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x} \mid \theta)\pi(\theta)}{f_{\mathbf{X}}(\mathbf{x})},$$

  where $f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta)d\theta$ for continuous θ, and $f_{\mathbf{X}}(\mathbf{x}) = \sum f_{\mathbf{X}}(\mathbf{x}|\theta_i)\pi(\theta_i)$ in the discrete case.

- Thus

$$\begin{aligned} \pi(\theta \,|\, \mathbf{x}) &\propto f_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior}, \end{aligned} \tag{1}$$

  where the constant of proportionality is chosen to make the total mass of the posterior distribution equal to one.

- In practice we use (**??**) and often we can recognise the family for $\pi(\theta \mid \mathbf{x})$.
- It should be clear that the data enter through the likelihood, and so the inference is automatically based on any sufficient statistic.

Inference about a discrete parameter

Suppose I have 3 coins in my pocket,

1. *biased 3:1 in favour of tails*
2. *a fair coin,*
3. *biased 3:1 in favour of heads*

I randomly select one coin and flip it once, observing a head. What is the probability that I have chosen coin 3?

- Let $X = 1$ denote the event that I observe a head, $X = 0$ if a tail
- θ denote the probability of a head: $\theta \in (0.25, 0.5, 0.75)$
- Prior: $p(\theta = 0.25) = p(\theta = 0.5) = p(\theta = 0.75) = 0.33$
- Probability mass function: $p(x|\theta) = \theta^x(1-\theta)^{(1-x)}$

Coin	θ	**Prior** $p(\theta)$	**Likelihood** $p(x=1\|\theta)$	**Un-normalised Posterior** $p(x=1\|\theta)p(\theta)$	**Normalised Posterior** $\frac{p(x=1\|\theta)p(\theta)}{p(x)^{\dagger}}$
1	0.25	0.33	0.25	0.0825	0.167
2	0.50	0.33	0.50	0.1650	0.333
3	0.75	0.33	0.75	0.2475	0.500
	Sum	1.00	1.50	0.495	1.000

† The normalising constant can be calculated as $p(x) = \sum_i p(x|\theta_i)p(\theta_i)$

So observing a head on a single toss of the coin means that there is now a 50% probability that the chance of heads is 0.75 and only a 16.7% probability that the chance of heads in 0.25.

Bayesian inference - how did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a ſingle trial lies ſomewhere between any two degrees of probability that can be named.

In modern language, given $r \sim$ Binomial(θ, n), what is $\mathbb{P}(\theta_1 < \theta < \theta_2 | r, n)$?

Example 6.1

Suppose we are interested in the true mortality risk θ in a hospital H which is about to try a new operation. On average in the country around 10% of people die, but mortality rates in different hospitals vary from around 3% to around 20%. Hospital H has no deaths in their first 10 operations. What should we believe about θ?

- Let $X_i = 1$ if the ith patient dies in H (zero otherwise), $i = 1, \ldots, n$.
- Then $f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$.
- Suppose a priori that $\theta \sim \text{Beta}(a, b)$ for some known $a > 0$, $b > 0$, so that $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, $0 < \theta < 1$.
- Then the posterior is

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto f_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta) \\ &\propto \theta^{\sum x_i + a - 1}(1-\theta)^{n-\sum x_i + b - 1}, \ 0 < \theta < 1. \end{aligned}$$

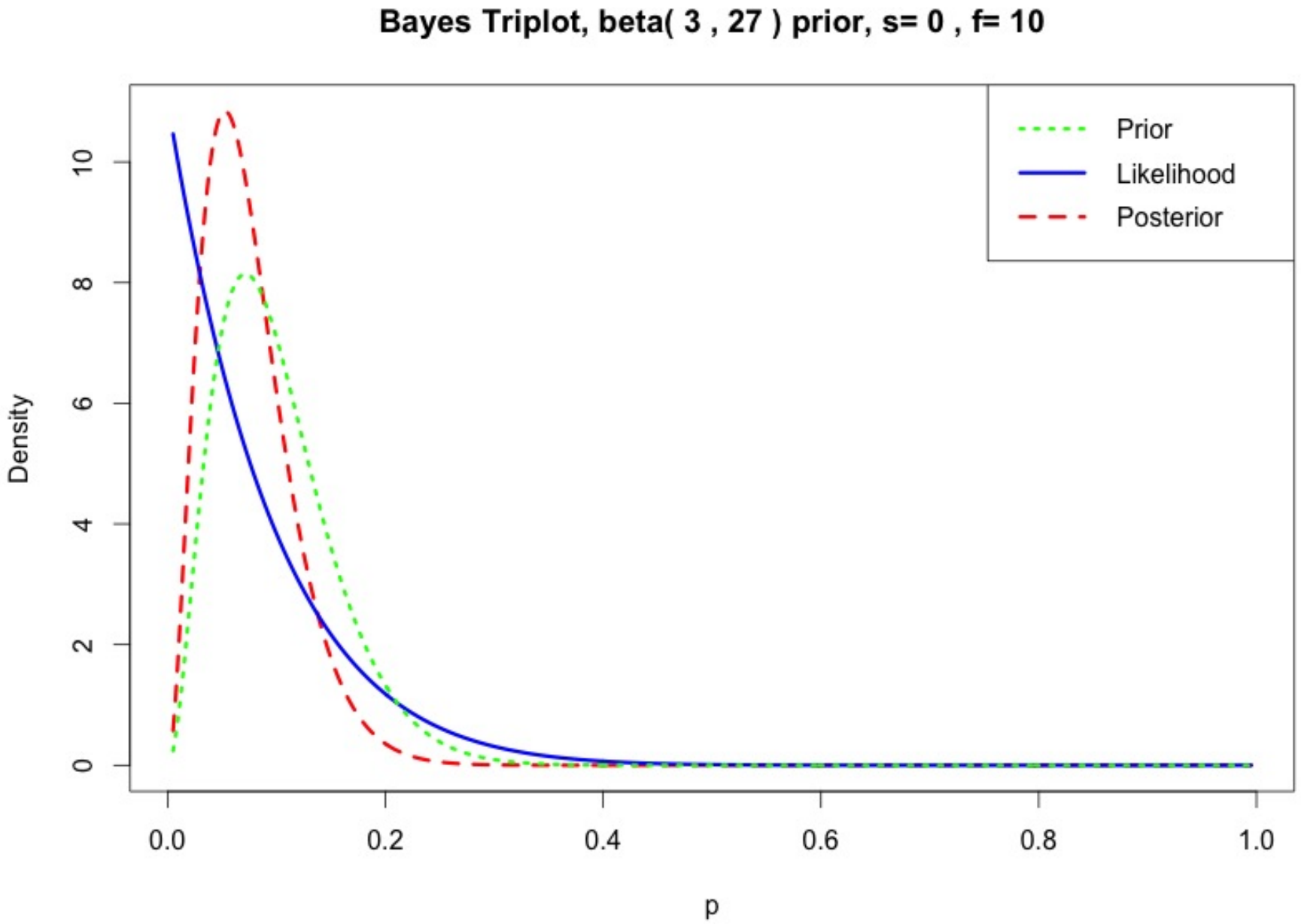
We recognise this as $\text{Beta}(\sum x_i + a, n - \sum x_i + b)$ and so

$$\pi(\theta|\mathbf{x}) = \frac{\theta^{\sum x_i + a - 1}(1-\theta)^{n-\sum x_i + b - 1}}{\text{B}(\sum x_i + a, n - \sum x_i + b)} \qquad \text{for } 0 < \theta < 1.$$

□

- In practice, we need to find a Beta prior distribution that matches our information from other hospitals.
- It turns out that a Beta(a=3,b=27) prior distribution has mean 0.1 and $\mathbb{P}(0.03 < \theta < 0.20) = 0.9$.
- The data is $\sum x_i = 0, n = 10$.
- So the posterior is Beta$(\sum x_i + a, n - \sum x_i + b)$ = Beta(3, 37)
- This has mean $3/40 = 0.075$.
- NB Even though nobody has died so far, the mle $\hat{\theta} = \sum x_i/n = 0$ (i.e. it is impossible that any will ever die) does not seem plausible.

```
install.packages("LearnBayes")
library(LearnBayes)
prior = c( a= 3, b = 27 )          # beta prior
data = c( s = 0, f = 10 ) #  s events out of f trials
triplot(prior,data)
```


Bayes Triplot, beta(3 , 27) prior, s= 0 , f= 10
Prior
Likelihood
Posterior
Density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
p

Conjugacy

- For this problem, a beta prior leads to a beta posterior. We say that the beta family is a **conjugate** family of prior distributions for Bernoulli samples.
- Suppose that $a = b = 1$ so that $\pi(\theta) = 1,\ 0 < \theta < 1$ - the uniform distribution (called the "principle of insufficient reason' by Laplace, 1774) .
- Then the posterior is Beta$(\sum x_i + 1, n - \sum x_i + 1)$, with properties.

	mean	mode	variance
prior	$1/2$	non-unique	$1/12$
posterior	$\frac{\sum x_i+1}{n+2}$	$\frac{\sum x_i}{n}$	$\frac{(\sum x_i+1)(n-\sum x_i+1)}{(n+2)^2(n+3)}$

- Notice that the mode of the posterior is the mle.
- The posterior mean estimator, $\frac{\sum x_i+1}{n+2}$ is discussed in Lecture 2, where we showed that this estimator had smaller mse than the mle for non-extreme values of θ. Known as Laplace's estimator.
- The posterior variance is bounded above by $1/(4(n+3))$, and this is smaller than the prior variance, and is smaller for larger n.
- Again, note the posterior automatically depends on the data through the sufficient statistic.

Bayesian approach to point estimation

- Let $L(\theta, a)$ be the loss incurred in estimating the value of a parameter to be a when the true value is θ.
- Common loss functions are quadratic loss $L(\theta, a) = (\theta - a)^2$, absolute error loss $L(\theta, a) = |\theta - a|$, but we can have others.
- When our estimate is a, the expected posterior loss is $h(a) = \int L(\theta, a)\pi(\theta \,|\, \mathbf{x})d\theta$.
- The **Bayes estimator $\hat{\theta}$ minimises the expected posterior loss**.
- For **quadratic loss**

$$h(a) = \int (a - \theta)^2 \pi(\theta \,|\, \mathbf{x})d\theta.$$

- $h'(a) = 0$ if

$$a \int \pi(\theta \,|\, \mathbf{x})d\theta = \int \theta\pi(\theta \,|\, \mathbf{x})d\theta.$$

- So $\hat{\theta} = \int \theta\pi(\theta \,|\, \mathbf{x})d\theta$, the **posterior mean**, minimises $h(a)$.

- For **absolute error loss**,

$$\begin{aligned} h(a) &= \int |\theta - a|\pi(\theta\,|\,\mathbf{x})d\theta = \int_{-\infty}^{a} (a-\theta)\pi(\theta\,|\,\mathbf{x})d\theta + \int_{a}^{\infty} (\theta - a)\pi(\theta\,|\,\mathbf{x})d\theta \\ &= a\int_{-\infty}^{a} \pi(\theta\,|\,\mathbf{x})d\theta - \int_{-\infty}^{a} \theta\pi(\theta\,|\,\mathbf{x})d\theta \\ &\quad + \int_{a}^{\infty} \theta\pi(\theta\,|\,\mathbf{x})d\theta - a\int_{a}^{\infty} \pi(\theta\,|\,\mathbf{x})d\theta \end{aligned}$$

  Now $h'(a) = 0$ if

$$\int_{-\infty}^{a} \pi(\theta\,|\,\mathbf{x})d\theta = \int_{a}^{\infty} \pi(\theta\,|\,\mathbf{x})d\theta.$$

- This occurs when each side is $1/2$ (since the two integrals must sum to 1) so $\hat{\theta}$ is the **posterior median**.

Example 6.2

Suppose that $X_1, \ldots, X_n$ are iid $\mathrm{N}(\mu, 1)$, and that a priori $\mu \sim \mathrm{N}(0, \tau^{-2})$ for known τ^{-2}.

- The posterior is given by

$$\begin{aligned}
\pi(\mu|\mathbf{x}) &\propto f_{\mathbf{X}}(\mathbf{x}|\mu)\pi(\mu) \\
&\propto \exp\left[-\frac{1}{2}\sum(x_i-\mu)^2\right]\exp\left[-\frac{\mu^2\tau^2}{2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(n+\tau^2\right)\left\{\mu-\frac{\sum x_i}{n+\tau^2}\right\}^2\right] \qquad \text{(check)}.
\end{aligned}$$

- So the posterior distribution of μ given $\mathbf{x}$ is a Normal distribution with mean $\sum x_i/(n+\tau^2)$ and variance $1/(n+\tau^2)$.
- The normal density is symmetric, and so the posterior mean and the posterior median have the same value $\sum x_i/(n+\tau^2)$.
- This is the optimal Bayes estimate of μ under both quadratic and absolute error loss.

Example 6.3

Suppose that $X_1, \ldots, X_n$ are iid Poisson(λ) rv's and that λ has an exponential distribution with mean 1, so that $\pi(\lambda) = e^{-\lambda}$, $\lambda > 0$.

- The posterior distribution is given by

$$\pi(\lambda|\mathbf{x}) \propto e^{-n\lambda}\lambda^{\sum x_i}e^{-\lambda} = \lambda^{\sum x_i}e^{-(n+1)\lambda}, \quad \lambda > 0,$$

  ie Gamma($\sum x_i + 1, n + 1$).
- Hence, under quadratic loss, $\hat{\lambda} = (\sum x_i + 1)/(n + 1)$, the posterior mean.
- Under absolute error loss, $\hat{\lambda}$ solves

$$\int_0^{\hat{\lambda}} \frac{(n+1)^{\sum x_i+1}\lambda^{\sum x_i}e^{-(n+1)\lambda}}{(\sum x_i)!}d\lambda = \frac{1}{2}.$$

Lecture 7. Simple Hypotheses

Introduction

Let $X_1, \ldots, X_n$ be iid, each taking values in $\mathcal{X}$, each with unknown pdf/pmf f, and suppose that we have two hypotheses, H_0 and H_1, about f.

On the basis of data $\mathbf{X} = \mathbf{x}$, we make a choice between the two hypotheses.

Examples

(a) A coin has $\mathbb{P}(\text{Heads}) = \theta$, and is thrown independently n times. We could have $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$.

(b) As in (a), with $H_0 : \theta = 1/2$ as before, but with $H_1 : \theta \neq 1/2$.

(c) Suppose $X_1, \ldots, X_n$ are iid discrete rv's. We could have H_0 :the distribution is Poisson with unknown mean, and H_1 :the distribution is not Poisson. This is a goodness-of-fit test.

(d) General parametric case: $X_1, \ldots, X_n$ are iid with density $f(x \mid \theta)$, with $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$ (we may or may not have $\Theta_0 \cup \Theta_1 = \Theta$).

(e) We could have $H_0 : f = f_0$ and $H_1 : f = f_1$ where f_0 and f_1 are densities that are completely specified but do not come from the same parametric family.

A **simple hypothesis** H specifies f completely (eg $H_0 : \theta = 1/2$ in (a)).

Otherwise H is a **composite hypothesis** (eg $H_1 : \theta \neq 1/2$ in (b)).

For testing H_0 against an alternative hypothesis H_1, a test procedure has to partition $\mathcal{X}^n$ into two disjoint and exhaustive regions C and $\bar{C}$, such that if $\mathbf{x} \in C$ then H_0 is rejected and if $\mathbf{x} \in \bar{C}$ then H_0 is not rejected.

The **critical region** (or **rejection region**) C defines the test.

When performing a test we may (i) arrive at a correct conclusion, or (ii) make one of two types of error:

(a) we may reject H_0 when H_0 is true (a **Type I error**),

(b) we may not reject H_0 when H_0 is false (a **Type II error**).

NB: When Neyman and Pearson developed the theory in the 1930s, they spoke of 'accepting' H_0. Now we generally refer to '*not rejecting* H_0'.

Testing a simple hypothesis against a simple alternative

When H_0 and H_1 are both simple, let

$$\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}(\mathbf{X} \in C \,|\, H_0 \text{ is true})$$
$$\beta = \mathbb{P}(\text{Type II error}) = \mathbb{P}(\mathbf{X} \notin C \,|\, H_1 \text{ is true}).$$

We define the **size** of the test to be α.

$1-\beta$ is also known as the **power** of the test to detect H_1.

Ideally we would like $\alpha = \beta = 0$, but typically it is not possible to find a test that makes both α and β arbitrarily small.

Definition 7.1

- The **likelihood** of a simple hypothesis $H : \theta = \theta^*$ given data $\mathbf{x}$ is $L_{\mathbf{x}}(H) = f_{\mathbf{X}}(\mathbf{x} \,|\, \theta = \theta^*)$.
- The **likelihood ratio** of two simple hypotheses H_0, H_1, given data $\mathbf{x}$, is $\Lambda_{\mathbf{x}}(H_0; H_1) = L_{\mathbf{x}}(H_1)/L_{\mathbf{x}}(H_0)$.
- A **likelihood ratio test** (LR test) is one where the critical region C is of the form $C = \{\mathbf{x} : \Lambda_{\mathbf{x}}(H_0; H_1) > k\}$ for some k. $\square$

Theorem 7.2

(The Neyman–Pearson Lemma) Suppose $H_0 : f = f_0$, $H_1 : f = f_1$, *where* f_0 *and* f_1 *are continuous densities that are nonzero on the same regions. Then among all tests of size less than or equal to* α, *the test with smallest probability of a Type II error is given by* $C = \{\mathbf{x} : f_1(\mathbf{x})/f_0(\mathbf{x}) > k\}$ *where* k *is chosen such that* $\alpha = \mathbb{P}(\textit{reject } H_0 \,|\, H_0) = \mathbb{P}(\mathbf{X} \in C \,|\, H_0) = \int_C f_0(\mathbf{x})d\mathbf{x}$.

Proof

The given C specifies a likelihood ratio test with size α.

Let $\beta = \mathbb{P}(\mathbf{X} \notin C \,|\, f_1) = \int_{\bar{C}} f_1(\mathbf{x})d\mathbf{x}$.

Let C^* be the critical region of any other test with size less than or equal to α.

Let $\alpha^* = \mathbb{P}(\mathbf{X} \in C^* \,|\, f_0)$, $\beta^* = \mathbb{P}(\mathbf{X} \notin C^* \,|\, f_1)$.

We want to show $\beta \leq \beta^*$.

We know $\alpha^* \leq \alpha$, ie $\int_{C^*} f_0(\mathbf{x})d\mathbf{x} \leq \int_C f_0(\mathbf{x})d\mathbf{x}$.

Also, on C we have $f_1(\mathbf{x}) > kf_0(\mathbf{x})$, while on $\bar{C}$ we have $f_1(\mathbf{x}) \leq kf_0(\mathbf{x})$.

Thus

$$\int_{\bar{C}^* \cap C} f_1(\mathbf{x})d\mathbf{x} \geq k \int_{\bar{C}^* \cap C} f_0(\mathbf{x})d\mathbf{x}, \qquad \int_{\bar{C} \cap C^*} f_1(\mathbf{x})d\mathbf{x} \leq k \int_{\bar{C} \cap C^*} f_0(\mathbf{x})d\mathbf{x}. \quad (1)$$

Hence

$$
\begin{aligned}
\beta - \beta^* &= \int_{\bar{C}} f_1(\mathbf{x})d\mathbf{x} - \int_{\bar{C}^*} f_1(\mathbf{x})d\mathbf{x} \\
&= \int_{\bar{C}\cap C^*} f_1(\mathbf{x})d\mathbf{x} + \int_{\bar{C}\cap\bar{C}^*} f_1(\mathbf{x})d\mathbf{x} - \int_{\bar{C}^*\cap C} f_1(\mathbf{x})d\mathbf{x} - \int_{\bar{C}\cap\bar{C}^*} f_1(\mathbf{x})d\mathbf{x} \\
&\leq k\int_{\bar{C}\cap C^*} f_0(\mathbf{x})d\mathbf{x} - k\int_{\bar{C}^*\cap C} f_0(\mathbf{x})d\mathbf{x} \qquad \text{by } (\mathbf{??}) \\
&= k\left\{\int_{\bar{C}\cap C^*} f_0(\mathbf{x})d\mathbf{x} + \int_{C\cap C^*} f_0(\mathbf{x})d\mathbf{x}\right\} \\
&\qquad -k\left\{\int_{\bar{C}^*\cap C} f_0(\mathbf{x})d\mathbf{x} + \int_{C\cap C^*} f_0(\mathbf{x})d\mathbf{x}\right\} \\
&= k(\alpha^* - \alpha) \\
&\leq 0.
\end{aligned}
$$

□

- We assume continuous densities to ensure that a LR test of exactly size α exists.
- The Neyman–Pearson Lemma shows that α and β cannot both be arbitrarily small.
- It says that the most powerful test (ie the one with the smallest Type II error probability), among tests with size smaller than or equal to α, is the size α likelihood ratio test.
- Thus we should fix $\mathbb{P}$(Type I error) at some level α and then use the Neyman–Pearson Lemma to find the best test.
- Here the hypotheses are not treated symmetrically; H_0 has precedence over H_1 and a Type I error is treated as more serious than a Type II error.
- H_0 is called the **null hypothesis** and H_1 is called the **alternative hypothesis**.
- The null hypothesis is a conservative hypothesis, ie one of "no change," "no bias," "no association," and is only rejected if we have clear evidence against it.
- H_1 represents the kind of departure from H_0 that is of interest to us.

Example 7.3

Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma_0^2)$, where σ_0^2 is known. We want to find the best size α test of $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, where μ_0 and μ_1 are known fixed values with $\mu_1 > \mu_0$.

$$\begin{aligned}
\Lambda_{\mathbf{x}}(H_0; H_1) &= \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_1)^2\right)}{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_0)^2\right)} \\
&= \exp\left(\frac{(\mu_1 - \mu_0)}{\sigma_0^2} n\bar{x} + \frac{n(\mu_0^2 - \mu_1^2)}{2\sigma_0^2}\right) \qquad \text{(check)}.
\end{aligned}$$

- This is an increasing function of $\bar{x}$, so for any k,

$$\Lambda_{\mathbf{x}} > k \Leftrightarrow \bar{x} > c \text{ for some } c.$$

- Hence we reject H_0 if $\bar{x} > c$ where c is chosen such that $\mathbb{P}(\bar{X} > c \,|\, H_0) = \alpha$.
- Under H_0, $\bar{X} \sim N(\mu_0, \sigma_0^2/n)$, so $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 \sim N(0, 1)$.
- Since $\bar{x} > c \Leftrightarrow z > c'$ for some c', the size α test rejects H_0 if $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 > z_\alpha$.

- Suppose $\mu_0 = 5$, $\mu_1 = 6$, $\sigma_0 = 1$, $\alpha = 0.05$, $n = 4$ and $\mathbf{x} = (5.1, 5.5, 4.9, 5.3)$, so that $\bar{x} = 5.2$.
- From tables, $z_{0.05} = 1.645$.
- We have $z = \frac{\sqrt{n}(\bar{x}-\mu_0)}{\sigma_0} = 0.4$ and this is less than 1.645, so $\mathbf{x}$ is not in the rejection region.
- We do not reject H_0 at the 5%- level; the data are consistent with H_0.
- This does not mean that H_0 is 'true', just that it cannot be ruled out.
- This is called a z-**test**. □

P-values

- In this example, LR tests reject H_0 if $z > k$ for some constant k.
- The size of such a test is $\alpha = \mathbb{P}(Z > k \,|\, H_0) = 1 - \Phi(k)$, and is decreasing as k increases.
- Our observed value z will be in the rejection region $\Leftrightarrow z > k \Leftrightarrow \alpha > p^* = \mathbb{P}(Z > z \,|\, H_0)$.
- The quantity p^* is called the p-**value** of our observed data **x**.
- For Example 7.3, $z = 0.4$ and so $p^* = 1 - \Phi(0.4) = 0.3446$.
- In general, the p-value is sometimes called the 'observed significance level' of **x** and is the probability under H_0 of seeing data that are 'more extreme' than our observed data **x**.
- Extreme observations are viewed as providing evidence againt H_0.

* The p-value has a Uniform(0,1) pdf under the null hypothesis. To see this for a z-test, note that

$$\mathbb{P}(p* < p \mid H_0) = \mathbb{P}\left([1 - \Phi(Z)] < p \mid H_0\right) = \mathbb{P}(Z > \Phi^{-1}(1-p) \mid H_0)$$

$$= 1 - \Phi\left(\Phi^{-1}(1-p)\right) = 1 - (1-p) = p.$$

Lecture 8. Composite hypotheses

Composite hypotheses, types of error and power

- For composite hypotheses like $H: \theta \geq 0$, the error probabilities do not have a single value.
- Define the **power function** $W(\theta) = \mathbb{P}(\mathbf{X} \in C \,|\, \theta) = \mathbb{P}(\text{reject } H_0 \,|\, \theta)$.
- We want $W(\theta)$ to be small on H_0 and large on H_1.
- Define the **size** of the test to be $\alpha = \sup_{\theta \in \Theta_0} W(\theta)$.
- For $\theta \in \Theta_1$, $1 - W(\theta) = \mathbb{P}(\text{Type II error} \,|\, \theta)$.
- Sometimes the Neyman–Pearson theory can be extended to one-sided alternatives.
- For example, in Example 7.3 we have shown that the most powerful size α test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ (where $\mu_1 > \mu_0$) is given by $C = \{\mathbf{x} : \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 > z_\alpha\}$.
- This critical region depends on μ_0, n, σ_0, α, on the fact that $\mu_1 > \mu_0$, but not on the particular value of μ_1.

- Hence this C defines the most powerful size α test of $H_0 : \mu = \mu_0$ against any μ_1 that is larger than μ_0.
- This test is then uniformly most powerful size α for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

Definition 8.1

A test specified by a critical region C is **uniformly most powerful** (UMP) size α test for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if

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

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

- UMP tests may not exist.
- However likelihood ratio tests are often UMP.

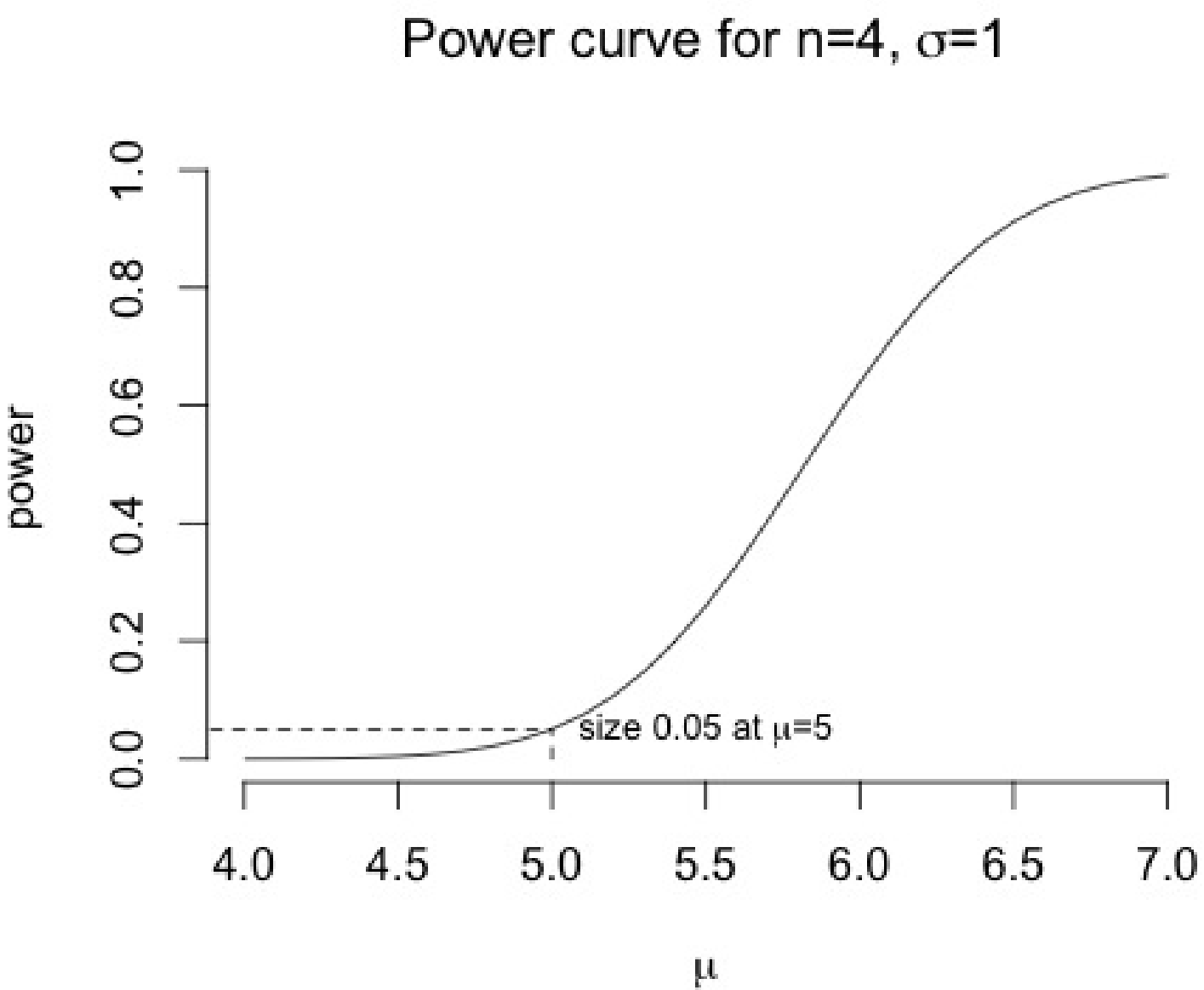
Example 8.2

Suppose $X_1, \ldots, X_n$ are iid $N(\mu_0, \sigma_0^2)$ where σ_0 is known, and we wish to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.

- First consider testing $H_0' : \mu = \mu_0$ against $H_1' : \mu = \mu_1$ where $\mu_1 > \mu_0$ (as in Example 7.3)
- As in Example 7.3, the Neyman-Pearson test of size α of H_0' against H_1' has $C = \{\mathbf{x} : \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 > z_\alpha\}$.
- We will show that C is in fact UMP for the composite hypotheses H_0 against H_1
- For $\mu \in \mathbb{R}$, the power function is

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

```
power= 1 - pnorm( qnorm(0.95) + sqrt(n) * (mu0-x) / sigma0 )
```



- We know $W(\mu_0) = \alpha$. (just plug in)
- $W(\mu)$ is an increasing function of μ.
- So $\sup_{\mu \leq \mu_0} W(\mu) = \alpha$, and (i) is satisfied.
- For (ii), observe that for any $\mu > \mu_0$, the Neyman Pearson size α test of H_0' vs H_1' has critical region C (the calculation in Example 7.3 depended only on the fact that $\mu > \mu_0$ and not on the particular value of μ_1.)
- Let C^* and W^* belong to any other test of H_0 vs H_1 of size $\leq \alpha$
- Then C^* can be regarded as a test of H_0 vs H_1 of size $\leq \alpha$, and NP-Lemma says that $W^*(\mu_1) \leq W(\mu_1)$
- This holds for all $\mu_1 > \mu_0$ and so (ii) is satisfied.
- So C is UMP size α for H_0 vs H_1. $\square$

Generalised likelihood ratio tests

- We now consider likelihood ratio tests for more general situations.
- Define the **likelihood of a composite hypothesis** $H : \theta \in \Theta$ given data $\mathbf{x}$ to be
$$L_{\mathbf{x}}(H) = \sup_{\theta \in \Theta} f(\mathbf{x} \mid \theta).$$
- So far we have considered disjoint hypotheses Θ_0, Θ_1, but often we are not interested in any specific alternative, and it is easier to take $\Theta_1 = \Theta$ rather than $\Theta_1 = \Theta \setminus \Theta_0$.
- Then
$$\Lambda_{\mathbf{x}}(H_0; H_1) = \frac{L_{\mathbf{x}}(H_1)}{L_{\mathbf{x}}(H_0)} = \frac{\sup_{\theta \in \Theta_1} f(\mathbf{x} \mid \theta)}{\sup_{\theta \in \Theta_0} f(\mathbf{x} \mid \theta)} (\geq 1), \tag{1}$$
with large values of $\Lambda_{\mathbf{x}}$ indicating departure from H_0.
- Notice that if $\Lambda^*_{\mathbf{x}} = \sup_{\theta \in \Theta \setminus \Theta_0} f(\mathbf{x} \mid \theta) / \sup_{\theta \in \Theta_0} f(\mathbf{x} \mid \theta)$, then $\Lambda_{\mathbf{x}} = \max\{1, \Lambda^*_{\mathbf{x}}\}$.

Example 8.3

Single sample: testing a given mean, known variance (z-test). Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma_0^2)$, with σ_0^2 known, and we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ (μ_0 is a given constant).

- Here $\Theta_0 = \{\mu_0\}$ and $\Theta = \mathbb{R}$.
- For the denominator in (1) we have $\sup_{\Theta_0} f(\mathbf{x}\,|\,\mu) = f(\mathbf{x}\,|\,\mu_0)$.
- For the numerator, we have $\sup_{\Theta} f(\mathbf{x}\,|\,\mu) = f(\mathbf{x}\,|\,\hat{\mu})$, where $\hat{\mu}$ is the mle, so $\hat{\mu} = \bar{x}$ (check).
- Hence

$$\Lambda_{\mathbf{x}}(H_0; H_1) = \frac{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum(x_i - \bar{x})^2\right)}{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\frac{1}{2\sigma_0^2}\sum(x_i - \mu_0)^2\right)},$$

  and we reject H_0 if $\Lambda_{\mathbf{x}}$ is 'large.'
- We find that

$$2\log\Lambda_{\mathbf{x}}(H_0; H_1) = \frac{1}{\sigma_0^2}\left[\sum(x_i - \mu_0)^2 - \sum(x_i - \bar{x})^2\right] = \frac{n}{\sigma_0^2}(\bar{x} - \mu_0)^2. \text{ (check)}$$

- Thus an equivalent test is to reject H_0 if $\left|\sqrt{n}(\bar{x} - \mu_0)/\sigma_0\right|$ is large.

- Under H_0, $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 \sim N(0,1)$ so the size α generalised likelihood test rejects H_0 if $\left|\sqrt{n}(\bar{x} - \mu_0)/\sigma_0\right| > z_{\alpha/2}$.
- Since $n(\bar{X} - \mu_0)^2/\sigma_0^2 \sim \chi_1^2$ if H_0 s true, this is equivalent to rejecting H_0 if $n(\bar{X} - \mu_0)^2/\sigma_0^2 > \chi_1^2(\alpha)$ (check that $z_{\alpha/2}^2 = \chi_1^2(\alpha)$). $\square$

Notes:

- This is a 'two-tailed' test - i.e. reject H_0 both for high and low values of $\bar{x}$.
- We reject H_0 if $\left|\sqrt{n}(\bar{x} - \mu_0)/\sigma_0\right| > z_{\alpha/2}$. A symmetric $100(1-\alpha)\%$ confidence interval for μ is $\bar{x} \pm z_{\alpha/2}\,\sigma_0/\sqrt{n}$. Therefore we reject H_0 iff μ_0 is not in this confidence interval (check).
- In later lectures the close connection between confidence intervals and hypothesis tests is explored further.

The 'generalised likelihood ratio test'

- The next theorem allows us to use likelihood ratio tests even when we cannot find the exact relevant null distribution.
- First consider the 'size' or 'dimension' of our hypotheses: suppose that H_0 imposes p independent restrictions on Θ, so for example, if and we have
  - $H_0 : \theta_{i_1} = a_1, \ldots, \theta_{i_p} = a_p$ ($a_1, \ldots, a_p$ given numbers),
  - $H_0 : A\theta = \mathbf{b}$ (A $p \times k$, $\mathbf{b}$ $p \times 1$ given),
  - $H_0 : \theta_i = f_i(\phi)$, $i = 1, \ldots, k$, $\phi = (\phi_1, \ldots, \phi_{k-p})$.
- Then Θ has 'k free parameters' and Θ_0 has '$k - p$ free parameters.'
- We write $|\Theta_0| = k - p$ and $|\Theta| = k$.

Theorem 8.4

(not proved)
Suppose $\Theta_0 \subseteq \Theta_1$, $|\Theta_1| - |\Theta_0| = p$. *Then under regularity conditions, as* $n \to \infty$, *with* $\mathbf{X} = (X_1, \ldots, X_n)$, X_i*'s iid, we have, if* H_0 *is true,*

$$2 \log \Lambda_{\mathbf{X}}(H_0; H_1) \sim \chi^2_p.$$

If H_0 *is not true, then* $2 \log \Lambda$ *tends to be larger. We reject* H_0 *if* $2 \log \Lambda > c$ *where* $c = \chi^2_p(\alpha)$ *for a test of approximately size* α.

In Example 8.3, $|\Theta_1| - |\Theta_0| = 1$, and in this case we saw that under H_0, $2 \log \Lambda \sim \chi^2_1$ *exactly* for all n in that particular case, rather than just approximately for large n as the Theorem shows.

(Often the likelihood ratio is calculated with the null hypothesis in the numerator, and so the test statistic is $-2 \log \Lambda_{\mathbf{X}}(H_1; H_0)$.)

Lecture 9. Tests of goodness-of-fit and independence

Goodness-of-fit of a fully-specified null distribution

Suppose the observation space $\mathcal{X}$ is partitioned into k sets, and let p_i be the probability that an observation is in set i, $i = 1, \ldots, k$.

Consider testing H_0 : the p_i's arise from a fully specified model against H_1: the p_i's are unrestricted (but we must still have $p_i \geq 0$, $\sum p_i = 1$).

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

Example 9.1

Birth month of admissions to Oxford and Cambridge in 2012

Month	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
n_i	470	515	470	457	473	381	466	457	437	396	384	394

Are these compatible with a uniform distribution over the year? □

- Out of n independent observations let N_i be the number of observations in the ith set.
- So $(N_1, \ldots, N_k) \sim \text{Multinomial}(n; p_1, \ldots, p_k)$.
- For a generalised likelihood ratio test of H_0, we need to find the maximised likelihood under H_0 and H_1.
- **Under $\mathbf{H_1}$**: $\text{like}(p_1, \ldots, p_k) \propto p_1^{n_1} \ldots p_k^{n_k}$ so the loglikelihood is $l = \text{constant} + \sum n_i \log p_i$.

  We want to maximise this subject to $\sum p_i = 1$.

  By considering the Lagrangian $\mathcal{L} = \sum n_i \log p_i - \lambda(\sum p_i - 1)$, we find mle's $\hat{p}_i = n_i/n$. Also $|\Theta_1| = k - 1$.
- **Under $\mathbf{H_0}$**: H_0 specifies the values of the p_i's completely, $p_i = \tilde{p}_i$ say, so $|\Theta_0| = 0$.
- Putting these two together, we find

$$2 \log \Lambda = 2 \log \left(\frac{\hat{p}_1^{n_1} \ldots \hat{p}_k^{n_k}}{\tilde{p}_1^{n_1} \ldots \tilde{p}_k^{n_k}} \right) = 2 \sum n_i \log \left(\frac{n_i}{n \tilde{p}_i} \right). \tag{1}$$

- Here $|\Theta_1| - |\Theta_0| = k - 1$, so we reject H_0 if $2 \log \Lambda > \chi^2_{k-1}(\alpha)$ for an approximate size α test.

Example 9.1 continued:

Under H_0 (no effect of month of birth), $\tilde{p}_i$ is the proportion of births in month i in 1993/1994 - this is *not* simply proportional to number of days in month, as there is for example an excess of September births (the 'Christmas effect').

Month	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
n_i	470	515	470	457	473	381	466	457	437	396	384	394
$100\tilde{p}_i$	8.8	8.5	7.9	8.3	8.3	7.6	8.6	8.3	8.6	8.5	8.5	8.3
$n\tilde{p}_i$	466.4	448.2	416.3	439.2	436.9	402.3	456.3	437.6	457.2	450.0	451.3	438.2

- $2 \log \Lambda = 2 \sum n_i \log \left(\frac{n_i}{n\tilde{p}_i} \right) = 44.9$
- $\mathbb{P}(\chi^2_{11} > 44.86) = 3x10^{-9}$, which is our p-value.
- Since this is certainly less than 0.001, we can reject H_0 at the 0.1% level, or can say 'significant at the 0.1% level'.
- NB The traditional levels for comparison are $\alpha = 0.05, 0.01, 0.001$, roughly corresponding to 'evidence', 'strong evidence', and 'very strong evidence'.

Likelihood ratio tests

A similar common situation has $H_0 : p_i = p_i(\theta)$ for some parameter θ and H_1 as before. Now $|\Theta_0|$ is the number of independent parameters to be estimated under H_0.

Under H_0: we find mle $\hat{\theta}$ by maximising $\sum n_i \log p_i(\theta)$, and then

$$2 \log \Lambda = 2 \log \left(\frac{\hat{p}_1^{n_1} \dots \hat{p}_k^{n_k}}{p_1(\hat{\theta})^{n_1} \dots p_k(\hat{\theta})^{n_k}} \right) = 2 \sum n_i \log \left(\frac{n_i}{n p_i(\hat{\theta})} \right). \qquad (2)$$

Now the degrees of freedom are $k - 1 - |\Theta_0|$, and we reject H_0 if $2 \log \Lambda > \chi^2_{k-1-|\Theta_0|}(\alpha)$.

Pearson's Chi-squared tests

Notice that (**??**) and (**??**) are of the same form.

Let $o_i = n_i$ (the observed number in ith set) and let e_i be $n\tilde{p}_i$ in (**??**) or $np_i(\hat{\theta})$ in (**??**). Let $\delta_i = o_i - e_i$. Then

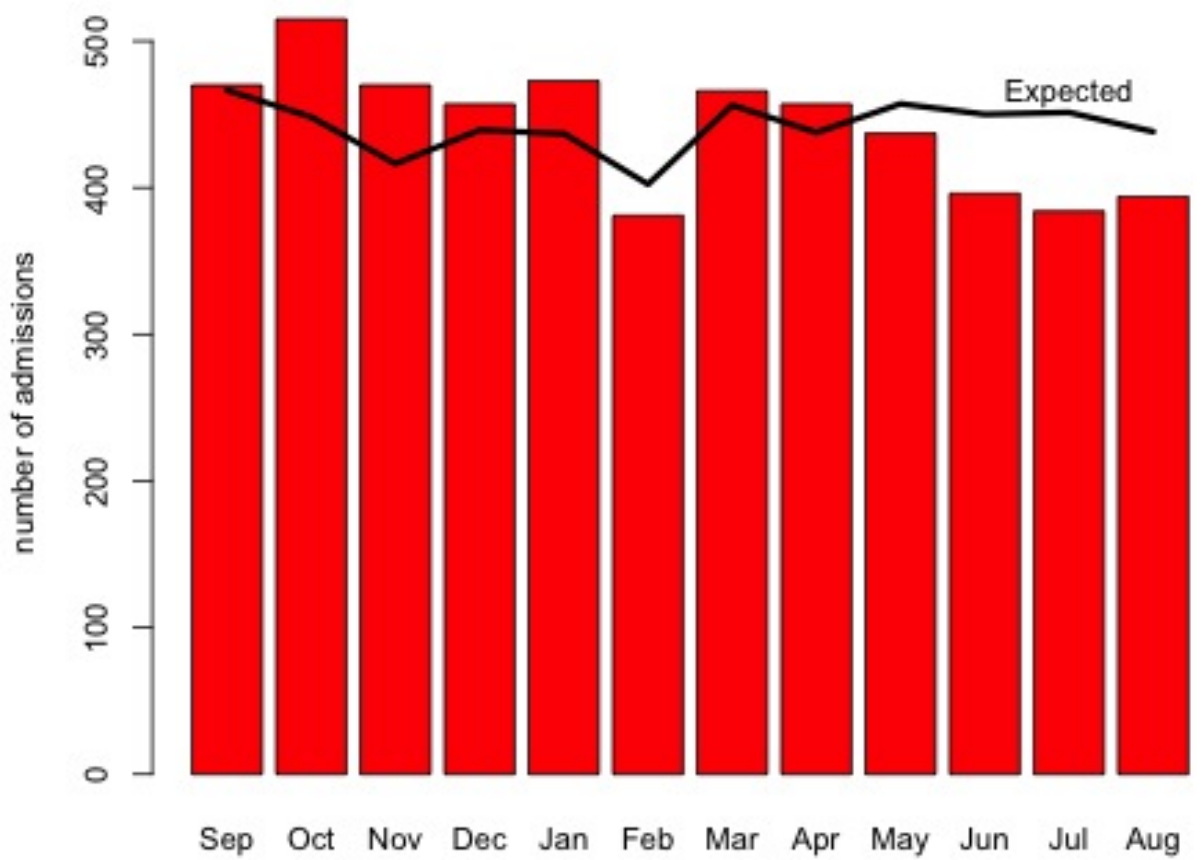
$$\begin{aligned} 2\log\Lambda &= 2\sum o_i \log\left(\frac{o_i}{e_i}\right) \\ &= 2\sum (e_i+\delta_i)\log\left(1+\frac{\delta_i}{e_i}\right) \\ &\approx 2\sum\left(\delta_i + \frac{\delta_i^2}{e_i} - \frac{\delta_i^2}{2e_i}\right) \\ &= \sum \frac{\delta_i^2}{e_i} = \sum \frac{(o_i-e_i)^2}{e_i}, \end{aligned}$$

where we have assumed $\log\left(1+\frac{\delta_i}{e_i}\right) \approx \frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i^2}$, ignored terms in δ_i^3, and used that $\sum \delta_i = 0$ (check).

This is **Pearson's chi-squared statistic**; we refer it to $\chi^2_{k-1-|\Theta_0|}$.

Example 9.1 continued using R:

```
chisq.test(n,p=ptilde)
data:  n
X-squared = 44.6912, df = 11, p-value = 5.498e-06
```



Example 9.2

Mendel crossed 556 smooth yellow male peas with wrinkled green female peas. From the progeny let

- N_1 be the number of smooth yellow peas,
- N_2 be the number of smooth green peas,
- N_3 be the number of wrinkled yellow peas,
- N_4 be the number of wrinkled green peas.

We wish to test the goodness of fit of the model
$H_0 : (p_1, p_2, p_3, p_4) = (9/16, 3/16, 3/16, 1/16)$, the proportions predicted by Mendel's theory.

Suppose we observe $(n_1, n_2, n_3, n_4) = (315, 108, 102, 31)$.

We find $(e_1, e_2, e_3, e_4) = (312.75, 104.25, 104.25, 34.75)$, $2 \log \Lambda = 0.618$ and $\sum \frac{(o_i - e_i)^2}{e_i} = 0.604$.

Here $|\Theta_0| = 0$ and $|\Theta_1| = 4 - 1 = 3$, so we refer our test statistics to χ^2_3.

Since $\chi^2_3(0.05) = 7.815$ we see that neither value is significant at 5% level, so there is no evidence against Mendel's theory.

In fact the p-value is approximately $\mathbb{P}(\chi^2_3 > 0.6) \approx 0.96$. $\square$

NB So in fact could be considered as a suspiciously good fit

Example 9.3

In a genetics problem, each individual has one of three possible genotypes, with probabilities p_1, p_2, p_3. Suppose that we wish to test $H_0 : p_i = p_i(\theta)$ $i = 1, 2, 3$, where $p_1(\theta) = \theta^2$, $p_2(\theta) = 2\theta(1-\theta)$, $p_3(\theta) = (1-\theta)^2$, for some $\theta \in (0, 1)$.

We observe $N_i = n_i$, $i = 1, 2, 3$ ($\sum N_i = n$).

Under H_0, the mle $\hat{\theta}$ is found by maximising

$$\sum n_i \log p_i(\theta) = 2n_1 \log \theta + n_2 \log(2\theta(1-\theta)) + 2n_3 \log(1-\theta).$$

We find that $\hat{\theta} = (2n_1 + n_2)/(2n)$ (check).

Also $|\Theta_0| = 1$ and $|\Theta_1| = 2$.

Now substitute $p_i(\hat{\theta})$ into (2), or find the corresponding Pearson's chi-squared statistic, and refer to χ^2_1. $\square$

Testing independence in contingency tables

A table in which observations or individuals are classified according to two or more criteria is called a **contingency table**.

Example 9.4

500 people with recent car changes were asked about their previous and new cars.

		New car		
		Large	Medium	Small
Previous car	Large	56	52	42
	Medium	50	83	67
	Small	18	51	81

This is a two-way contingency table: each person is classified according to previous car size and new car size. □

- Consider a two-way contingency table with r rows and c columns.
- For $i = 1, \ldots, r$ and $j = 1, \ldots, c$ let p_{ij} be the probability that an individual selected at random from the population under consideration is classified in row i and column j (ie in the (i, j) cell of the table).
- Let $p_{i+} = \sum_j p_{ij} = \mathbb{P}(\text{in row } i)$, and $p_{+j} = \sum_i p_{ij} = \mathbb{P}(\text{in column } j)$.
- We must have $p_{++} = \sum_i \sum_j p_{ij} = 1$, ie $\sum_i p_{i+} = \sum_j p_{+j} = 1$.
- Suppose a random sample of n individuals is taken, and let n_{ij} be the number of these classified in the (i, j) cell of the table.
- Let $n_{i+} = \sum_j n_{ij}$ and $n_{+j} = \sum_i n_{ij}$, so $n_{++} = n$.
- We have

$$(N_{11}, N_{12}, \ldots, N_{1c}, N_{21}, \ldots, N_{rc}) \sim \text{Multinomial}(n; p_{11}, p_{12}, \ldots, p_{1c}, p_{21}, \ldots, p_{rc})$$

- We may be interested in testing the null hypothesis that the two classifications are independent, so test
  - $H_0 : p_{ij} = p_{i+}p_{+j}$, $i = 1, \ldots, r$, $j = 1, \ldots, c$ (with $\sum_i p_{i+} = 1 = \sum_j p_{+j}$, $p_{i+}, p_{+j} \geq 0$),
  - $H_1 : p_{ij}$'s unrestricted (but as usual need $p_{++} = 1$, $p_{ij} \geq 0$).
- Under H_1 the mle's are $\hat{p}_{ij} = n_{ij}/n$.
- Under H_0, using Lagrangian methods, the mle's are $\hat{p}_{i+} = n_{i+}/n$ and $\hat{p}_{+j} = n_{+j}/n$.
- Write o_{ij} for n_{ij} and let $e_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n_{i+}n_{+j}/n$.
- Then

$$2 \log \Lambda = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} o_{ij} \log \left(\frac{o_{ij}}{e_{ij}} \right) \approx \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  using the same approximating steps as for Pearson's Chi-squared test.
- We have $|\Theta_1| = rc - 1$, because under H_1 the p_{ij}'s sum to one.
- Further, $|\Theta_0| = (r-1) + (c-1)$, because $p_{1+}, \ldots, p_{r+}$ must satisfy $\sum_i p_{i+} = 1$ and $p_{+1}, \ldots, p_{+c}$ must satisfy $\sum_j p_{+j} = 1$.
- So $|\Theta_1| - |\Theta_0| = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$.

Example 9.5

In Example 9.4, suppose we wish to test H_0 : the new and previous car sizes are independent.

We obtain:

	o_{ij}	New car Large	Medium	Small	
Previous car	Large	56	52	42	150
	Medium	50	83	67	200
	Small	18	51	81	150
		124	186	190	500

	e_{ij}	New car Large	Medium	Small	
Previous car	Large	37.2	55.8	57.0	150
	Medium	49.6	74.4	76.0	200
	Small	37.2	55.8	57.0	150
		124	186	190	500

Note the margins are the same.

Then $\sum\sum \frac{(o_{ij}-e_{ij})^2}{e_{ij}} = 36.20$, and df $= (3-1)(3-1) = 4$.

From tables, $\chi^2_4(0.05) = 9.488$ and $\chi^2_4(0.01) = 13.28$.

So our observed value of 36.20 is significant at the 1% level, ie there is strong evidence against H_0, so we conclude that the new and present car sizes are not independent.

It may be informative to look at the contributions of each cell to Pearson's chi-squared:

		New car		
		Large	Medium	Small
Previous car	Large	9.50	0.26	3.95
	Medium	0.00	0.99	1.07
	Small	9.91	0.41	10.11

It seems that more owners of large cars than expected under H_0 bought another large car, and more owners of small cars than expected under H_0 bought another small car.

Fewer than expected changed from a small to a large car. □

Lecture 10. Tests of homogeneity, and connections to confidence intervals

Tests of homogeneity

Example 10.1

150 patients were randomly allocated to three groups of 50 patients each. Two groups were given a new drug at different dosage levels, and the third group received a placebo. The responses were as shown in the table below.

	Improved	No difference	Worse	
Placebo	18	17	15	50
Half dose	20	10	20	50
Full dose	25	13	12	50
	63	40	47	150

Here the row totals are fixed in advance, in contrast to Example 9.4 where the row totals are random variables.

For the above table, we may be interested in testing H_0: the probability of "improved" is the same for each of the three treatment groups, and so are the probabilities of "no difference" and "worse," ie H_0 says that we have homogeneity down the rows. □

- In general, we have independent observations from r multinomial distributions each of which has c categories,
- ie we observe an $r \times c$ table (n_{ij}), $i = 1, \ldots, r$, $j = 1, \ldots, c$, where $(N_{i1}, \ldots, N_{ic}) \sim \text{Multinomial}(n_{i+}; p_{i1}, \ldots, p_{ic})$ independently for $i = 1, \ldots, r$.
- We test $H_0 : p_{1j} = p_{2j} = \ldots = p_{rj} = p_j$ say, $j = 1, \ldots, c$ where $p_+ = 1$, and $H_1 : p_{ij}$ are unrestricted (but with $p_{ij} \geq 0$ and $p_{i+} = 1$, $i = 1, \ldots, r$).
- **Under H_1**: $\text{like}((p_{ij})) = \prod_{i=1}^{r} \frac{n_{i+}!}{n_{i1}! \ldots n_{ic}!} p_{i1}^{n_{i1}} \ldots p_{ic}^{n_{ic}}$, and
  $\text{loglike} = \text{constant} + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log p_{ij}$.
  Using Lagrangian methods (with constraints $p_{i+} = 1$, $i = 1, \ldots, r$) we find $\hat{p}_{ij} = n_{ij}/n_{i+}$.
- **Under H_0**:
  $\text{loglike} = \text{constant} + \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \log p_j = \text{constant} + \sum_{j=1}^{c} n_{+j} \log p_j$.
  Lagrangian techniques here (with constraint $\sum p_j = 1$) give $\hat{p}_j = n_{+j}/n_{++}$.

- Hence

$$\begin{aligned} 2\log\Lambda &= 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{\hat{p}_{ij}}{\hat{p}_j}\right) \\ &= 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{n_{ij}}{n_{i+}n_{+j}/n_{++}}\right), \end{aligned}$$

  ie the same as in Example 9.5.
- We have $|\Theta_1| = r(c-1)$ (because there are $c-1$ free parameters for each of r distributions).
- Also $|\Theta_0| = c-1$ (because H_0 has c parameters $p_1, \ldots, p_c$ with constraint $p_+ = 1$).
- So df $= |\Theta_1| - |\Theta_0| = r(c-1) - (c-1) = (r-1)(c-1)$, and under H_0, $2\log\Lambda$ is approximately $\chi^2_{(r-1)(c-1)}$ (ie same as in Example 9.5).
- We reject H_0 if $2\log\Lambda > \chi^2_{(r-1)(c-1)}(\alpha)$ for an approximate size α test.
- Let $o_{ij} = n_{ij}$, $e_{ij} = n_{i+}n_{+j}/n_{++}$ $\delta_{ij} = o_{ij} - e_{ij}$, and use the same approximating steps as for Pearson's Chi-squared to see that $2\log\Lambda \approx \sum_i \sum_j \frac{(o_{ij}-e_{ij})^2}{e_{ij}}$.

Example 10.2

Example 10.1 continued

o_{ij}	Improved	No difference	Worse	
Placebo	18	17	15	50
Half dose	20	10	20	50
Full dose	25	13	12	50
	63	40	47	150

e_{ij}	Improved	No difference	Worse	
Placebo	21	13.3	15.7	50
Half dose	21	13.3	15.7	50
Full dose	21	13.3	15.7	50

We find $2 \log \Lambda = 5.129$, and we refer this to χ^2_4.

From tables, $\chi^2_4(0.05) = 9.488$, so our observed value is not significant at 5% level, and the data are consistent with H_0.

We conclude that there is no evidence for a difference between the drug at the given doses and the placebo.

For interest, $\sum\sum(o_{ij} - e_{ij})^2/e_{ij} = 5.173$, leading to the same conclusion. □

Confidence intervals and hypothesis tests

- Confidence intervals or sets can be obtained by inverting hypothesis tests, and vice versa.
- Define the **acceptance region** A of a test to be the complement of the critical region C.
- NB By 'acceptance', we really mean 'non-rejection'
- Suppose $X_1, \ldots, X_n$ have joint pdf $f_{\mathbf{X}}(\mathbf{x}|\theta)$, $\theta \in \Theta$.

Theorem 10.3

(i) *Suppose that for every $\theta_0 \in \Theta$ there is a size α test of $H_0 : \theta = \theta_0$. Denote the acceptance region by $A(\theta_0)$. Then the set $I(\mathbf{X}) = \{\theta : \mathbf{X} \in A(\theta)\}$ is a $100(1-\alpha)\%$ confidence set for θ.*

(ii) *Suppose $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set for θ. Then $A(\theta_0) = \{\mathbf{X} : \theta_0 \in I(\mathbf{X})\}$ is an acceptance region for a size α test of $H_0 : \theta = \theta_0$.*

Proof:

- First note that $\theta_0 \in I(\mathbf{X}) \Leftrightarrow \mathbf{X} \in A(\theta_0)$.
- For (i), since the test is size α, we have $\mathbb{P}(\text{accept } H_0 \,|\, H_0 \text{ is true}) = \mathbb{P}(\mathbf{X} \in A(\theta_0) \,|\, \theta = \theta_0) = 1 - \alpha$
- And so $\mathbb{P}(I(\mathbf{X}) \ni \theta_0 \,|\, \theta = \theta_0) = \mathbb{P}(\mathbf{X} \in A(\theta_0) \,|\, \theta = \theta_0) = 1 - \alpha$.
- For (ii), since $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence set, we have $\mathbb{P}(I(\mathbf{X}) \ni \theta_0 \,|\, \theta = \theta_0) = 1 - \alpha$.
- So $\mathbb{P}(\mathbf{X} \in A(\theta_0) \,|\, \theta = \theta_0) = \mathbb{P}(I(\mathbf{X}) \ni \theta_0 \,|\, \theta = \theta_0) = 1 - \alpha$. $\square$

In words,

(i) says that a $100(1-\alpha)\%$ confidence set for θ consists of all those values of θ_0 for which $H_0 : \theta = \theta_0$ is not rejected at level α on the basis of $\mathbf{X}$,

(ii) says that given a confidence set, we define the test by rejecting θ_0 if it is not in the confidence set.

Example 10.4

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ random variables and we want a 95% confidence set for μ.

- One way is to use the above theorem and find the confidence set that belongs to the hypothesis test that we found in Example 10.1.
- Using Example 8.3 (with $\sigma_0^2 = 1$), we find a test of size 0.05 of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ that rejects H_0 when $|\sqrt{n}(\bar{x} - \mu_0)| > 1.96$ (1.96 is the upper 2.5% point of $N(0, 1)$).
- Then $I(\mathbf{X}) = \{\mu : \mathbf{X} \in A(\mu)\} = \{\mu : |\sqrt{n}(\bar{X} - \mu)| < 1.96\}$ so a 95% confidence set for μ is $(\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$.
- This is the same confidence interval we found in Example 5.2. □

Simpson's paradox*

For five subjects in 1996, the admission statistics for Cambridge were as follows:

	Women			**Men**		
	Applied	Accepted	%	Applied	Accepted	%
Total	1184	274	23 %	2470	584	24%

This looks like the acceptance rate is higher for men. But by subject...

	Women			**Men**		
	Applied	Accepted	%	Applied	Accepted	%
Computer Science	26	7	27%	228	58	25%
Economics	240	63	26%	512	112	22%
Engineering	164	52	32%	972	252	26%
Medicine	416	99	24%	578	140	24%
Veterinary medicine	338	53	16%	180	22	12%
Total	1184	274	23 %	2470	584	24%

In all subjects, the acceptance rate was higher for women!

Explanation: women tend to apply for subjects with the lowest acceptance rates.

This shows the danger of pooling (or collapsing) contingency tables.

Lecture 11. Multivariate Normal theory

Properties of means and covariances of vectors

- A random (column) vector $\mathbf{X} = (X_1, .., X_n)^T$ has mean

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_n))^T = (\mu_1, .., \mu_n)^T$$

  and covariance matrix

$$\text{cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T] = (\text{cov}(X_i, X_j))_{i,j},$$

  provided the relevant expectations exist.

- For $m \times n$ A,

$$\mathbb{E}[A\mathbf{X}] = A\boldsymbol{\mu},$$

  and

$$\text{cov}(A\mathbf{X}) = A\,\text{cov}(\mathbf{X})\,A^T, \tag{1}$$

  since $\text{cov}(A\mathbf{X}) = \mathbb{E}\left[(AX - \mathbb{E}(AX))(AX - \mathbb{E}(AX))^T)\right] =$
  $\mathbb{E}\left[A(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T A^T\right].$

- Define $\text{cov}(V, W)$ to be a matrix with (i, j) element $\text{cov}(V_i, W_j)$.
  Then $\text{cov}(A\mathbf{X}, B\mathbf{X}) = A\,\text{cov}(\mathbf{X})\,B^T$. (check. Important for later)

Multivariate normal distribution

- Recall that a univariate normal $X \sim \mathrm{N}(\mu, \sigma^2)$ has density

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

  and mgf

$$M_X(s) = \mathbb{E}[e^{sX}] = \exp\left(\mu s + \tfrac{1}{2}\sigma^2 s^2\right).$$

- $\mathbf{X}$ has a **multivariate normal distribution** if, for every $\mathbf{t} \in \mathbb{R}^n$, the rv $\mathbf{t}^T\mathbf{X}$ has a normal distribution.
- If $\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{X}) = \Sigma$, we write $\mathbf{X} \sim \mathrm{N}_n(\boldsymbol{\mu}, \Sigma)$.
- Note Σ is symmetric and is non-negative definite because by (1), $\mathbf{t}^T\Sigma\mathbf{t} = \mathrm{var}(\mathbf{t}^T\mathbf{X}) \geq 0$.
- By (1), $\mathbf{t}^T X \sim \mathrm{N}(\mathbf{t}^T\mu, \mathbf{t}^T\Sigma\mathbf{t})$ and so has mgf

$$M_{\mathbf{t}^T\mathbf{X}}(s) = \mathbb{E}[e^{s\mathbf{t}^T\mathbf{X}}] = \exp\left(\mathbf{t}^T\boldsymbol{\mu}s + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}s^2\right).$$

- Hence $\mathbf{X}$ has mgf

$$M_{\mathbf{X}}(\mathbf{t}) == \mathbb{E}[e^{\mathbf{t}^T\mathbf{X}}] = M_{\mathbf{t}^T\mathbf{X}}(1) = \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \tfrac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right). \qquad (2)$$

Proposition 11.1

(i) *If* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$ *and* A *is* $m \times n$, *then* $A\mathbf{X} \sim N_m(A\boldsymbol{\mu}, A\Sigma A^T)$

(ii) *If* $\mathbf{X} \sim N_n(\mathbf{0}, \sigma^2 I)$ *then*

$$\frac{\|\mathbf{X}\|^2}{\sigma^2} = \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} = \sum \frac{X_i^2}{\sigma^2} \sim \chi_n^2.$$

Proof:

(i) from exercise sheet 3.

(ii) Immediate from definition of χ_n^2. $\square$

Note that we often write $||X||^2 \sim \sigma^2\chi_n^2$.

Proposition 11.2

Let $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$, $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, *where* $\mathbf{X}_i$ *is a* $n_i \times 1$ *column vector, and* $n_1 + n_2 = n$. *Write similarly* $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$, *and* $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, *where* Σ_{ij} *is* $n_i \times n_j$. *Then*

(i) $\mathbf{X}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \Sigma_{ii})$,

(ii) $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent iff* $\Sigma_{12} = 0$.

Proof:

(i) See Example sheet 3.

(ii) From (2), $M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right)$, $\mathbf{t} \in \mathbb{R}^n$. Write

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\mathbf{t}_1{}^T\boldsymbol{\mu}_1 + \mathbf{t}_2{}^T\boldsymbol{\mu}_2 + \frac{1}{2}\mathbf{t}_1{}^T\Sigma_{11}\mathbf{t}_1 + \frac{1}{2}\mathbf{t}_2{}^T\Sigma_{22}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_1{}^T\Sigma_{12}\mathbf{t}_2 + \frac{1}{2}\mathbf{t}_2{}^T\Sigma_{21}\mathbf{t}_1\right).$$

From (i), $M_{\mathbf{X}_i}(\mathbf{t}_i) = \exp\left(\mathbf{t}_i{}^T\boldsymbol{\mu}_i + \frac{1}{2}\mathbf{t}_i^T\Sigma_{ii}\mathbf{t}_i\right)$ so $M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}_1}(\mathbf{t}_1)M_{\mathbf{X}_2}(\mathbf{t}_2)$, for all $\mathbf{t} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix}$ iff $\Sigma_{12} = 0$.

□

Density for a multivariate normal

When Σ is positive definite, then $\mathbf{X}$ has pdf

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\mu},\Sigma) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\left(\frac{1}{\sqrt{2\pi}}\right)^{n} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \qquad \mathbf{x}\in\mathbb{R}^{n}.$$

Normal random samples

We now consider $\bar{X} = \frac{1}{n}\sum X_i$, and $S_{XX} = \sum (X_i - \bar{X})^2$ for univariate normal data.

Theorem 11.3

(Joint distribution of $\bar{X}$ and S_{XX}) Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, $\bar{X} = \frac{1}{n}\sum X_i$, and $S_{XX} = \sum (X_i - \bar{X})^2$. Then

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

Proof

We can write the joint density as $\mathbf{X} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 I)$, where $\boldsymbol{\mu} = \mu \mathbf{1}$ ($\mathbf{1}$ is a $n \times 1$ column vector of 1's).

Let A be the $n \times n$ orthogonal matrix

$$A = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2\times 1}} & \frac{-1}{\sqrt{2\times 1}} & 0 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{3\times 2}} & \frac{1}{\sqrt{3\times 2}} & \frac{-2}{\sqrt{3\times 2}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{bmatrix}.$$

So $A^T A = A A^T = I$. (check)

(Note that the rows form an orthonormal basis of $\mathbb{R}^n$.)

(Strictly, we just need an orthogonal matrix with the first row matching that of A above.)

- By Proposition 11.1(i), $\mathbf{Y} = A\mathbf{X} \sim \mathrm{N}_n(A\boldsymbol{\mu}, A\sigma^2 I A^T) \sim \mathrm{N}_n(A\boldsymbol{\mu}, \sigma^2 I)$, since $AA^T = I$.
- We have $A\boldsymbol{\mu} = \begin{pmatrix} \sqrt{n}\mu \\ 0 \\ . \\ . \\ 0 \end{pmatrix}$, so $Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i = \sqrt{n}\bar{X} \sim \mathrm{N}(\sqrt{n}\mu, \sigma^2)$ (Prop 11.1 (ii))

  and $Y_i \sim \mathrm{N}(0, \sigma^2), i = 2, ..., n$ and $Y_1, ..., Y_n$ are independent.
- Note also that

$$\begin{aligned} Y_2^2 + \ldots + Y_n^2 &= \mathbf{Y}^T\mathbf{Y} - Y_1^2 = \mathbf{X}^T A^T A \mathbf{X} - Y_1^2 = \mathbf{X}^T\mathbf{X} - n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = S_{XX}. \end{aligned}$$

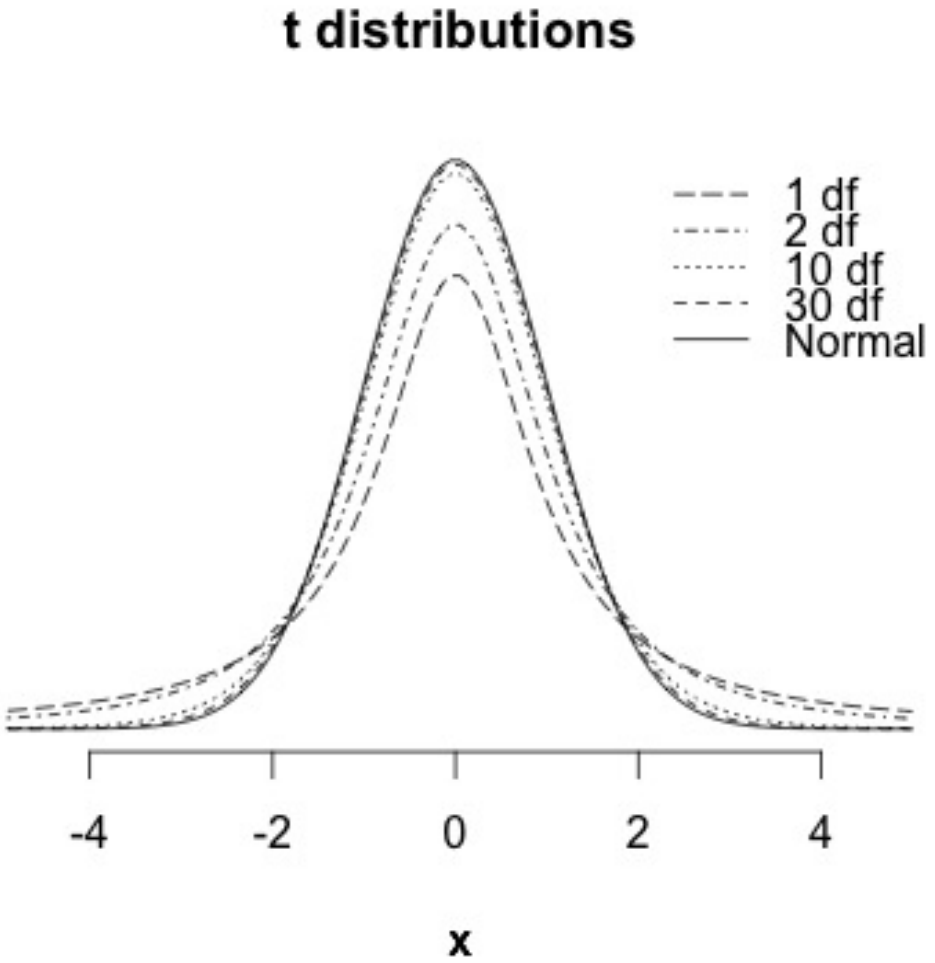
- To prove (ii), note that $S_{XX} = Y_2^2 + \ldots + Y_n^2 \sim \sigma^2 \chi^2_{n-1}$ (from definition of χ^2_{n-1}).
- FInally, for (iii), since Y_1 and $Y_2, ..., Y_n$ are independent (Prop 11.2 (ii)), so are $\bar{X}$ and S_{XX}. $\square$

Student's t-distribution

- Suppose that Z and Y are independent, $Z \sim N(0,1)$ and $Y \sim \chi^2_k$.
- Then $T = \frac{Z}{\sqrt{Y/k}}$ is said to have a t-**distribution** on k degrees of freedom, and we write $T \sim t_k$.
- The density of t_k turns out to be

$$f_T(t) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \frac{1}{\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}, \qquad t \in \mathbb{R}.$$

- This density is symmetric, bell-shaped, and has a maximum at $t = 0$, rather like the standard normal density.
- However, it can be shown that $\mathbb{P}(T > t) > \mathbb{P}(Z > t)$ for all $t > 0$, and that the t_k distribution approaches a normal distribution as $k \to \infty$.
- $\mathbb{E}_k(T) = 0$ for $k > 1$, otherwise undefined.
- $\mathrm{var}_k(T) = \frac{k}{k-2}$ for $k > 2$, $= \infty$ if $k = 2$, otherwise undefined.
- $k = 1$ is known as the Cauchy distribution, and has an undefined mean and variance.



Let $t_k(\alpha)$ be the upper $100\alpha\%$ point of the t_k- distribution, so that $\mathbb{P}(T > t_k(\alpha)) = \alpha$. There are tables of these percentage points.

Application of Student's t-distribution to normal random samples

- Let $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$.
- From Theorem 11.3 $\bar{X} \sim N(\mu, \sigma^2/n)$ so $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$.
- Also $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$ independently of $\bar{X}$ and hence of Z.
- Hence

$$\frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{S_{XX}/((n-1)\sigma^2)}} \sim t_{n-1}, \text{ ie } \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1}. \tag{3}$$

- Let $\tilde{\sigma}^2 = \frac{S_{XX}}{n-1}$. Note this is an unbiased estimator, as $\mathbb{E}(S_{XX}) = (n-1)\sigma^2$.
- Then a $100(1-\alpha)\%$ CI for μ is found from

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

and has endpoints

$$\bar{X} \pm \frac{\tilde{\sigma}}{\sqrt{n}} t_{n-1}(\tfrac{\alpha}{2}).$$

- See example sheet 3 for use of t distributions in hypothesis tests.

Lecture 12. The linear model

Introduction to linear models

- Linear models can be used to explain or model the relationship between a *response*, or *dependent*, variable and one or more *explanatory variables*, or *covariates* or *predictors*.
- For example, how do motor insurance claims depend on the age and sex of the driver, and where they live?

  Here the claim rate is the response, and age, sex and region are explanatory variables, assumed known.

- In the *linear model*, we assume our n observations (responses) are $Y_1, .., Y_n$ are modelled as

$$Y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where

- $\beta_1, .., \beta_p$ are unknown parameters, $n > p$
- $x_{i1}, .., x_{ip}$ are the values of p covariates for the ith response (assumed known)
- $\varepsilon_1, .., \varepsilon_n$ are independent (or possible just uncorrelated) random variables with mean 0 and variance σ^2.

From (1),

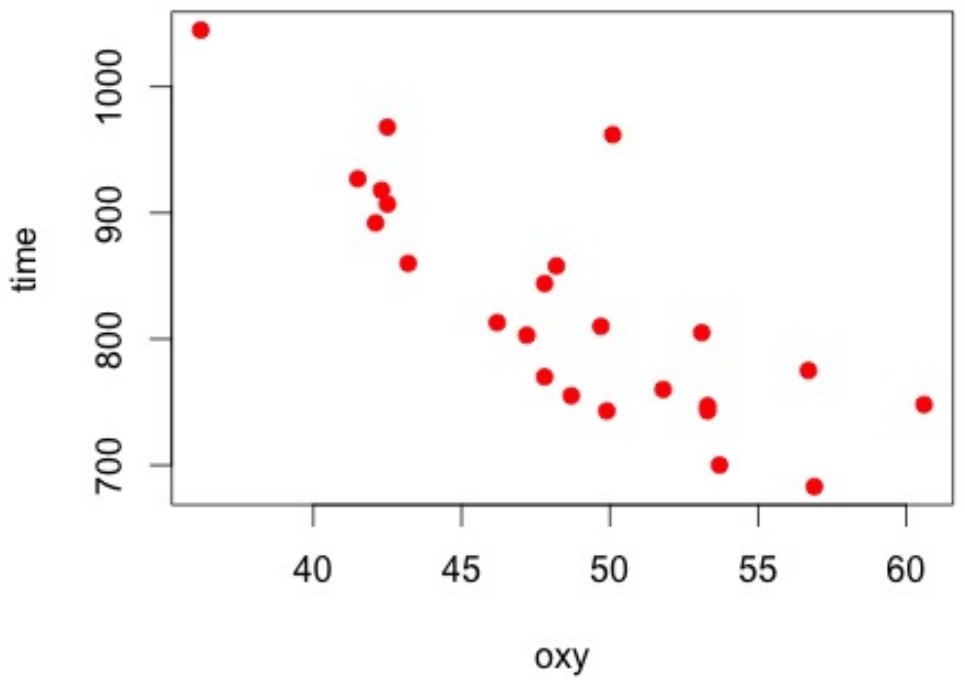
- $\mathbb{E}(Y_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$
- $\text{var}(Y_i) = \text{var}(\varepsilon_i) = \sigma^2$
- $Y_1, .., Y_n$ are independent (or uncorrelated).

Note that (1) is linear in the parameters $\beta_1, .., \beta_p$ (there are a wide range of more complex models which are non-linear in the parameters).

Example 12.1

For each of 24 males, the maximum volume of oxygen uptake in the blood and the time taken to run 2 miles (in minutes) were measured. Interest lies on how the time to run 2 miles depends on the oxygen uptake.

```
oxy=c(42.3,53.1,42.1,50.1,42.5,42.5,47.8,49.9,
      36.2,49.7,41.5,46.2,48.2,43.2,51.8,53.3,
      53.3,47.2,56.9,47.8,48.7,53.7,60.6,56.7)
time=c(918, 805, 892, 962, 968, 907, 770, 743,
      1045, 810, 927, 813, 858, 860, 760, 747,
       743, 803, 683, 844, 755, 700, 748, 775)
plot(oxy, time)
```



- For individual i, let Y_i be the time to run 2 miles, and x_i be the maximum volume of oxygen uptake, $i = 1, ..., 24$.
- A possible model is

$$Y_i = a + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, 24,$$

where ε_i are independent random variables with variance σ^2, and a and b are constants.

Matrix formulation

The linear model may be written in matrix form. Let

$$\underset{n\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ . \\ . \\ Y_n \end{pmatrix}, \quad \underset{n\times p}{X} = \begin{pmatrix} x_{11} & . & . & x_{1p} \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & . & . & x_{np} \end{pmatrix}, \quad \underset{p\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix}, \quad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{pmatrix},$$

Then from (1),

$$\begin{aligned} \mathbf{Y} &= X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \mathbb{E}(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ \mathrm{cov}(\mathbf{Y}) &= \sigma^2 I \end{aligned} \tag{2}$$

We assume throughout that X has full rank p.

We also assume the error variance is the same for each observation: this is the *homoscedastic* case (as opposed to *heteroscedastic*).

Example 12.1 continued

- Recall $Y_i = a + bx_i + \varepsilon_i, \quad i = 1, .., 24.$
- In matrix form:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ . \\ . \\ Y_{24} \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_1 \\ . & . \\ . & . \\ 1 & x_{24} \end{pmatrix}, \; \boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}, \; \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_{24} \end{pmatrix},$$

- Then

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Least squares estimation

- In a linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the *least squares estimator* $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimises

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \end{aligned}$$

- So

$$\left.\frac{\partial S}{\partial \beta_k}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0, \;\; k = 1, .., p.$$

- So $\quad -2\sum_{i=1}^{n} x_{ik}(Y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j) = 0, \;\; k = 1, .., p.$
- i.e. $\quad \sum_{i=1}^{n} x_{ik} \sum_{j=1}^{p} x_{ij}\hat{\beta}_j = \sum_{i=1}^{n} x_{ik} Y_i, \;\; k = 1, .., p.$
- In matrix form,

$$X^T X\hat{\boldsymbol{\beta}} = X^T\mathbf{Y} \tag{3}$$

the *least squares equation*.

- Recall we assume X is of full rank p.
- This implies
$$\mathbf{t}^T X^T X \mathbf{t} = (X\mathbf{t})^T (X\mathbf{t}) = \|X\mathbf{t}\|^2 > 0$$
for $\mathbf{t} \neq \mathbf{0}$ in $\mathbb{R}^p$.
- i.e. $X^T X$ is *positive definite*, and hence has an inverse.
- Hence
$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} \tag{4}$$
which is linear in the Y_i's.
- We also have that
$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1} X^T \mathbb{E}(\mathbf{Y}) = (X^T X)^{-1} X^T X \boldsymbol{\beta} = \boldsymbol{\beta}$$
so $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.
- And
$$\text{cov}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1} X^T \text{cov}(\mathbf{Y}) X (X^T X)^{-1} = (X^T X)^{-1} \sigma^2 \tag{5}$$
since $\text{cov}(\mathbf{Y}) = \sigma^2 I$.

Simple linear regression using standardised x's

- The model

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

can be reparametrised to

$$Y_i = a' + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $\bar{x} = \sum x_i/n$ and $a' = a + b\bar{x}$.

- Since $\sum_i (x_i - \bar{x}) = 0$, this leads to simplified calculations.

- In matrix form, $X = \begin{pmatrix} 1 & (x_1 - \bar{x}) \\ . & . \\ 1 & (x_{24} - \bar{x}) \end{pmatrix}$, so that $X^T X = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$, where $S_{xx} = \sum_i (x_i - \bar{x})^2$.
- Hence

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix},$$

so that

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y} = \begin{pmatrix} \bar{Y} & 0 \\ 0 & \frac{S_{xY}}{S_{xx}} \end{pmatrix},$$

where $S_{xY} = \sum_i Y_i (x_i - \bar{x})$.

- We note that the estimated intercept is $\hat{a}' = \bar{y}$, and the estimated gradient $\hat{b}$ is

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2)}} \times \sqrt{\frac{S_{yy}}{S_{xx}}}$$
$$= r \times \sqrt{\frac{S_{yy}}{S_{xx}}}$$

- Thus the gradient is the *Pearson product-moment correlation coefficient* r, times the ratio of the empirical standard deviations of the y's and x's.

  (Note this gradient is the same whether the x's are standardised to have mean 0 or not.)
- From (5), $\text{cov}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1}\sigma^2$, and so

$$\text{var}(\hat{a}') = \text{var}(\bar{Y}) = \frac{\sigma^2}{n}; \qquad \text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}};$$

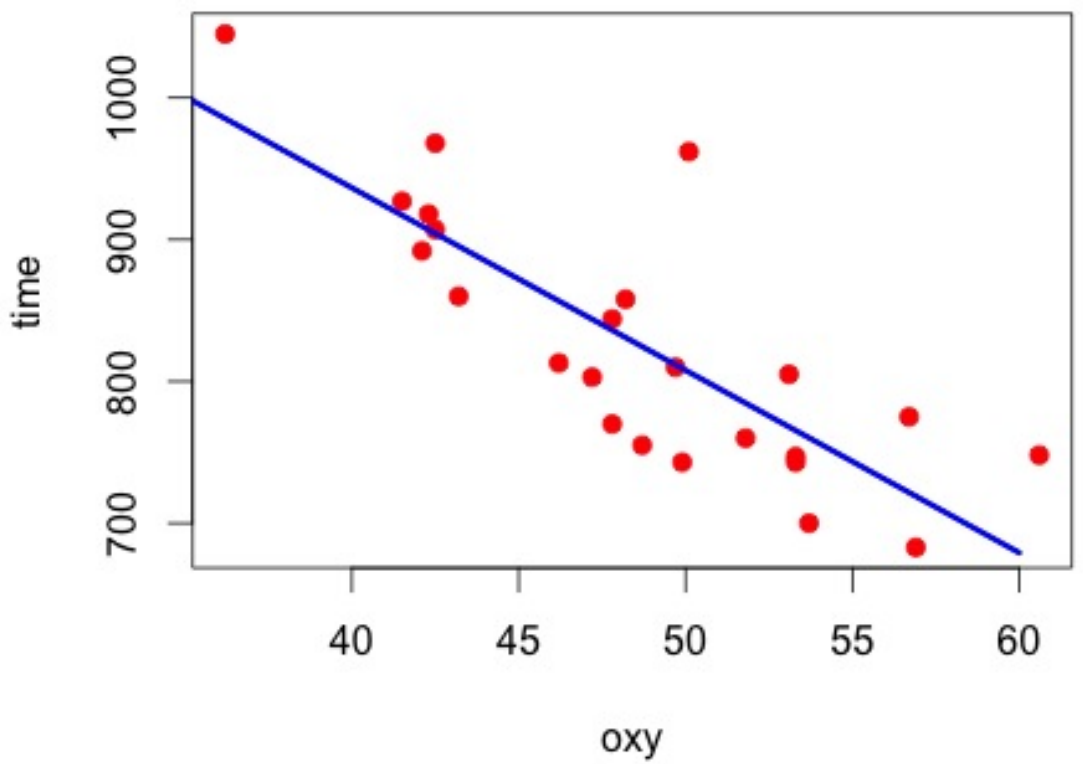
- These estimators are uncorrelated.

All these results are obtained without any explicit distributional assumptions.

Example 12.1 continued

$n = 24, \hat{a}' = \bar{y} = 826.5.$

$S_{xx} = 783.5 = 28.0^2, S_{xy} = -10077, S_{yy} = 444^2, r = -0.81, \hat{b} = -12.9.$



Line goes through $(\bar{x}, \bar{y})$.

'Gauss Markov' theorem

Theorem 12.2

In the full rank linear model, let $\hat{\boldsymbol{\beta}}$ be the least squares estimator of $\boldsymbol{\beta}$ and let $\boldsymbol{\beta}^*$ be any other unbiassed estimator for $\boldsymbol{\beta}$ which is linear in the Y_i's.
Then $var(\mathbf{t}^T\hat{\boldsymbol{\beta}}) \leq var(\mathbf{t}^T\boldsymbol{\beta}^*)$ for all $\mathbf{t} \in \mathbb{R}^p$.
We say that $\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator of $\boldsymbol{\beta}$ (BLUE).

Proof:

- Since $\boldsymbol{\beta}^*$ is linear in the Y_i's, $\boldsymbol{\beta}^* = A\mathbf{Y}$ for some $\underset{p\times n}{A}$.
- Since $\boldsymbol{\beta}^*$ is unbiased, we have that $\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{\beta}^*) = AX\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, and so $AX = I_p$.
- Now

$$\begin{aligned}
\text{cov}(\boldsymbol{\beta}^*) &= \mathbb{E}\left(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})^T\right) \\
&= \mathbb{E}\left(AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta})(AX\boldsymbol{\beta} + A\boldsymbol{\varepsilon} - \boldsymbol{\beta})^T\right) \\
&= \mathbb{E}\left(A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T A^T\right) \quad \text{since } AX\boldsymbol{\beta} = \boldsymbol{\beta} \\
&= A(\sigma^2 I)A^T = \sigma^2 AA^T
\end{aligned}$$

- Now $\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} = (A - (X^T X)^{-1} X^T)\mathbf{Y} = \underset{p \times n}{B}\,\mathbf{Y}$, say.
- And $BX = AX - (X^T X)^{-1} X^T X = I_p - I_p = 0$.
- So

$$\begin{aligned} \text{cov}(\boldsymbol{\beta}^*) &= \sigma^2 (B + (X^T X)^{-1} X^T)(B + (X^T X)^{-1} X^T)^T \\ &= \sigma^2 (BB^T + (X^T X)^{-1}) \\ &= \sigma^2 BB^T + \text{cov}(\hat{\boldsymbol{\beta}}) \end{aligned}$$

- So for $\mathbf{t} \in \mathbb{R}^p$,

$$\begin{aligned} \text{var}(\mathbf{t}^T \boldsymbol{\beta}^*) &= \mathbf{t}^T \text{cov}(\boldsymbol{\beta}^*)\mathbf{t} = \mathbf{t}^T \text{cov}(\boldsymbol{\beta})\mathbf{t} + \mathbf{t}^T BB^T \mathbf{t}\ \sigma^2 \\ &= \text{var}(\mathbf{t}^T \hat{\boldsymbol{\beta}}) + \sigma^2 \| B^T \mathbf{t} \|^2 \\ &\geq \text{var}(\mathbf{t}^T \hat{\boldsymbol{\beta}}). \end{aligned}$$

- Taking $\mathbf{t} = (0, .., 1, 0, .., 0)^T$ with a 1 in the ith position, gives

$$\text{var}(\hat{\beta}_i) \leq \text{var}(\beta_i^*).$$

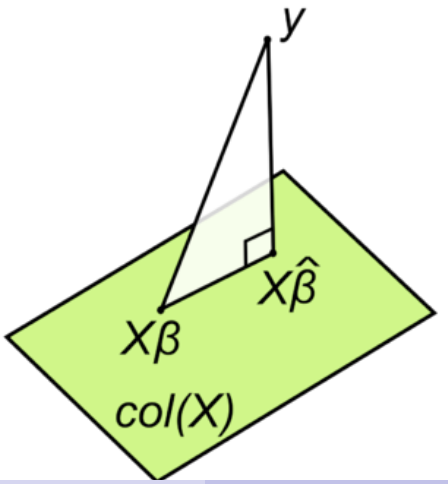
□

Fitted values and residuals

Definition 12.3

- $\hat{\mathbf{Y}} = X\hat{\beta}$ is the vector of *fitted values*.
- $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the vector of *residuals*.
- The *residual sum of squares* is $\text{RSS} = \|\mathbf{R}\|^2 = \mathbf{R}^T\mathbf{R} = (\mathbf{Y} - X\hat{\beta})^T(\mathbf{Y} - X\hat{\beta})$

- Note $X^T\mathbf{R} = X^T(\mathbf{Y} - \hat{\mathbf{Y}}) = X^T\mathbf{Y} - X^TX\hat{\beta} = 0$ by (3).
- So $\mathbf{R}$ is orthogonal to the column space of X.
- Write $\hat{\mathbf{Y}} = X\hat{\beta} = X(X^TX)^{-1}X^T\mathbf{Y} = P\mathbf{Y}$, where $P = X(X^TX)^{-1}X^T$.
- P represents an orthogonal projection of $\mathbb{R}^n$ onto the space spanned by columns of X. We have $P^2 = P$ (P is idempotent) and $P^T = P$ (symmetric).

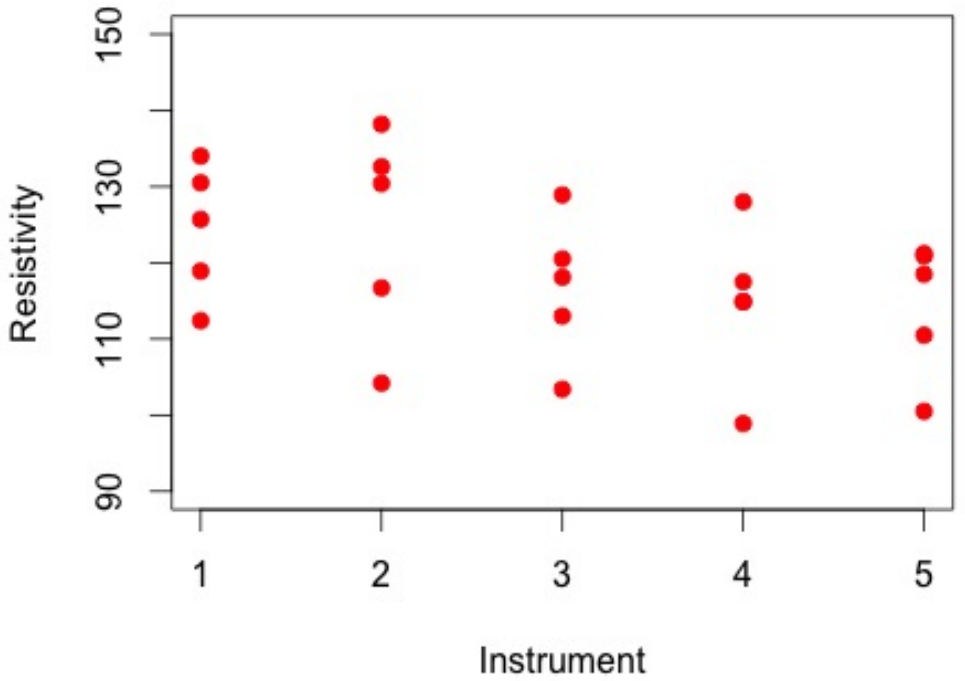


Lecture 13. Linear models with normal assumptions

One way analysis of variance

Example 13.1

Resistivity of silicon wafers was measured by five instruments.
Five wafers were measured by each instrument (25 wafers in all).



```
y=c(130.5,112.4,118.9,125.7,134.0,
    130.4,138.2,116.7,132.6,104.2,
    113.0,120.5,128.9,103.4,118.1,
    128.0,117.5,114.9,114.9, 98.9,
    121.2,110.5,118.5,100.5,120.9)
```

Let $Y_{i,j}$ be the resistivity of the jth wafer measured by instrument i, where $i, j = 1, .., 5$.

A possible model is, for $i, j = 1, .., 5$.

$$Y_{i,j} = \mu_i + \varepsilon_{i,j},$$

where $\varepsilon_{i,j}$ are independent $\mathrm{N}(0, \sigma^2)$ random variables, and the μ_i's are unknown constants.

This can be written in matrix form: Let

$$\underset{25\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_{1,1} \\ . \\ . \\ Y_{1,5} \\ Y_{2,1} \\ . \\ . \\ Y_{2,5} \\ . \\ . \\ Y_{5,1} \\ . \\ . \\ Y_{5,5} \end{pmatrix}, \quad \underset{25\times 5}{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \end{pmatrix}, \quad \underset{5\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}, \quad \underset{25\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_{1,1} \\ . \\ . \\ \varepsilon_{1,5} \\ \varepsilon_{2,1} \\ . \\ . \\ \varepsilon_{2,5} \\ . \\ . \\ \varepsilon_{5,1} \\ . \\ . \\ \varepsilon_{5,5} \end{pmatrix},$$

Then

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$$X^T X = \begin{pmatrix} 5 & 0 & \dots & 0 \\ 0 & 5 & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & .. & 5 \end{pmatrix}.$$

Hence

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{5} & 0 & \dots & 0 \\ 0 & \frac{1}{5} & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & .. & \frac{1}{5} \end{pmatrix},$$

so that

$$\hat{\boldsymbol{\mu}} = (X^T X)^{-1} X^T \mathbf{Y} = \begin{pmatrix} \overline{Y_{1.}} \\ .. \\ \overline{Y_{5.}} \end{pmatrix}$$

$\text{RSS} = \sum_{i=1}^{5} \sum_{j=1}^{5} (Y_{i,j} - \hat{\mu}_i)^2 = \sum_{i=1}^{5} \sum_{j=1}^{5} (Y_{i,j} - \overline{Y_{i.}})^2$ on $n - p = 25 - 5 = 20$ degrees of freedom.

For these data, $\tilde{\sigma} = \sqrt{\text{RSS}/(n-p)} = \sqrt{2170/20} = 10.4$.

Assuming normality

- We now make a Normal assumption

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I), \qquad \text{rank } (X) = p(< n).$$

- This is a special case of the linear model of Lecture 12, so all results hold.
- Since $\mathbf{Y} \sim \mathrm{N}_n(X\boldsymbol{\beta}, \sigma^2 I)$, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}S(\boldsymbol{\beta}),$$

  where $S(\boldsymbol{\beta}) = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})$.
- Maximising ℓ wrt $\boldsymbol{\beta}$ is equivalent to minimising $S(\boldsymbol{\beta})$, so MLE is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y},$$

  the same as for least squares.

- For the MLE of σ^2, we require

$$\begin{aligned} \left.\frac{\partial \ell}{\partial \sigma^2}\right|_{\hat{\boldsymbol{\beta}},\hat{\sigma}^2} &= 0, \\ \text{i.e.} \qquad -\frac{n}{2\hat{\sigma}^2} + \frac{S(\hat{\boldsymbol{\beta}})}{2\hat{\sigma}^4} &= 0 \end{aligned}$$

- . So

$$\hat{\sigma}^2 = \frac{1}{n} S(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) = \frac{1}{n}\text{RSS},$$

  where RSS is 'residual sum of squares' - see last lecture.
- See example sheet for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ for simple linear regression and one-way analysis of variance.

Lemma 13.2

(i) *If* $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I)$, *and* A *is* $n \times n$, *symmetric, idempotent with rank* r, *then* $\mathbf{Z}^T A \mathbf{Z} \sim \sigma^2 \chi^2_r$.

(ii) *For a symmetric idempotent matrix* A, *rank*(A) = *trace*(A)

Proof:

- (i) $A^2 = A$ since idempotent, and so eigenvalues of A are $\lambda_i \in \{0, 1\}, i = 1, .., n, \qquad [\lambda_i \mathbf{x} = A\mathbf{x} = A^2\mathbf{x} = \lambda_i^2 \mathbf{x}]$.
- A is also symmetric, and so there exists an orthogonal Q such that

$$Q^T A Q = \text{diag}\,(\lambda_1, .., \lambda_n) = \text{diag}\,(\underbrace{1, .., 1}_{r}, \underbrace{0, ..., 0}_{n-r}) = \Lambda \text{ (say)}.$$

- Let $\mathbf{W} = Q^T \mathbf{Z}$, and so $\mathbf{Z} = Q\mathbf{W}$. Then $\mathbf{W} \sim \text{N}_n(\mathbf{0}, \sigma^2 I)$ by Proposition 11.1(i). (since $\text{cov}(\mathbf{W}) = Q^T \sigma^2 I Q = \sigma^2 I$).
- Then

$$\mathbf{Z}^T A \mathbf{Z} = \mathbf{W}^T Q^T A Q \mathbf{W} = \mathbf{W}^T \Lambda \mathbf{W} = \sum_{i=1}^{r} w_i^2 \sim \chi^2_r,$$

from the definition of χ^2.

- (ii)

$$\begin{aligned}
\text{rank}\,(A) &= \text{rank}\,(Q^T AQ) && \text{if } Q \text{ orthogonal} \\
&= \text{rank}\,(\Lambda) \\
&= \text{trace}\,(\Lambda) \\
&= \text{trace}\,(Q^T AQ) \\
&= \text{trace}\,(AQ^T Q) && \text{since } \text{tr}(AB) = \text{tr}(BA) \\
&= \text{trace}\,(A)
\end{aligned}$$

□

Theorem 13.3

For the normal linear model $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$,

(i) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$.

(ii) $RSS \sim \sigma^2\chi^2_{n-p}$, *and so* $\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$.

(iii) $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}^2$ *are independent.*

(also may be referred to as Theorem 3.6 in Example Sheets)

Proof:

- (i) $\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{Y}$, say $C\mathbf{Y}$.

  Then from Proposition 11.1(i), $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})$.

- (ii) We can apply Lemma 13.2(i) with $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\beta} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I_n)$ and $A = (I_n - P)$, where $P = X(X^TX)^{-1}X^T$ is the projection matrix covered after Definition 12.3.
- (P is also known as the 'hat' matrix since it projects from the observation $\mathbf{Y}$ onto the fitted values $\hat{\mathbf{Y}}$.)
- P is symmetric and idempotent, so $I_n - P$ is also symmetric and idempotent (check).
- By Lemma 13.2(ii),

$$\mathrm{rank}(P) = \mathrm{trace}(P) = \mathrm{trace}(X(X^TX)^{-1}X^T) = \mathrm{trace}((X^TX)^{-1}X^TX) = p,$$

so $\mathrm{rank}(I_n - P) = \mathrm{trace}(I_n - P) = n - p$.

- Note that $(I_n - P)X = 0$ (check) so that

$$\mathbf{Z}^TA\mathbf{Z} = (\mathbf{Y} - X\boldsymbol{\beta})^T(I_n - P)(\mathbf{Y} - X\boldsymbol{\beta}) = \mathbf{Y}^T(I_n - P)\mathbf{Y} \text{ since } (I_n - P)X = 0.$$

- We know $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = (I_n - P)\mathbf{Y}$ and $(I_n - P)$ is symmetric and idempotent, and so

$$\mathrm{RSS} = \mathbf{R}^T\mathbf{R} = \mathbf{Y}^T(I_n - P)\mathbf{Y} \qquad (= \mathbf{Z}^TA\mathbf{Z}).$$

- Hence by Lemma 13.2(i), $\mathrm{RSS} \sim \sigma^2\chi^2_{n-p}$ and $\hat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n} \sim \dfrac{\sigma^2}{n}\chi^2_{n-p}$.

- (iii) Let $\underset{(p+n)\times 1}{V} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{R} \end{pmatrix} = D\mathbf{Y}$, where $D = \begin{pmatrix} C \\ I_n - P \end{pmatrix}$ is a $(p+n) \times n$ matrix.
- By Proposition 11.1(i), V is multivariate normal with

$$\begin{aligned} \text{cov}(V) = \sigma^2 DD^T &= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P)^T \\ (I_n - P)C^T & (I_n - P)(I_n - P)^T \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P) \\ (I_n - P)C^T & (I_n - P) \end{pmatrix}. \end{aligned}$$

- We have $C(I_n - P) = 0$ (check) $[(X^TX)^{-1}X^T(I_n - P) = 0$ because $(I_n - P)X = 0]$.
- Hence $\hat{\boldsymbol{\beta}}$ and $\mathbf{R}$ are independent by Proposition 11.2(ii).
- Hence $\hat{\boldsymbol{\beta}}$ and RSS$=\mathbf{R}^T\mathbf{R}$ are independent, and so $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. □.

From (ii), $\mathbb{E}(\text{RSS}) = \sigma^2(n-p)$, and so $\tilde{\sigma}^2 = \frac{\text{RSS}}{n-p}$ is an unbiased estimator of σ^2. $\tilde{\sigma}$ is often known as the *residual standard error on $n-p$ degrees of freedom*.

Example 12.1 continued

The RSS = residual sum of squares is the sum of the squared vertical distances from the data-points to the fitted straight line.

$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{a}' - \hat{b}(x_i - \bar{x})^2 = 67968.$

So the estimate of

$$\tilde{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{67968}{(24-2)} = 3089.$$

Residual standard error is $\tilde{\sigma} = \sqrt{3089} = 55.6$ on 22 degrees of freedom.

The F distribution

- Suppose that U and V are independent with $U \sim \chi^2_m$ and $V \sim \chi^2_n$.
- Then $X = (U/m)/(V/n)$ is said to have an F **distribution** on m and n degrees of freedom.
- We write $X \sim F_{m,n}$.
- Note that, if $X \sim F_{m,n}$ then $1/X \sim F_{n,m}$.
- Let $F_{m,n}(\alpha)$ be the upper $100\alpha\%$ point for the $F_{m,n}$-distribution so that if $X \sim F_{m,n}$ then $\mathbb{P}(X > F_{m,n}(\alpha)) = \alpha$. These are tabulated.
- If we need, say, the lower 5% point of $F_{m,n}$, then find the upper 5% point x of $F_{n,m}$ and use $\mathbb{P}(F_{m,n} < 1/x) = \mathbb{P}(F_{n,m} > x)$.
- Note further that it is immediate from the definitions of t_n and $F_{1,n}$ that if $Y \sim t_n$ then $Y^2 \sim F_{1,n}$, since ratio of independent χ^2_1 and χ^2_n variables.

Lecture 14. Applications of the distribution theory

Inference for $\boldsymbol{\beta}$

We know that $\hat{\boldsymbol{\beta}} \sim \mathsf{N}_p(\boldsymbol{\beta}, \sigma^2(X^T X)^{-1})$, and so

$$\hat{\beta}_j \sim \mathsf{N}(\beta_j, \sigma^2(X^T X)^{-1}_{jj}).$$

The *standard error* of $\hat{\boldsymbol{\beta}}_j$ is

$$\text{s.e.}(\hat{\beta}_j) = \sqrt{\tilde{\sigma}^2(X^T X)^{-1}_{jj}},$$

where $\tilde{\sigma}^2 = \text{RSS}/(n-p)$, as in Theorem 13.3.

Then

$$\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\tilde{\sigma}^2(X^T X)^{-1}_{jj}}} = \frac{(\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2(X^T X)^{-1}_{jj}}}{\sqrt{\text{RSS}/((n-p)\sigma^2)}}.$$

The numerator is a standard normal $\mathsf{N}(0,1)$, the denominator is an independent $\sqrt{\chi^2_{n-p}/(n-p)}$, and so $\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} \sim \mathsf{t}_{n-p}$.

So a $100(1-\alpha)\%$ CI for β_j has endpoints $\hat{\beta}_j \pm \text{s.e.}(\hat{\beta}_j)\ t_{n-p}(\frac{\alpha}{2})$.

To test $H_0 : \beta_j = 0$, use the fact that, under H_0, $\dfrac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)} \sim \text{t}_{n-p}$.

Simple linear regression

We assume that

$$Y_i = a' + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\bar{x} = \sum x_i/n$, and $\varepsilon_i, i = 1, ..., n$ are iid $\mathrm{N}(0, \sigma^2)$.

Then from Lecture 12 and Theorem 13.3 we have that

$$\hat{a}' = \overline{Y} \sim \mathrm{N}\left(a', \frac{\sigma^2}{n}\right), \qquad \hat{b} = \frac{S_{xY}}{S_{xx}} \sim \mathrm{N}\left(b, \frac{\sigma^2}{S_{xx}}\right),$$

$$\hat{Y}_i = \hat{a}' + \hat{b}(x_i - \bar{x}), \qquad \mathrm{RSS} = \sum_i (Y_i - \hat{Y}_i)^2 \sim \sigma^2 \chi^2_{n-2},$$

and $(\hat{a}', \hat{b})$ and $\hat{\sigma}^2 = \mathrm{RSS}/n$ are independent.

Example 12.1 continued

- We have seen that $\tilde{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{67968}{(24-2)} = 3089 = 55.6^2$.
- So the standard error of $\hat{b}$ is

$$\text{s.e.}(\hat{b}) = \sqrt{\tilde{\sigma}^2 (X^T X)^{-1}_{22}}, = \sqrt{\frac{3089}{S_{xx}}} = \frac{55.6}{28.0} = 1.99.$$

- So a 95% interval for b has endpoints
$\hat{b} \pm \text{s.e.}(\hat{b}) \times t_{n-p}(0.025) = -12.9 \pm 1.99 * t_{22}(0.025) = (-17.0, -8.8)$,
where $t_{22}(0.025) = 2.07$.
- This does not contain 0. Hence if carry out a size 0.05 test of $H_0 : b = 0$ *vs* $H_1 : b \neq 0$, the test statistic would be $\frac{\hat{b}}{\text{s.e.}(\hat{b})} = \frac{-12.9}{1.99} = -6.48$, and we would reject H_0 since this is less than $-t_{22}(0.025) = -2.07$.
-

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  826.500     11.346  72.846  < 2e-16 ***
oxy.s        -12.869      1.986  -6.479 1.62e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 55.58 on 22 degrees of freedom
```

Expected response at $\mathbf{x}^*$

- Let $\mathbf{x}^*$ be a new vector of values for the explanatory variables
- The expected response at $\mathbf{x}^*$ is $\mathbb{E}(Y|\mathbf{x}^*) = \mathbf{x}^{*T}\boldsymbol{\beta}$.
- We estimate this by $\mathbf{x}^{*T}\hat{\boldsymbol{\beta}}$.
- By Theorem 13.3 and Proposition 11.1(i),

$$\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{N}(0, \sigma^2\, \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*).$$

- Let $\tau^2 = \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*$.
- Then

$$\frac{\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\tilde{\sigma}\tau} \sim t_{n-p}.$$

- A $100(1-\alpha)\%$ confidence interval for the expected response $\mathbf{x}^{*T}\boldsymbol{\beta}$ has endpoints

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \pm \tilde{\sigma}\tau t_{n-p}(\tfrac{\alpha}{2}).$$

Example 12.1 continued

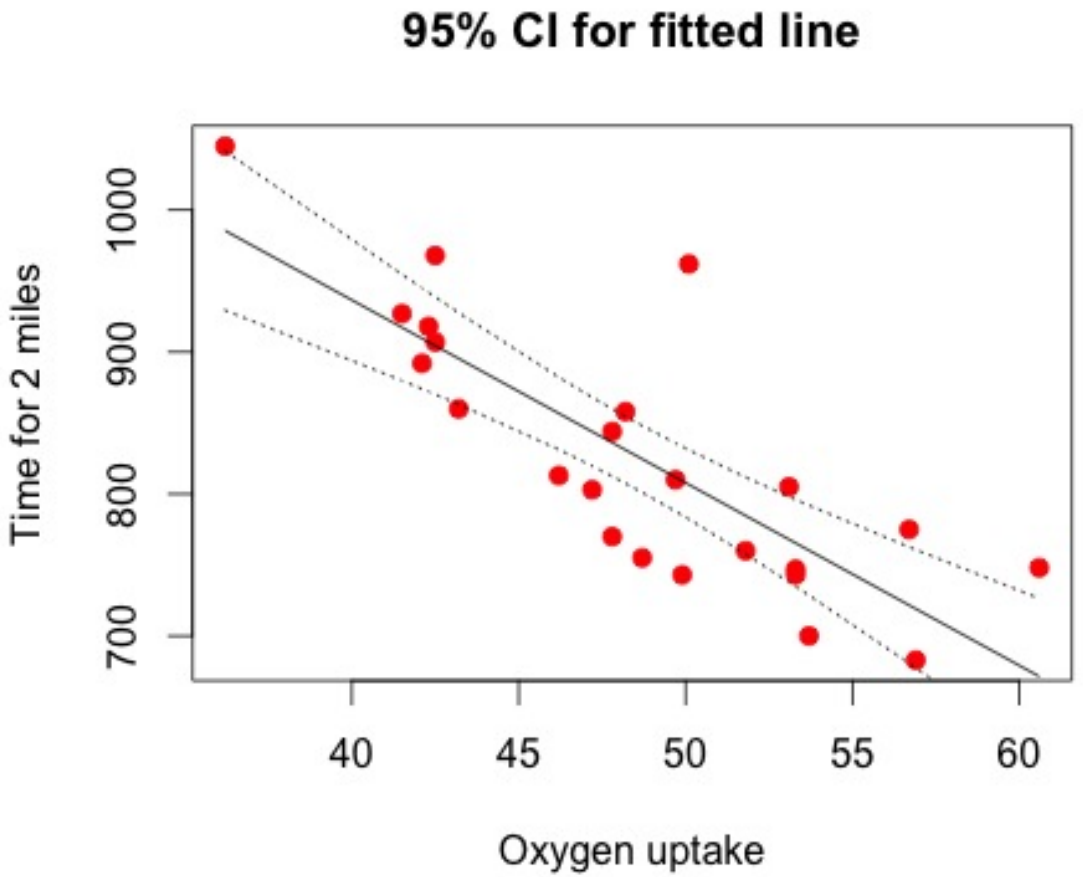
- Suppose we wish to estimate the time to run 2 miles for a man with an oxygen take-up measurement of 50.
- Here $\mathbf{x}^{*T} = (1, (50 - \bar{x}))$, where $\bar{x} = 48.6$.
- The estimated expected response at $\mathbf{x}^{*T}$ is

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} = \hat{a}' + (50 - 48.6) \times \hat{b} = 826.5 - 1.4 \times 12.9 = 808.5.$$

- We find $\tau^2 = \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^* = \frac{1}{n} + \frac{\mathbf{x}^{*2}}{S_{xx}} = \frac{1}{24} + \frac{1.4^2}{783.5} = 0.044 = 0.21^2$.
- So a 95% CI for $\mathbb{E}(Y|\mathbf{x}^* = 50 - \bar{x})$ is

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \pm \tilde{\sigma}\tau t_{n-p}(\tfrac{\alpha}{2}) = 808.5 \pm 55.6 \times 0.21 \times 2.07 = (783.6, 832.2).$$

```
oxy.s = oxy - mean(oxy)
fit=lm(time~ oxy.s  )
pred=predict.lm(fit,  interval="confidence")
plot(oxy, time,col="red", pch=19, xlab="Oxygen uptake",ylab="Time for 2 miles", mai
lines(oxy, pred[, "fit"])
lines(oxy, pred[, "lwr"], lty = "dotted")
lines(oxy, pred[, "upr"], lty = "dotted")
```



Predicted response at $x*$

- The response at $\mathbf{x}^*$ is $Y^* = \mathbf{x}^*\boldsymbol{\beta} + \varepsilon^*$, where $\varepsilon^* \sim \mathrm{N}(0, \sigma^2)$, and Y^* is independent of $Y_1, .., Y_n$.
- We predict $\hat{Y}^*$ by $\mathbf{x}^{*T}\hat{\boldsymbol{\beta}}$.
- A $100(1-\alpha)\%$ *prediction interval* for Y^* is an interval $I(\mathbf{Y})$ such that $\mathbb{P}(Y^* \in I(\mathbf{Y})) = 1 - \alpha$, where the probability is over the joint distribution of $(Y^*, Y_1, ..., Y_n)$.
- Observe that $\hat{Y}^* - Y^* = \mathbf{x}^{*T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon^*$.
- So $\mathbb{E}(\hat{Y}^* - Y^*) = \mathbf{x}^{*T}(\boldsymbol{\beta} - \boldsymbol{\beta}) = 0$.
- And

$$\begin{aligned} \mathrm{var}(\hat{Y}^* - Y^*) &= \mathrm{var}(\mathbf{x}^{*T}(\hat{\boldsymbol{\beta}})) + \mathrm{var}(\varepsilon^*) \\ &= \sigma^2 \mathbf{x}^{*T}(X^T X)^{-1}\mathbf{x}^* + \sigma^2 \\ &= \sigma^2(\tau^2 + 1) \end{aligned}$$

- So

$$\hat{Y}^* - Y^* \sim \mathrm{N}(0, \sigma^2(\tau^2 + 1)).$$

- We therefore find that

$$\frac{\hat{Y}^* - Y^*}{\tilde{\sigma}\sqrt{(\tau^2+1)}} \sim t_{n-p}.$$

- So the interval with endpoints

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \pm \tilde{\sigma}\sqrt{(\tau^2+1)}\; t_{n-p}(\tfrac{\alpha}{2}).$$

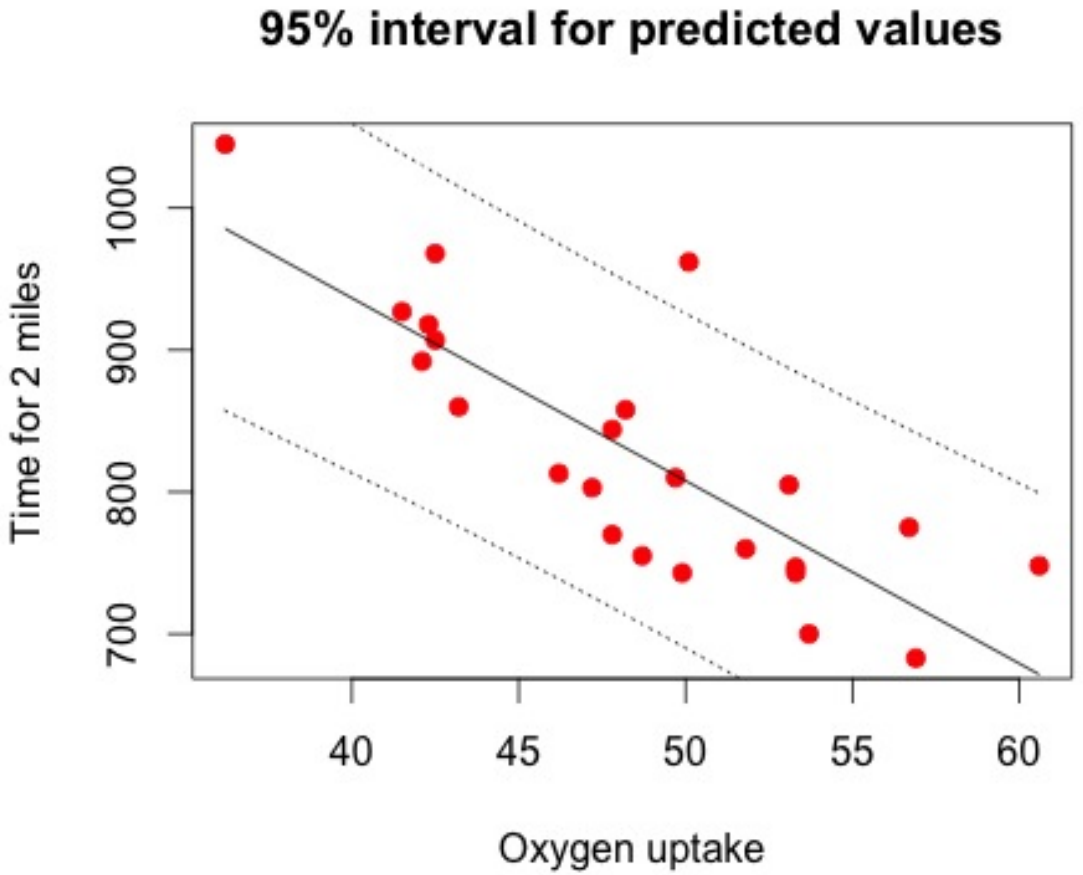
  is a 95% prediction interval for Y^*.

Example 12.1 continued

A 95% prediction interval for Y^* at $\mathbf{x}^{*T} = (1, (50 - \bar{x}))$ is

$$\mathbf{x}^{*T}\hat{\boldsymbol{\beta}} \pm \tilde{\sigma}\sqrt{(\tau^2+1)}\; t_{n-p}(\tfrac{\alpha}{2}) = 808.5 \pm 55.6 \times 1.02 \times 2.07 = (691.1, 925.8).$$

```
pred=predict.lm(fit,  interval="prediction")
```



Note wide prediction intervals for individual points, with the width of the interval dominated by the residual error term $\tilde{\sigma}$ rather than the uncertainty about the fitted line.

Example 13.1 continued. One-way analysis of variance

- Suppose we wish to estimate the expected resistivity of a new wafer in the first instrument.
- Here $\mathbf{x}^{*T} = (1, 0, .., 0)$.
- The estimated expected response at $\mathbf{x}^{*T}$ is

$$\mathbf{x}^{*T}\hat{\boldsymbol{\mu}} = \hat{\mu}_1 = \overline{Y}_{1.} = 124.3$$

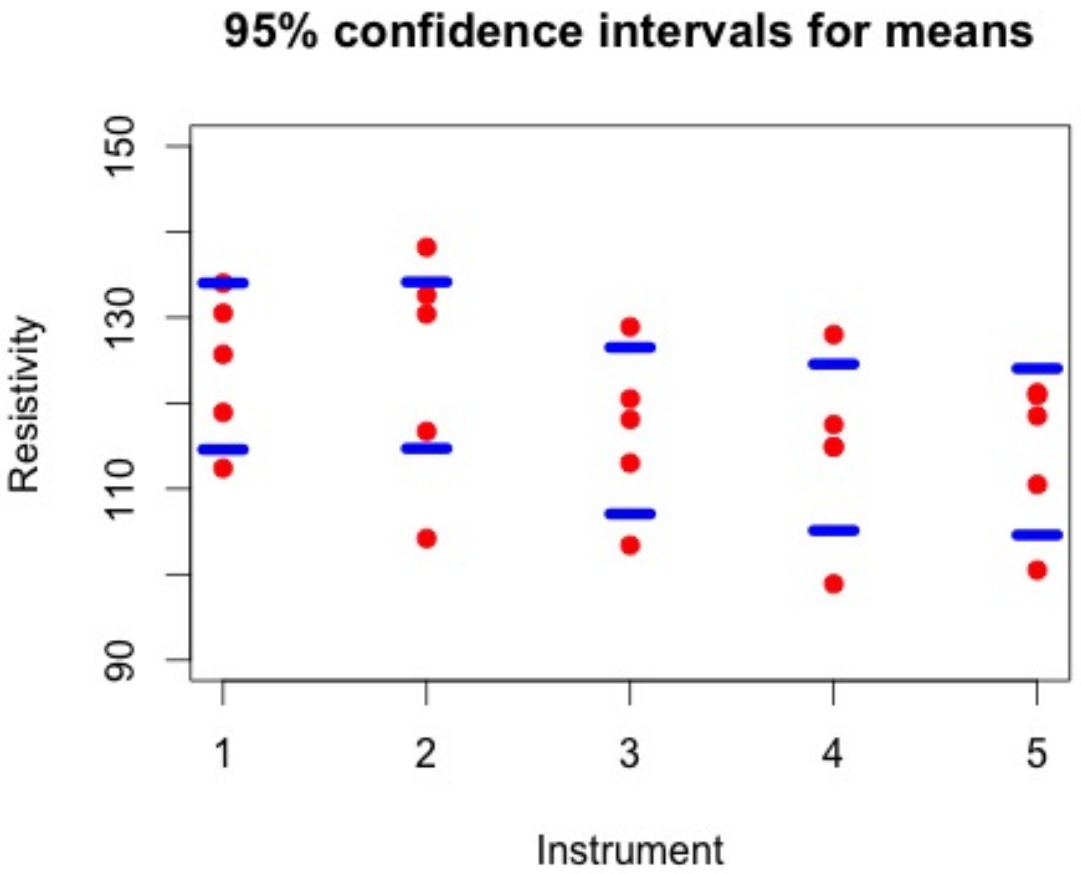
- We find $\tau^2 = \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^* = \frac{1}{5}$.
- So a 95% CI for $\mathbb{E}(Y_{1*})$ is $\mathbf{x}^{*T}\hat{\boldsymbol{\mu}} \pm \tilde{\sigma}\,\tau\,t_{n-p}(\frac{\alpha}{2})$

$$= 124.3 \pm 10.4/\sqrt{5} \times 2.09 = 124.3 \pm 4.66 \times 2.09 = (114.6, 134.0).$$

- Note that we are using an estimate of σ obtained from all five instruments. If we had only used the data from the first instrument, σ would be estimated as $\tilde{\sigma}_1 = \sqrt{\sum_{j=1}^{5}(y_{1,j} - \overline{y}_{1.})^2/(5-1)} = 8.74$.
- The observed 95% confidence interval for μ_1 would have been

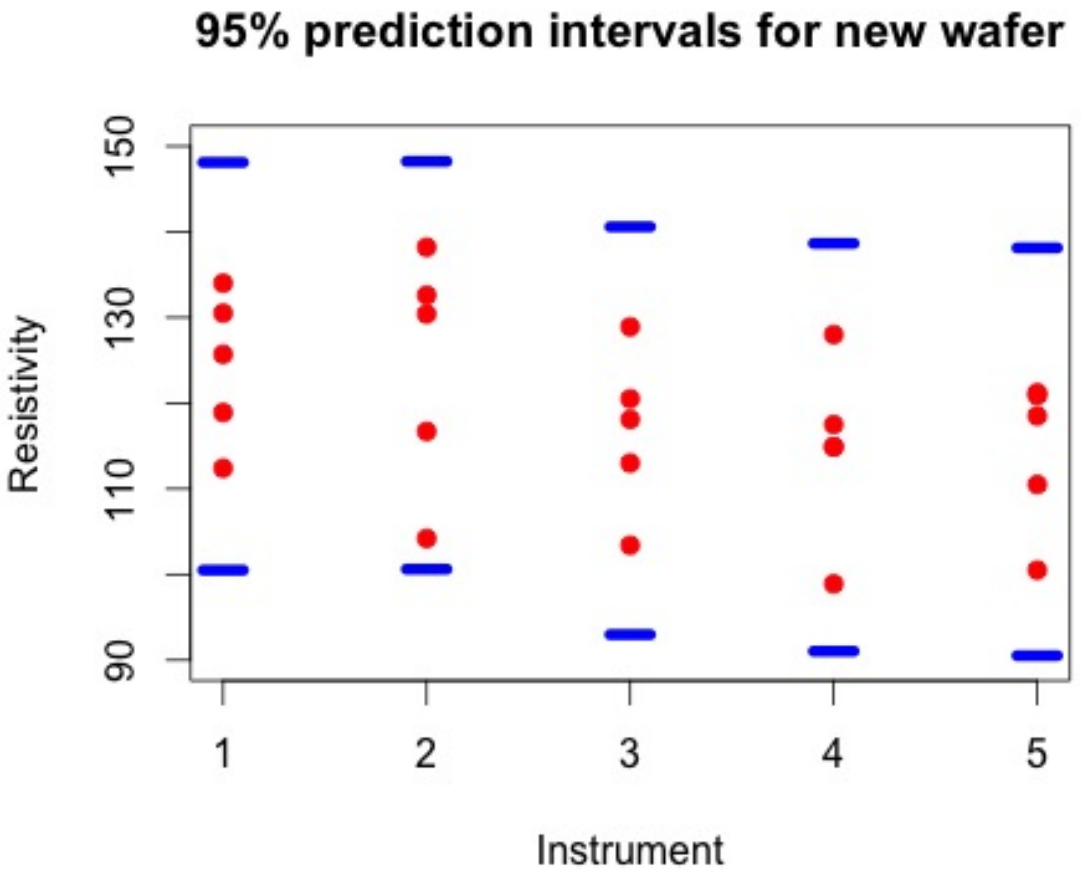
$$\overline{y_{1.}} \pm \frac{\tilde{\sigma}_1}{\sqrt{5}}\ t_4(\tfrac{\alpha}{2}) = 124.3 \pm 3.91 \times 2.78 = (113.5, 135.1).$$

- The 'pooled' analysis gives a slightly narrower interval.


95% confidence intervals for means
Resistivity
150
130
110
90
1
2
3
4
5
Instrument

A 95% prediction interval for Y_{1*} at $\mathbf{x}^{*T} = (1, 0, ..., 0)$ is

$$\mathbf{x}^{*T}\hat{\boldsymbol{\mu}} \pm \tilde{\sigma}\sqrt{(\tau^2 + 1)}\; t_{n-p}(\tfrac{\alpha}{2}) = 124.3 \pm 10.42 \times 1.1 \times 2.07 = (100.5, 148.1).$$



Lecture 15. Hypothesis testing in the linear model

Preliminary lemma

Lemma 15.1

Suppose $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I_n)$ *and* A_1 *and* A_2 *and symmetric, idempotent* $n \times n$ *matrices with* $A_1 A_2 = 0$. *Then* $\mathbf{Z}^T A_1 \mathbf{Z}$ *and* $\mathbf{Z}^T A_2 \mathbf{Z}$ *are independent.*

Proof:

- Let $\mathbf{W}_i = A_i \mathbf{Z}, i = 1, 2$ and $\underset{2n \times 1}{\mathbf{W}} = \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix} = A\mathbf{Z}$, where $\underset{2n \times n}{A} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$.
- By Proposition 11.1(i), $\mathbf{W} \sim \mathrm{N}_{2n}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \sigma^2 \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}\right)$ check.
- So $\mathbf{W}_1$ and $\mathbf{W}_2$ are independent, which implies ${\mathbf{W}_1}^T \mathbf{W}_1 = \mathbf{Z}^T A_1 \mathbf{Z}$ and ${\mathbf{W}_2}^T \mathbf{W}_2 = \mathbf{Z}^T A_2 \mathbf{Z}$ are independent. $\square$.

Hypothesis testing

- Suppose $\underset{n\times p}{X} = (\underset{n\times p_0}{X_0} \quad \underset{n\times(p-p_0)}{X_1})$ and $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, where $\text{rank}(X) = p, \text{rank}(X_0) = p_0$.
- We want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.
- Under H_0, $\mathbf{Y} = X_0\beta_0 + \varepsilon$.
- Under H_0, MLEs of β_0 and σ^2 are

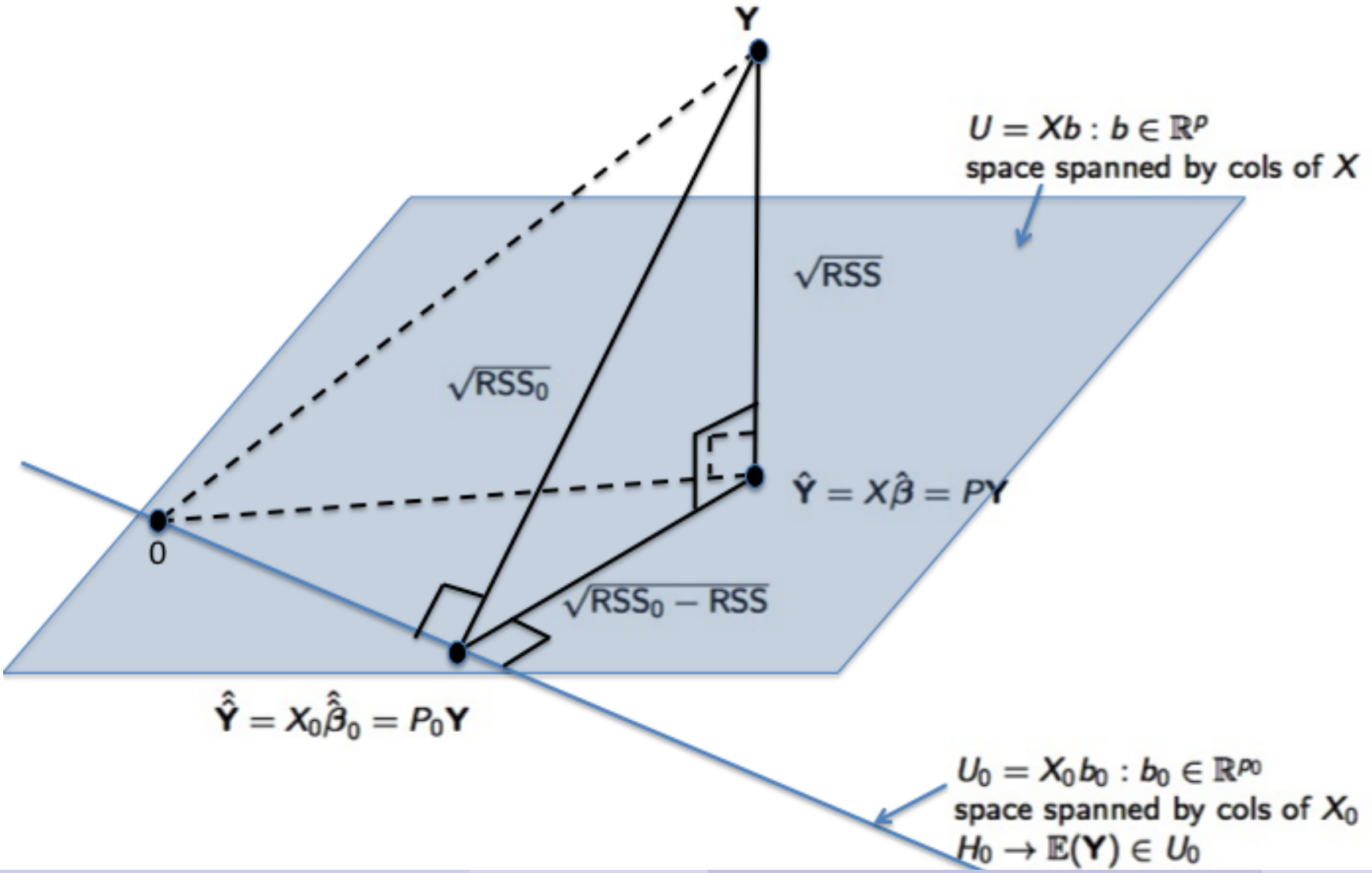
$$\begin{aligned} \hat{\hat{\beta}}_0 &= (X_0{}^T X_0)^{-1} X_0{}^T \mathbf{Y} \\ \hat{\hat{\sigma}}^2 &= \frac{\text{RSS}_0}{n} = \frac{1}{n}(\mathbf{Y} - X_0\hat{\hat{\beta}}_0)^T(\mathbf{Y} - X_0\hat{\hat{\beta}}_0) \end{aligned}$$

  and these are independent, by Theorem 13.3.
- So fitted values under H_0 are

$$\hat{\hat{\mathbf{Y}}} = X_0(X_0{}^T X_0)^{-1} X_0{}^T \mathbf{Y} = P_0\mathbf{Y},$$

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

Geometric interpretation



Generalised likelihood ratio test

- The generalised likelihood ratio test of H_0 against H_1 is

$$\begin{aligned}\Lambda_{\mathbf{Y}}(H_0, H_1) &= \frac{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma}^2}(\mathbf{Y}-X\hat{\beta})^T(\mathbf{Y}-X\hat{\beta})\right)}{\left(\frac{1}{\sqrt{2\pi\hat{\hat{\sigma}}^2}}\right)^n \exp\left(-\frac{1}{2\hat{\hat{\sigma}}^2}(\mathbf{Y}-X\hat{\hat{\beta}}_0)^T(\mathbf{Y}-X\hat{\hat{\beta}}_0)\right)} \\ &= \left(\frac{\hat{\hat{\sigma}}^2}{\hat{\sigma}^2}\right)^{\frac{n}{2}} = \left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}}\right)^{\frac{n}{2}} = \left(1+\frac{\mathrm{RSS}_0-\mathrm{RSS}}{\mathrm{RSS}}\right)^{\frac{n}{2}}\end{aligned}$$

- We reject H_0 when $2\log\Lambda$ is large, equivalently when $\frac{(\mathrm{RSS}_0-\mathrm{RSS})}{\mathrm{RSS}}$ is large.
- Using the results in Lecture 8, under H_0

$$2\log\Lambda = n\log\left(1+\frac{\mathrm{RSS}_0-\mathrm{RSS}}{\mathrm{RSS}}\right)$$

  is approximately a $\chi^2_{p_1-p_0}$ rv.
- But we can get an exact null distribution.

Null distribution of test statistic

- We have $\text{RSS} = \mathbf{Y}^T(I_n - P)\mathbf{Y}$ (see proof of Theorem 13.3 (ii)), and so

$$\text{RSS}_0 - \text{RSS} = \mathbf{Y}^T(I_n - P_0)\mathbf{Y} - \mathbf{Y}^T(I_n - P)\mathbf{Y} = \mathbf{Y}^T(P - P_0)\mathbf{Y}.$$

- Now $I_n - P$ and $P - P_0$ are symmetric and idempotent, and therefore $\text{rank}(I_n - P) = n - p$, and

$$\text{rank}(P - P_0) = \text{tr}(P - P_0) = \text{tr}(P) - \text{tr}(P_0) = \text{rank}(P) - \text{rank}(P_0) = p - p_0.$$

- Also

$$(I_n - P)(P - P_0) = (I_n - P)P - (I_n - P)P_0 = 0.$$

- Finally,

$$\begin{aligned} \mathbf{Y}^T(I_n - P)\mathbf{Y} &= (\mathbf{Y} - X_0\boldsymbol{\beta}_0)^T(I_n - P)(\mathbf{Y} - X_0\boldsymbol{\beta}_0) \text{ since } (I_n - P)X_0 = 0, \\ \mathbf{Y}^T(P - P_0)\mathbf{Y} &= (\mathbf{Y} - X_0\boldsymbol{\beta}_0)^T(P - P_0)(\mathbf{Y} - X_0\boldsymbol{\beta}_0) \text{ since } (P - P_0)X_0 = 0, \end{aligned}$$

- Applying Lemmas 13.2 ($\mathbf{Z}^T A_i \mathbf{Z} \sim \sigma^2 \chi^2_r$) and 15.1 to $\mathbf{Z} = \mathbf{Y} - X_0\beta_0, A_1 = I_n - P, A_2 = P - P_0$ to get that under H_0,

$$\begin{aligned} \text{RSS} = \mathbf{Y}^T(I_n - P)\mathbf{Y} &\sim \chi^2_{n-p} \\ \text{RSS}_0 - \text{RSS} = \mathbf{Y}^T(P - P_0)\mathbf{Y} &\sim \chi^2_{p-p_0} \end{aligned}$$

  and these rvs are independent.

- So under H_0,

$$F = \frac{\mathbf{Y}^T(P - P_0)\mathbf{Y}/(p - p_0)}{\mathbf{Y}^T(I_n - P)\mathbf{Y}/(n - p)} = \frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n - p)} \sim F_{p-p_0, n-p}.$$

- Hence we reject H_0 if $F > F_{p-p_0, n-p}(\alpha)$.
- RSS_0 - RSS is the 'reduction in the sum of squares due to fitting β_1.

Arrangement as an 'analysis of variance' table

Source of variation	degrees of freedom (df)	sum of squares	mean square	F statistic
Fitted model	$p - p_0$	RSS_0 - RSS	$\frac{(\text{RSS}_0 - \text{RSS})}{(p - p_0)}$	$\frac{(\text{RSS}_0 - \text{RSS})/(p - p_0)}{\text{RSS}/(n-p)}$
Residual	$n - p$	RSS	$\frac{\text{RSS}}{(n-p)}$	
	$n - p_0$	RSS_0		

The ratio $\frac{(\text{RSS}_0 - \text{RSS})}{\text{RSS}_0}$ is sometimes known as the *proportion of variance explained by* β_1, and denoted R^2.

Simple linear regression

- We assume that

$$Y_i = a' + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

  where $\bar{x} = \sum x_i/n$, and $\varepsilon_i, i = 1, ..., n$ are iid $N(0, \sigma^2)$.
- Suppose we want to test the hypothesis $H_0 : b = 0$, i.e. no linear relationship. From Lecture 14 we have seen how to construct a confidence interval, and so could simply see if it included 0.
- Alternatively , under H_0, the model is $Y_i \sim N(a', \sigma^2)$, and so $\hat{a}' = \overline{Y}$, and the fitted values are $\hat{Y}_i = \overline{Y}$.
- The observed RSS_0 is therefore

$$RSS_0 = \sum_i (y_i - \overline{y})^2 = S_{yy}.$$

- The fitted sum of squares is therefore

$$RSS_0 - RSS = \sum_i \left((y_i - \overline{y})^2 - (y_i - \overline{y} - \hat{b}(x_i - \bar{x}))^2 \right) = \hat{b}^2 (x_i - \bar{x})^2 = \hat{b}^2 S_{xx}.$$

Source of variation	d.f.	sum of squares	mean square	F statistic
Fitted model	1	$\text{RSS}_0 - \text{RSS} = \hat{b}^2 S_{xx}$	$\hat{b}^2 S_{xx}$	$F = \hat{b}^2 S_{xx}/\tilde{\sigma}^2$
Residual	$n-2$	$\text{RSS} = \sum_i (y_i - \hat{y})^2$	$\tilde{\sigma}^2$	
	$n-1$	$\text{RSS}_0 = \sum_i (y_i - \overline{y})^2$		

- Note that the proportion of variance explained is $\hat{b}^2 S_{xx}/S_{yy} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$, where r is Pearson's Product Moment Correlation coefficient $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$.
- From lecture 14, slide 5, we see that under H_0, $\frac{\hat{b}}{\text{s.e.}(\hat{b})} \sim t_{n-2}$, where $\text{s.e.}(\hat{b}) = \tilde{\sigma}/\sqrt{S_{xx}}$.
  So $\frac{\hat{b}}{\text{s.e.}(\hat{b})} = \frac{\hat{b}\sqrt{S_{xx}}}{\tilde{\sigma}} = t$.
- Checking whether $|t| > t_{n-2}(\frac{\alpha}{2})$ is precisely the same as checking whether $t^2 = F > F_{1,n-2}(\alpha)$, since a $F_{1,n-2}$ variable is t_{n-2}^2.
- Hence the same conclusion is reached, whether based on a t-distribution or the F statistic derived from an analysis-of-variance table.

Example 12.1 continued

As R code

```
> fit=lm(time~ oxy.s  )
> summary.aov(fit)

            Df Sum Sq Mean Sq F value   Pr(>F)
oxy.s        1 129690  129690   41.98 1.62e-06 ***
Residuals   22  67968    3089
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note that the F statistic, 41.98, is -6.48^2, the square of the t statistic on Slide 5 in Lecture 14.

One way analysis of variance with equal numbers in each group

- Assume J measurements taken in each of I groups, and that

$$Y_{i,j} = \mu_i + \varepsilon_{i,j},$$

  where $\varepsilon_{i,j}$ are independent $\mathrm{N}(0, \sigma^2)$ random variables, and the μ_i's are unknown constants.
- Fitting this model gives
  $\mathrm{RSS} = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i,j} - \hat{\mu}_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{i,j} - \overline{Y}_{i.})^2$ on $n - I$ degrees of freedom.
- Suppose we want to test the hypothesis $H_0 : \mu_i = \mu$, i.e. no difference between groups.
- Under H_0, the model is $Y_{i,j} \sim \mathrm{N}(\mu, \sigma^2)$, and so $\hat{\mu} = \overline{Y}_{..}$, and the fitted values are $\hat{Y}_{i,j} = \overline{Y}_{..}$.
- The observed RSS_0 is therefore

$$\mathrm{RSS}_0 = \sum_i \sum_j (y_{i,j} - \overline{y}_{..})^2.$$

- The fitted sum of squares is therefore

$$\mathrm{RSS}_0 - \mathrm{RSS} = \sum_i \sum_j \left((y_{i,j} - \overline{y}_{..})^2 - (y_{i,j} - \overline{y}_{i.})^2 \right) = J \sum_i (\overline{y}_{i.} - \overline{y}_{..})^2.$$

Source of variation	d.f.	sum of squares	mean square	F statistic
Fitted model	$I-1$	$J\sum_i(\overline{y}_{i.}-\overline{y}_{..})^2$	$\frac{J\sum_i(\overline{y}_{i.}-\overline{y}_{..})^2}{(I-1)}$	$F = \frac{J\sum_i(\overline{y}_{i.}-\overline{y}_{..})^2}{(I-1)\tilde{\sigma}^2}$
Residual	$n-I$	$\sum_i\sum_j(y_{i,j}-\overline{y}_{i.})^2$	$\tilde{\sigma}^2$	
	$n-1$	$\sum_i\sum_j(y_{i,j}-\overline{y}_{..})^2$		

Example 13.1

As R code

```
> summary.aov(fit)

             Df Sum Sq Mean Sq F value Pr(>F)
x             4  507.9   127.0    1.17  0.354
Residuals    20 2170.1   108.5
```

The p-value is 0.35, and so there is no evidence for a difference between the instruments.

Lecture 16. Linear model examples, and 'rules of thumb'

Two samples: testing equality of means, unknown common variance.

- Suppose we have two independent samples,
  $X_1, \ldots, X_m$ iid $N(\mu_X, \sigma^2)$, and $Y_1, \ldots Y_n$ iid $N(\mu_Y, \sigma^2)$,
  with σ^2 unknown.
- We wish to test $H_0 : \mu_X = \mu_Y = \mu$ against $H_1 : \mu_X \neq \mu_Y$.

Using the generalised likelihood ratio test

- $L_{\mathbf{x},\mathbf{y}}(H_0) = \sup_{\mu,\sigma^2} f_{\mathbf{X}}(\mathbf{x}\,|\,\mu,\sigma^2) f_{\mathbf{Y}}(\mathbf{y}\,|\,\mu,\sigma^2)$.
- Under H_0 the mle's are

$$\begin{aligned}\hat{\mu} &= (m\bar{x} + n\bar{y})/(m+n)\\ \hat{\sigma}_0^2 &= \tfrac{1}{m+n}\left(\sum(x_i-\hat{\mu})^2 + \sum(y_i-\hat{\mu})^2\right) = \tfrac{1}{m+n}\left(S_{xx} + S_{yy} + \tfrac{mn}{m+n}(\bar{x}-\bar{y})^2\right)\end{aligned}$$

so

$$\begin{aligned}L_{\mathbf{x},\mathbf{y}}(H_0) &= (2\pi\hat{\sigma}_0^2)^{-(m+n)/2} e^{-\frac{1}{2\hat{\sigma}_0^2}\left(\sum(x_i-\hat{\mu})^2+\sum(y_i-\hat{\mu})^2\right)}\\ &= (2\pi\hat{\sigma}_0^2)^{-(m+n)/2} e^{-\frac{m+n}{2}}.\end{aligned}$$

- Similarly

$$L_{\mathbf{x},\mathbf{y}}(H_1) = \sup_{\mu_X,\mu_Y,\sigma^2} f_{\mathbf{X}}(\mathbf{x}\,|\,\mu_X,\sigma^2) f_{\mathbf{Y}}(\mathbf{y}\,|\,\mu_Y,\sigma^2) = (2\pi\hat{\sigma}_1^2)^{-(m+n)/2} e^{-\frac{m+n}{2}},$$

  achieved by $\hat{\mu}_X = \bar{x}$, $\hat{\mu}_Y = \bar{y}$ and $\hat{\sigma}_1^2 = (S_{xx} + S_{yy})/(m+n)$.

- Hence

$$\Lambda_{\mathbf{x},\mathbf{y}}(H_0; H_1) = \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{(m+n)/2} = \left(1 + \frac{mn(\bar{x}-\bar{y})^2}{(m+n)(S_{xx}+S_{yy})}\right)^{(m+n)/2}.$$

- We reject H_0 if $mn(\bar{x}-\bar{y})^2/\big((S_{xx}+S_{yy})(m+n)\big)$ is large, or equivalently if

$$|t| = \frac{|\bar{x}-\bar{y}|}{\sqrt{\frac{S_{xx}+S_{yy}}{n+m-2}\left(\frac{1}{m}+\frac{1}{n}\right)}}$$

  is large.

- Under H_0, $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{m})$, $\bar{Y} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ and so

$$(\bar{X} - \bar{Y}) / \left(\sigma \sqrt{\tfrac{1}{m} + \tfrac{1}{n}} \right) \sim \mathrm{N}(0, 1).$$

- From Theorem 16.3 we know $S_{XX}/\sigma^2 \sim \chi^2_{m-1}$ independently of $\bar{X}$ and $S_{YY}/\sigma^2 \sim \chi^2_{n-1}$ independently of $\bar{Y}$.
- Hence $(S_{XX} + S_{YY})/\sigma^2 \sim \chi^2_{n+m-2}$, from additivity of independent χ^2 distributions.
- Since our two random samples are independent, we have $\bar{X} - \bar{Y}$ and $S_{XX} + S_{YY}$ are independent.
- This means that under H_0,

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_{XX}+S_{YY}}{n+m-2} \left(\frac{1}{m} + \frac{1}{n} \right)}} \sim t_{n+m-2}.$$

- A size α test is to reject H_0 if $|t| > t_{n+m-2}(\alpha/2)$.

Example 16.1

Seeds of a particular variety of plant were randomly assigned either to a nutritionally rich environment (the treatment) or to the standard conditions (the control). After a predetermined period, all plants were harvested, dried and weighed, with weights as shown below in grams.

Control	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Treatment	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

- Control observations are realisations of $X_1, \ldots, X_{10}$ iid $N(\mu_X, \sigma^2)$, and for the treatment we have $Y_1, \ldots, Y_{10}$ iid $N(\mu_Y, \sigma^2)$.
- We test $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X \neq \mu_Y$.
- Here $m = n = 10$, $\bar{x} = 5.032$, $S_{xx} = 3.060$, $\bar{y} = 4.661$ and $S_{yy} = 5.669$, so $\tilde{\sigma}^2 = (S_{xx} + S_{yy})/(m + n - 2) = 0.485$.
- Then $|t| = |\bar{x} - \bar{y}|/\sqrt{\tilde{\sigma}^2(\frac{1}{m} + \frac{1}{n})} = 1.19$.
- From tables $t_{18}(0.025) = 2.101$, so we do not reject H_0. We conclude that there is no evidence for a difference between the mean weights due to the environmental conditions.

Arranged as analysis of variance:

Source of variation	d.f.	sum of squares	mean square	F statistic
Fitted model	1	$\frac{mn}{m+n}(\bar{x}-\bar{y})^2$	$\frac{mn}{m+n}(\bar{x}-\bar{y})^2$	$F = \frac{mn}{m+n}(\bar{x}-\bar{y})^2/\tilde{\sigma}^2$
Residual	$m+n-2$	$S_{xx}+S_{yy}$	$\tilde{\sigma}^2$	
	$m+n-1$			

Seeing if $F > F_{1,m+n-2}(\alpha)$ is exactly the same as checking if $|t| > t_{n+m-2}(\alpha/2)$.

Notice that although we have equal size samples here, they are not paired; there is nothing to connect the first plant in the control sample with the first plant in the treatment sample.

Paired observations

- Suppose the observations *were* paired: say because pairs of plants were randomised.
- We can introduce a parameter γ_i for the ith pair, where $\sum_i \gamma_i = 0$, so that we assume

$$X_i \sim \mathrm{N}(\mu_X + \gamma_i, \sigma^2), \quad Y_i \sim \mathrm{N}(\mu_Y + \gamma_i, \sigma^2),\ i = 1, .., n,$$

and all independent.

- Working through the generalised likelihood ratio test, or expressing in matrix form, leads to the intuitive conclusion that we should work with the differences $D_i = X_i - Y_i, i = 1, .., n$, where

$$D_i \sim \mathrm{N}(\mu_X - \mu_Y, \phi^2), \text{ where } \phi^2 = 2\sigma^2.$$

- Thus $\overline{D} \sim \mathrm{N}(\mu_X - \mu_Y, \frac{\phi^2}{n})$,, and we test $H_0 : \mu_X - \mu_Y = 0$ by the t statistic

$$t = \frac{\overline{D}}{\tilde{\phi}/\sqrt{n}},$$

where $\tilde{\phi}^2 = S_{DD}/(n-1) = \sum_i (D_i - \overline{D})^2/(n-1)$, and $t \sim t_{n-1}$ distribution under H_0.

Example 16.2

Pairs of seeds of a particular variety of plant were sampled, and then one of each pair randomly assigned either to a nutritionally rich environment (the treatment) or to the standard conditions (the control).

Pair	1	2	3	4	5	6	7	8	9	10
Control	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Treatment	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
Difference	- 0.64	1.41	0.77	2.52	-1.37	0.78	-0.86	-0.36	1.01	0.45

- Observed statistics are $\overline{d} = 0.37, S_{dd} = 12.54, n = 10$, so that $\tilde{\phi} = \sqrt{S_{dd}/(n-1)} = \sqrt{2.33/9} = 1.18$.
- Thus $t = \frac{\overline{d}}{\tilde{\phi}/\sqrt{n}} = \frac{0.37}{1.18/\sqrt{10}} = 0.99$.
- This can be compared to $t_{18}(0.025) = 2.262$ to show that we cannot reject $H_0 : \mathbb{E}(D) = 0$, i.e. that there is no effect of the treatment.
- Alternatively, we see that the observed p-value is the probability of getting such an extreme result, under H_0, i.e.

$$\mathbb{P}(|t_9| > |t||H_0) = 2\mathbb{P}(t_9 > |t|) = 2 \times 0.17 = 0.34.$$

In R code:

```
> t.test(x,y,paired=T)

t = 0.9938, df = 9, p-value = 0.3463
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4734609  1.2154609
sample estimates:
mean of the differences
                  0.371
```

Rules of thumb: the 'rule of three'*

Rules of Thumb 16.3

If there have been n opportunities for an event to occur, and yet it has not occurred yet, then we can be 95% confident that the chance of it occurring at the next opportunity is less than $3/n$.

- Let p be the chance of it occurring at each opportunity. Assume these are independent Bernoulli trials, so essentially we have $X \sim \text{Binom}(n, p)$, we have observed $X = 0$, and want a one-sided 95% CI for p.
- Base this on the set of values that cannot be rejected at the 5% level in a one-sided test.
- i.e. the 95% interval is $(0, p')$ where the one-sided p-value for p' is 0.05, so

$$0.05 = \mathbb{P}(X = 0|p') = (1 - p')^n.$$

- Hence

$$p' = 1 - e^{\log(0.05)/n} \approx \frac{-\log(0.05)}{n} \approx \frac{3}{n},$$

since $\log(0.05) = -2.9957$.

- For example, suppose we have given a drug to 100 people and none of them have had a serious adverse reaction.
- Then we can be 95% confident that the chance the next person has a serious reaction is less than 3%.
- The exact p' is $1 - e^{\log(0.05)/100} = 0.0295$.

Rules of thumb: the 'rule of root n'*

Rules of Thumb 16.4

After n observations, if the number of events differs from that expected under a null hypothesis H_0 by more than $\sqrt{n}$, reject H_0.

- We assume $X \sim \text{Binom}(n, p)$, and $H_0 : p = p_0$, so the expected number of events is $\mathbb{E}(X|H_0) = np_0$.
- Then the probability of the difference between observed and expected exceeding $\sqrt{n}$, given H_0 is true, is

$$\begin{aligned}
\mathbb{P}(|X - np_0| > \sqrt{n}|H_0) &= \mathbb{P}\left(\frac{|X - np_0|}{\sqrt{np_0(1-p_0)}} > \frac{1}{\sqrt{p_0(1-p_0)}}\middle| H_0\right) \\
&< \mathbb{P}\left(\frac{|X - np_0|}{\sqrt{np_0(1-p_0)}} > 2\middle| H_0\right) \text{ since } \frac{1}{\sqrt{p_0(1-p_0)}} > 2 \\
&\approx \mathbb{P}(|Z| > 2) \\
&\approx 0.05
\end{aligned}$$

- For example, suppose we flip a coin 1000 times and it comes up heads 550 times, do we think the coin is odd?
- We expect 500 heads, and observe 50 more. $\sqrt{n} = \sqrt{1000} \approx 32$, which is less than 50, so this suggests the coin is odd.
- The 2-sided p-value is actually $2 \times \mathbb{P}(X \geq 550) = 2 \times (1 - \mathbb{P}(X \leq 549))$, where $X \sim \text{Binom}(1000, 0.5)$, which according to R is

```
> 2 * (1 - pbinom(549,1000,0.5))

0.001730536
```

Rules of thumb: the 'rule of root 4 × expected'*

The 'rule of root n' is fine for chances around 0.5, but is too lenient for rarer events, in which case the following can be used.

Rules of Thumb 16.5

After n observations, if the number of rare events differs from that expected under a null hypothesis H_0 by more than '4× expected', reject H_0.

- We assume $X \sim \text{Binom}(n, p)$, and $H_0 : p = p_0$, so the expected number of events is $\mathbb{E}(X|H_0) = np_0$.
- Under H_0, the critical difference is $\approx 2 \times \text{s.e.}(X - np_0) = \sqrt{4np_0(1-p_0)}$, which is less than $\sqrt{n}$: this is the rule of root n.
- But $\sqrt{4np_0(1-p_0)} < \sqrt{4np_0}$, which will be less than $\sqrt{n}$ if $p_0 < 0.25$.
- So for smaller p_0, a more powerful rule is to reject H_0 if the difference between observed and expected is greater than $\sqrt{4 \times \text{expected}}$.
- This is essentially a Poisson approximation.

- For example, suppose we throw a die 120 times and it comes up 'six' 30 times; is this 'significant'?
- We expect 20 sixes, and so the difference between observed and expected is 10.
- Since $\sqrt{n} = \sqrt{120} \approx 11$, which is more than 10, the 'rule of root n' does not suggest a significant difference.
- But since $\sqrt{4 \times \text{ expected}} = \sqrt{80} \approx 9$, the second rule does suggest significance.
- The 2-sided p-value is actually $2 \times \mathbb{P}(X \geq 30) = 2 \times (1 - \mathbb{P}(X \leq 29))$, where $X \sim \text{Binom}(120, \frac{1}{6})$, which according to R is

```
> 2 * (1 - pbinom(29,120, 1/6 ))

0.02576321
```

Rules of thumb: non-overlapping confidence intervals*

Rules of Thumb 16.6

Suppose we have 95% confidence intervals for μ_1 and μ_2 based on independent estimates $\overline{y}_1$ and $\overline{y}_2$. Let $H_0 : \mu_1 = \mu_2$.

(1) *If the confidence intervals* **do not** *overlap, then we can reject H_0 at $p < 0.05$.*

(2) *If the confidence intervals* **do** *overlap, then this does not necessarily imply that we cannot reject H_0 at $p < 0.05$.*

- Assume for simplicity that the confidence intervals are based on assuming $\overline{Y}_1 \sim \mathrm{N}(\mu_1, s_1^2)$, $\overline{Y}_2 \sim \mathrm{N}(\mu_2, s_2^2)$, where s_1 and s_2 are known standard errors.
- Suppose wlg that $\overline{y}_1 > \overline{y}_2$. Then since $\overline{Y}_1 - \overline{Y}_2 \sim \mathrm{N}(\mu_1 - \mu_2, s_1^2 + s_2^2)$, we can reject H_0 at $\alpha = 0.05$ if

$$\overline{y}_1 - \overline{y}_2 > 1.96\sqrt{s_1^2 + s_2^2}.$$

- The two CIs will not overlap if

$$\overline{y}_1 - 1.96 s_1 > \overline{y}_2 + 1.96 s_2, \text{ i.e. } \overline{y}_1 - \overline{y}_2 > 1.96(s_1 + s_2).$$

- But since $s_1 + s_2 > \sqrt{s_1^2 + s_2^2}$ for positive s_1, s_2, we have the 'rule of thumb'.

- Non-overlapping CIs is a more stringent criterion: we cannot conclude 'not significantly different' just because CIs overlap.
- So, if 95% CIs just touch, what is the p-value?
- Suppose $s_1 = s_2 = s$. Then CIs just touch if $|\overline{y}_1 - \overline{y}_2| = 1.96 \times 2s = 3.92 \times s$.
- So p-value $=$

$$\begin{aligned}\mathbb{P}(|\overline{Y}_1 - \overline{Y}_2| > 3.92s) &= \mathbb{P}\left(\left|\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{2}s}\right| > \frac{3.92}{\sqrt{2}}\right) \\ &= \mathbb{P}(|Z| > 2.77) = 2 \times \mathbb{P}(Z > 2.77) = 0.0055.\end{aligned}$$

- And if 'just not touching' $100(1-\alpha)\%$ CIs were to be equivalent to 'just rejecting H_0', then we would need to set α so that the critical difference between $\overline{y}_1 - \overline{y}_2$ was exactly the width of each of the CIs, and so

$$1.96 \times \sqrt{2} \times s = s \times \Phi^{-1}(1 - \tfrac{\alpha}{2}).$$

- Which means $\alpha = 2 \times \Phi(-1.96/\sqrt{2}) = 0.16$.
- So in these specific circumstances, we would need to use 84% intervals in order to make non-overlapping CIs the same as rejecting H_0 at the 5% level.