1. Introduction and probability review

1.1. 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

1.2. 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 $\theta$, taking values in some parameter space $\Theta$
- 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 $\theta$ is unknown.
Let \( X_1, X_2, \ldots, X_n \) be independent and identically distributed (iid) with the same distribution as \( X \), so that \( X = (X_1, X_2, \ldots, X_n) \) is a simple random sample (our data).

We use the observed \( X = x \) to make inferences about \( \theta \), such as,

(a) giving an estimate \( \hat{\theta}(x) \) of the true value of \( \theta \) (point estimation);
(b) giving an interval estimate \( (\hat{\theta}_1(x), \hat{\theta}_2(x)) \) for \( \theta \);
(c) testing a hypothesis about \( \theta \), e.g., 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

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

The distribution function of a r.v. \( X \) is \( F_X(x) = P(X \leq x) \), 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) \( f_X(x) = P(X = x) \).

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

$$P(X \in A) = \int_A f_X(t) \, dt$$

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 $p$.]

### Independence

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

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n).$$

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 $X = (X_1, \ldots, X_n)$ has pdf or pmf

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

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

$$F_Y(y) = P(Y \leq y) = P(\max(X_1, \ldots, X_n) \leq y) = P(X_1 \leq y, \ldots, X_n \leq y) = P(X_1 \leq y)^n = [F_X(y)]^n$$

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,
\[
E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)
\]
\[
E(a_1X_1 + b_1) = a_1E(X_1) + b_1
\]
\[
E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n)
\]
\[
\text{var}(a_1X_1 + b_1) = a_1^2 \text{var}(X_1)
\]
For independent random variables,
\[
E(X_1 \times \cdots \times X_n) = E(X_1) \times \cdots \times E(X_n),
\]
\[
\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n),
\]
and
\[
\text{var}(a_1X_1 + \cdots + a_nX_n) = a_1^2 \text{var}(X_1) + \cdots + a_n^2 \text{var}(X_n).
\]

Moment generating functions

The **moment generating function** for a r.v. $X$ is
\[
M_X(t) = E(e^{tX}) = \left\{ \begin{array}{ll}
\sum_{x \in \mathbb{R}} e^{tx}P(X = x) & \text{if } X \text{ is discrete} \\
\int e^{tx}f_X(x)dx & \text{if } X \text{ is continuous.}
\end{array} \right.
\]
provided $M$ exists for $t$ in a neighbourhood of 0.

Can use this to obtain moments of $X$, since
\[
E(X^n) = M_X^{(n)}(0),
\]
i.e. $n$th derivative of $M$ evaluated at $t = 0$.
Under broad conditions, $M_X(t) = M_Y(t)$ implies $F_X = F_Y$.

Standardised statistics

Suppose $X_1, \ldots, X_n$ are iid with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2$.
Write their sum as
\[
S_n = \sum_{i=1}^{n} X_i
\]
From preceding slide, $E(S_n) = n\mu$ and $\text{var}(S_n) = n\sigma^2$.
Let $\bar{X}_n = S_n/n$ be the sample mean.
Then $E(\bar{X}_n) = \mu$ 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 $E(Z_n) = 0$ and $\text{var}(Z_n) = 1$.
$Z_n$ is known as a **standardised statistic**.

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

**Example:** sum of Poissons
If $X_i \sim \text{Poisson}(\mu)$, then
\[
M_{X_i}(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx}e^{-\mu}x! = e^{-\mu(1-e^t)} \sum_{x=0}^{\infty} e^{-\mu e^t}x!/x! = e^{-\mu(1-e^t)}.
\]
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. \( \square \)
Convergence

The Weak Law of Large Numbers (WLLN) states that for all \( \epsilon > 0 \),
\[
P \left( \left| \bar{X}_n - \mu \right| > \epsilon \right) \to 0 \text{ as } n \to \infty.
\]

The Strong Law of Large Numbers (SLLN) says that
\[
P \left( \bar{X}_n \to \mu \right) = 1.
\]

The Central Limit Theorem tells us that
\[
Z_n = \frac{S_n - n\mu}{\sigma \sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \text{ is approximately } N(0,1) \text{ for large } n.
\]

Conditioning

Let \( X \) and \( Y \) be discrete random variables with joint pmf
\[
p_{X,Y}(x,y) = P(X=x, Y=y).
\]

Then the marginal pmf of \( Y \) is
\[
p_Y(y) = P(Y=y) = \sum_x p_{X,Y}(x,y).
\]

The conditional pmf of \( X \) given \( Y = y \) is
\[
p_{X|Y}(x|y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)},
\]
if \( p_Y(y) \neq 0 \) (and is defined to be zero if \( p_Y(y) = 0 \)).

In the continuous case, suppose that \( X \) and \( Y \) have joint pdf \( f_{X,Y}(x,y) \), so that for example
\[
P(X \leq x_1, Y \leq y_2) = \int_{-\infty}^{y_2} \int_{-\infty}^{x_1} f_{X,Y}(x,y) \, dx \, dy.
\]

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
\[
f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},
\]
if \( f_Y(y) \neq 0 \) (and is defined to be zero if \( f_Y(y) = 0 \)).

Proof [discrete case]:
\[
E[X | Y] = \sum_y \left( \sum_x f_{X|Y}(x|y) \right) f_Y(y) = \sum_y \sum_x x \cdot f_{X|Y}(x|y) f_Y(y) = \sum_x \left( \sum_y y \cdot f_{Y|X}(y|x) \right) f_Y(y).
\]

Thus \( E[X | Y] \) is a function of \( y \), and \( E[X | Y] \) is a function of \( Y \) and hence a r.v.

The conditional expectation formula says
\[
E[X] = E[E[X | Y]].
\]
The conditional variance of $X$ given $Y = y$ is defined by
\[
\text{var}(X \mid Y = y) = \mathbb{E}[ (X - \mathbb{E}(X \mid Y = y))^2 \mid Y = y],
\]
and this is equal to $\mathbb{E}(X^2 \mid Y = y) - (\mathbb{E}(X \mid Y = y))^2$.

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:**
\[
\text{var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2
= \mathbb{E}[\mathbb{E}(X^2 \mid Y)] - [\mathbb{E}[\mathbb{E}(X \mid Y)]]^2
= \mathbb{E}[\mathbb{E}(X^2 \mid Y)] - \mathbb{E}(X \mid Y)^2 + \mathbb{E}[\mathbb{E}[\mathbb{E}(X \mid Y)^2] - [\mathbb{E}[\mathbb{E}(X \mid Y)]]^2
= \mathbb{E}[\text{var}(X \mid Y)] - \text{var}[\mathbb{E}(X \mid Y)].
\]

---

### Change of variable (illustrated in 2-d)

Let the joint density of random variables $(X, Y)$ be $f_{X,Y}(x, y)$.
Consider a 1-1 (bijective) differentiable transformation to random variables
$(U(X, Y), V(X, Y))$, with inverse $(X(U, V), Y(U, V))$.
Then the joint density of $(U, V)$ is given by
\[
f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v)) |J|,
\]
where $J$ is the Jacobian
\[
J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}
\]

---

### 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

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

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

R code:

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

![Graph showing the distribution of the number of sixes in 10 throws](image.png)
1. Introduction and probability review

1.15. Some important discrete distributions: Poisson

Some important discrete distributions: Poisson

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

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

(zero otherwise).

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

In a **Poisson process** the number of events $X(t)$ in an interval of length $t$ is $\text{Poisson}(\mu t)$, where $\mu$ is the rate per unit time.

The Poisson($\mu$) 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 $X \sim \text{Poisson}(6)$.

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:

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 $X \sim \text{Poisson}(6)$.

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

---

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 $X \sim \text{Poisson}(6)$.

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 $X \sim \text{Poisson}(6)$.

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 $X \sim \text{Poisson}(6)$.

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 $X \sim \text{Poisson}(6)$.

```
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

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

(zero otherwise). Then $E(X) = k/p$ and $\text{var}(X) = k(1-p)/p^2$. This is the distribution of the number of trials up to and including the $k$th 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

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

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

\[
P(\mathbf{N} = n_1, \ldots, n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \cdots p_k^{n_k}, \quad \text{if} \quad \sum_i n_i = n,
\]

and is zero otherwise.

The r.v’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 \( \frac{6!}{1!1!1!1!1!1!} (\frac{1}{6})^6 = 0.015 \)

\[
dmultinom(x=c(1,1,1,1,1,1), size=6, prob=rep(1/6,6))
\]

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

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

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 \((\alpha, \lambda)\) distribution \((\alpha > 0, \lambda > 0)\) if it has pdf

\[
f_X(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0,
\]

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

\[
M_X(t) = E(e^{tx}) = \left( \frac{\lambda}{\lambda-t} \right)^\alpha, \quad \text{for} \ t < \lambda.
\]

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 $\lambda$ ($\lambda > 0$) if

$X \sim \text{Gamma}(1, \lambda)$, so that $X$ has pdf

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

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

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

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

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, $X \sim \chi^2_k$.

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

Further, the moment generating function of $Z_i^2$ is

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

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

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

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

Some Gamma distributions:

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

```R
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 $X \sim \text{Gamma}(k/2, 1/2)$.
2. If $Y \sim \text{Gamma}(n, \lambda)$ 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 $\chi^2$.
4. We denote the upper $100\alpha$% point of $\chi^2_k$ by $\chi^2_k(\alpha)$, so that, if $X \sim \chi^2_k$ then $P\left(X > \chi^2_k(\alpha)\right) = \alpha$. These are tabulated. The above connections between gamma and $\chi^2$ means that sometimes we can use $\chi^2$-tables to find percentage points for gamma distributions.

Some important continuous distributions: Beta

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

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

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

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

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)~ U[0,1].
Lecture 2. Estimation, bias, and mean squared error

2. Estimation and bias

2.1. Estimators

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

Example
Let \( X_1, \ldots, X_n \) be iid \( N(\mu, 1) \). A possible estimator for \( \mu \) is \( T(X) = \frac{1}{n} \sum_{i=1}^{n} x_i \).

We have \( T(X) \sim N(\mu, 1/n) \).

2.2. Bias

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

\[
\text{bias}(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta.
\]

An estimator \( T(X) \) is unbiased for \( \theta \) if \( E_\theta T(X) = \theta \) for all \( \theta \), otherwise it is biased.

In the above example, \( E_\mu(T) = \mu \) so \( T \) is unbiased for \( \mu \).

[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]

2.3. Mean squared error

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 \( \theta \), whatever that value might be.

Definition 2.1
The mean squared error (mse) of an estimator \( \hat{\theta} \) is \( E_\theta[(\hat{\theta} - \theta)^2] \).

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

\[
E_\theta[(\hat{\theta} - \theta)^2] = E_\theta[(\hat{\theta} - E_\theta \hat{\theta} + E_\theta \hat{\theta} - \theta)^2] = E_\theta[(\hat{\theta} - E_\theta \hat{\theta})^2] + \{E_\theta(\hat{\theta} - \theta)\}^2 + 2\{E_\theta(\hat{\theta} - \theta)\}E_\theta(\hat{\theta} - E_\theta \hat{\theta}) = \text{var}_\theta(\hat{\theta}) + \text{bias}^2(\hat{\theta}).
\]

where bias(\( \hat{\theta} \)) = \( E_\theta(\hat{\theta} - \theta) \).

[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 $\theta$.
- The standard estimator is $T_U = X/n$, which is Unbiased since $E(\theta(T_U)) = n\theta/n = \theta$.
- $T_U$ has variance $\text{var}(T_U) = \text{var}(X)/n^2 = \theta(1-\theta)/n$.
- 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 $1/2$.)
- 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 $E(\theta(T_B)) = \theta(1 - w)/(n + 2) - \theta = (1 - w) \left(\frac{1}{2} - \theta\right)$, and so it is biased.
- $\text{var}(T_B) = \frac{\text{var}(X)}{(n + 2)^2} = w^2 \theta(1 - \theta)/n$.
- Now $\text{mse}(T_U) = \text{var}(T_U) + \text{bias}^2(T_U) = \theta(1 - \theta)/n$.
- $\text{mse}(T_B) = \text{var}(T_B) + \text{bias}^2(T_B) = w^2 \theta(1 - \theta)/n + (1 - w)^2 (\frac{1}{2} - \theta)^2$

So the biased estimator has smaller MSE in much of the range of $\theta$ $T_B$ may be preferable if we do not think $\theta$ is near 0 or 1.
So our prior judgement about $\theta$ 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 $\theta = [P(X = 0)]^2 = e^{-2\lambda}$.
Then any unbiased estimator $T(X)$ must satisfy $E_\theta(T(X)) = \theta$, or equivalently

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

The only function $T$ that can satisfy this equation is $T(X) = (-1)^X$ [coefficients of polynomial must match].
Thus the only unbiased estimator estimates $e^{-2\lambda}$ to be 1 if $X$ is even, -1 if $X$ is odd.
This is not sensible.
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 \( \theta \) that is in the sample?"

Example 3.1

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

So \( f_X(x | \theta) = \prod_{i=1}^{n} \theta^x (1 - \theta)^{1-x} = \theta^x \sum_{i=1}^{n} (1 - \theta)^{n-x} \).

This depends on the data only through \( T(x) = \sum_{i=1}^{n} x_i \), the total number of ones.

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

Let \( T(\mathbf{X}) = \sum_{i=1}^{n} x_i \) be the sufficient statistic.

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

\[
f_{\mathbf{X}|T=t}(\mathbf{x} | T=t) = \frac{f_{\mathbf{X}}(\mathbf{x} | \theta = \theta) \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T = t)}{\mathbb{P}_\theta(T = t)} = \frac{\theta^x \sum_{i=1}^{n} (1 - \theta)^{n-x}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} = \left( \frac{n}{t} \right)^{-t},
\]

i.e. the conditional distribution of \( X \) given \( T = t \) does not depend on \( \theta \).

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 \( \theta \). \( \square \)

Theorem 3.2

(The Factorisation criterion) \( T \) is sufficient for \( \theta \) iff \( f_X(x | \theta) = g(T(x), \theta)h(x) \) for suitable functions \( g \) and \( h \).

Proof (Discrete case only)

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

If \( T(X) = t \) then

\[
f_{\mathbf{X}|T=t}(\mathbf{x} | T=t) = \frac{f_{\mathbf{X}}(\mathbf{x} | \theta = \theta) \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T = t)}{\mathbb{P}_\theta(T = t)} = \frac{g(T(x), \theta)h(x)}{g(t, \theta)h(x)} = \frac{g(t, \theta)h(x)}{g(t, \theta) \sum_{x': T(x') = t} h(x')}
\]

which does not depend on \( \theta \), so \( T \) is sufficient.

Now suppose that \( T \) is sufficient so that the conditional distribution of \( \mathbf{X} | T = t \) does not depend on \( \theta \). Then

\[
\mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = t(\mathbf{x})) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x} | T = t) \mathbb{P}_\theta(T = t).
\]

The first factor does not depend on \( \theta \) by assumption; call it \( h(x) \). Let the second factor be \( g(t, \theta) \), and so we have the required factorisation. \( \square \)
Example 3.1 continued

For Bernoulli trials, \( f_X(x | \theta) = \theta^x (1 - \theta)^{n-x} \).

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

Example 3.2

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

Write \( 1_A(x) \) for the indicator function, \( = 1 \) if \( x \in A \), \( = 0 \) otherwise.

We have

\[
f_X(x | \theta) = \prod_{i=1}^n \frac{1}{\theta} 1_{[0, \theta]}(x_i) = \frac{1}{\theta^n} (\max_{i=1}^n x_i) (\min_{i=1}^n x_i).
\]

Then \( T(X) = \max_i X_i \) is sufficient for \( \theta \). \( \square \)

Minimal sufficient statistics can be found using the following theorem.

**Theorem 3.4**

Suppose \( T = T(X) \) is a statistic such that \( f_X(x; \theta)/f_Y(y; \theta) \) is constant as a function of \( \theta \) if and only if \( T(x) = T(y) \). Then \( T \) is minimal sufficient for \( \theta \).

**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 \( x \sim y \) when \( T(x) = T(y) \). (Check that this is indeed an equivalence relation.) Let \( \mathcal{U} = \{ T(X) : x \in \mathcal{X}^n \} \), and for each \( u \in \mathcal{U} \), choose a representative \( x_u \) from the equivalence class \( \{ x : T(x) = u \} \). Let \( x \in \mathcal{X}^n \) and suppose that \( T(x) = t \). Then \( x \) is in the equivalence class \( \{ x' : T(x') = t \} \), which has representative \( x_t \), and this representative may also be written \( x_{T(x)} \). We have \( x \sim x_t \), so that \( T(x) = T(x_t) \), i.e. \( T(x) = T(x_{T(x)}) \). Hence, by hypothesis, the ratio \( \frac{f(x; \theta)}{f_y(y; \theta)} \) does not depend on \( \theta \), so let this be \( h(x) \). Let \( g(t; \theta) = f(x_{T(x)}; \theta) \). Then

\[
f_X(x; \theta) = f(x_{T(x)}; \theta) \frac{f(x; \theta)}{f(x_{T(x)}; \theta)} = g(t, \theta)h(x),
\]

and so \( T = T(X) \) is sufficient for \( \theta \) by the Factorisation Criterion.

Minimal sufficient statistics

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

\( X \) itself is always sufficient for \( \theta \); take \( T(X) = X \). \( g(t, \theta) = f_X(t | \theta) \) and \( h(x) = 1 \). But this is not much use.

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

If \( T \) is sufficient, then this data reduction does not lose any information on \( \theta \).

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

**Definition 3.3**

A sufficient statistic \( T(X) \) is **minimal sufficient** if it is a function of every other sufficient statistic:

i.e. if \( T'(X) \) is also sufficient, then \( T'(X) = T'(Y) \rightarrow T(X) = T(Y) \)

i.e. the partition for \( T \) is coarser than that for \( T' \).
Example 3.3
Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$.

Then

$$f_{\tilde{X}}(x | \mu, \sigma^2) = \frac{1}{2\pi \sigma^2} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \sum_i (x_i - \mu)^2 \right] \right\}$$

This is constant as a function of $(\mu, \sigma^2)$ if $\sum_i x_i^2 = \sum_i y_i^2$ and $\sum_i x_i = \sum_i y_i$.

So $T(X) = (\sum_i X_i^2, \sum_i X_i)$ is minimal sufficient for $(\mu, \sigma^2)$. □

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

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 $\theta$ and let $\hat{\theta}$ be an estimator for $\theta$ with $E(\hat{\theta}^2) < \infty$ for all $\theta$. Let $\tilde{\theta} = E(\hat{\theta} | T)$. Then for all $\theta$,

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

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

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

$$\text{var}(\tilde{\theta}) = E[\text{var}(\hat{\theta} | T)] + \text{var}[E(\hat{\theta} | T)] = E[\text{var}(\hat{\theta} | T)] + \text{var}(\tilde{\theta}).$$

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

Notes
(i) Since $T$ is sufficient for $\theta$, the conditional distribution of $X$ given $T = t$ does not depend on $\theta$. Hence $\tilde{\theta} = E[\theta(X) | T]$ does not depend on $\theta$, 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 $\tilde{\theta} = \hat{\theta}$. 

3. Sufficiency 3.3. The Rao–Blackwell Theorem

Example 3.4
Suppose \( X_1, \ldots, X_n \) are iid Poisson(\( \lambda \)), and let \( \theta = e^{-\lambda} \) (\( = P(X_1 = 0) \)).
Then \( p_X(x|\lambda) = (e^{-n\lambda} \lambda^{\sum x_i}) / \prod x_i! \), so that \( p_X(x|\theta) = (\theta^n (-\log \theta)^{\sum x_i}) / \prod x_i! \).
We see that \( T = \sum X_i \) is sufficient for \( \theta \), and \( \sum X_i \sim \text{Poisson}(n\lambda) \).
An easy estimator of \( \theta \) is \( \tilde{\theta} = 1_{[X_i = 0]} \) (unbiased) [i.e. if do not observe any events in first observation period, assume the event is impossible!]

Then

\[
E[\tilde{\theta} | T = t] = \frac{P(X_1 = 0 \mid \sum_{i=1}^{n} X_i = t)}{P(\sum_{i=1}^{n} X_i = t)} \left( \frac{n-1}{n} \right)^t \text{(check)}.
\]

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

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

Example 3.5
Let \( X_1, \ldots, X_n \) be iid \( U[0, \theta] \), and suppose that we want to estimate \( \theta \). From Example 3.2, \( T = \max X_i \) is sufficient for \( \theta \). Let \( \tilde{\theta} = 2X_i \), an unbiased estimator for \( \theta \) [check].

Then

\[
E[\tilde{\theta} | T = t] = 2E[X_1 | \max X_i = t] = 2\left( \frac{n}{n+1} \max X_i \right) \text{(check)}.
\]

so that \( \hat{\theta} = \frac{n+1}{n} \max X_i \). \( \Box \)

In Lecture 4 we show directly that this is unbiased.

N.B. Why is \( E[X_1 | \max X_i = t, X_1 \neq \max X_i] = \frac{t}{2} \)?

Because

\[
f_{X_1}(x_1 | X_1 < t) = \frac{f_{X_1} (x_1) 1_{[0 < x_1 < t]}}{P(X_1 < t)} = \frac{1/\theta 1_{[0 < x_1 < t]}}{t/\theta} = \frac{1/\theta 1_{[0 < x_1 < t]}}{t/\theta} = \frac{1}{t} 1_{[0 < x_1 < t]}, \text{ and so } X_1 | X_1 < t \sim U[0, t].
\]
Lecture 4. Maximum Likelihood Estimation

4. Maximum likelihood estimation

4.1. 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 $f_X(x \mid \theta)$. We observe $X = x$.

**Definition 4.1**

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

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

$$\log\text{like}(\theta) = \sum_{i=1}^{n} \log f_X(x_i \mid \theta).$$

**Example 4.1**

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

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

Thus

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

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

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

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

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 $\frac{S_{XX}}{\sigma^2} \sim \chi^2_{n-1}$, 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.]
4. Maximum likelihood estimation

4.1. Likelihood

Example 4.3
Let $X_1, \ldots, X_n$ be iid $U[0, \theta]$. Then

$$\text{like}(\theta) = \frac{1}{\theta^n} 1_{\{\text{max}_i X_i \leq \theta\}}(\text{max}_i X_i).$$

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

Hence the value $\hat{\theta} = \text{max}_i X_i$ maximises the likelihood.

Is $\hat{\theta}$ unbiased? First we need to find the distribution of $\hat{\theta}$. For $0 < \text{max}_i X_i$, $\text{like}(\theta) = \frac{1}{\theta^n} > 0$ and decreasing as $\theta$ increases, while for $\theta < \text{max}_i X_i$, $\text{like}(\theta) = 0$.

Hence the value $\hat{\theta} = \text{max}_i X_i$ maximises the likelihood.

Properties of mle's

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

To maximise this as a function of $\theta$, 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 $(1 \rightarrow 1)$, then the mle of $\phi$ 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)

$$\text{like}(k) = \prod_{i=1}^k P_k(1st \text{ is a new colour}) \prod_{i=2}^k P_k(2nd \text{ is a new colour}) \prod_{i=3}^k P_k(3rd \text{ matches 1st}) \prod_{i=4}^k P_k(4th \text{ is a new colour})$$

$$= 1 \times \frac{k-1}{k} \times \frac{1}{k} \times \frac{k-2}{k} \times \frac{(k-1)(k-2)}{k^3}$$

(Alternatively, can think of Multinomial likelihood $\propto \frac{1}{k^3}$, 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 $k = 5$. 

Fairly flat! Not a lot of information.
We now consider interval estimation for $\theta$.

**Definition 5.1**
A 100$\gamma$% ($0 < \gamma < 1$) confidence interval (CI) for $\theta$ is a random interval $(A(X), B(X))$ such that $\Pr(A(X) < \theta < B(X)) = \gamma$, no matter what the true value of $\theta$ may be.

Notice that it is the endpoints of the interval that are random quantities (not $\theta$).

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

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

The above example illustrates a common procedure for finding CIs.

- Find a quantity $R(X, \theta)$ such that the $F_{R(\theta)}$-distribution of $R(X, \theta)$ does not depend on $\theta$. This is called a pivot.
- In Example 5.2, $R(X, \theta) = \sqrt{n}(\bar{X} - \theta)$.
- Write down a probability statement of the form $\Pr(\theta < R(X, \theta) < c_2) = \gamma$.
- Rearrange the inequalities inside $\Pr(\theta)$ 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 $(A(x), B(x))$ is a 100$\gamma$% CI for $\theta$, and $T(\theta)$ is a monotone increasing function of $\theta$, then $(T(A(x)), T(B(x)))$ is a 100$\gamma$% CI for $T(\theta)$.
- If $T$ is monotone decreasing, then $(T(B(x)), T(A(x)))$ 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 $\sigma^2$.

- Thus $X_i/\sigma \sim N(0, 1)$. So, from the Probability review, $\frac{1}{\sigma^2} \sum_{i=1}^{n} X_i^2 \sim \chi^2_{50}$.
- So $R(X, \sigma^2) = \sum_{i=1}^{n} X_i^2 / \sigma^2$ is a pivot.
- Recall that $\chi^2_{n}(\alpha)$ is the upper 100$\alpha$% point of $\chi^2_{n}$, i.e. $P(\chi^2_{n} \leq \chi^2_{n}(\alpha)) = 1 - \alpha$.
- From $\chi^2$-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 $P_{\sigma^2} (c_1 < \sum X_i^2 / c_2 < c_2) = 0.99$, and so $P_{\sigma^2} (\sum X_i^2 / c_2 < \sigma^2 < \sum X_i^2 / c_1) = 0.99$ which gives a confidence interval $(\sum X_i^2 / 79.49, \sum X_i^2 / 27.99)$.
- Further, a 99% confidence interval for $\sigma$ is then $(\sqrt{\sum X_i^2 / 79.49}, \sqrt{\sum X_i^2 / 27.99})$. \(\Box\)

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

$$(\hat{p} - z_{1-\gamma}/\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{1-\gamma}/\sqrt{\hat{p}(1-\hat{p})/n})$$

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 be the true proportion?

- We assume we have an observation of $x = 200$ from a Binomial$(n, \theta)$ distribution with $n = 1,000$.
- Then $\hat{p} = x/n = 0.2$ is an unbiased estimate, also the mle.
- Now var $(\hat{X}) = \frac{\theta(1-\theta)}{n} \approx \frac{0.2(1-0.2)}{1000} = 0.00016$.
- So a 95% CI is $\left(\hat{p} - 1.96 \sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96 \sqrt{\hat{p}(1-\hat{p})/n}\right) = 0.20 \pm 1.96 \times 0.013 = (0.175, 0.225)$, or around 17% to 23%.
- Special case of common procedure for an unbiased estimator $T$: 95% CI $\approx T \pm 2 \times \text{var } T = T \pm 2 \times \text{SE}$, where SE = ‘standard error’, $\approx \sqrt{\text{var } T}$
- 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’

Example 5.6
Suppose $X_1$ and $X_2$ are iid from Uniform$(\theta - 1/2, \theta + 1/2)$. What is a sensible 50% CI for $\theta$?

- Consider the probability of getting one observation each side of $\theta$,

$P_{\theta} (\min(X_1, X_2) \leq \theta \leq \max(X_1, X_2)) = P_{\theta} (X_1 \leq \theta \leq X_2) + P_{\theta} (X_2 \leq \theta \leq X_1) = \left(\frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2}\right) = 1/2$.

So $(\min(X_1, X_2), \max(X_1, X_2))$ is a 50% CI for $\theta$.
- But suppose $|X_1 - X_2| \geq \frac{1}{2}$, e.g. $x_1 = 0.2, x_2 = 0.9$. Then we know that, in this particular case, $\theta$ 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 $\theta$ are interpreted in terms of repeat sampling.
- In contrast, the Bayesian approach treats $\theta$ as a random variable taking values in $\Theta$.
- The investigator’s information and beliefs about the possible values for $\theta$, before any observation of data, are summarised by a **prior distribution** $\pi(\theta)$.
- When data $X=x$ are observed, the extra information about $\theta$ is combined with the prior to obtain the **posterior distribution** $\pi(\theta | x)$ for $\theta$ given $X=x$.
- 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,
$$\pi(\theta | x) = \frac{f_X(x | \theta) \pi(\theta)}{f_X(x)}.$$
- where $f_X(x) = \int f_X(x | \theta) \pi(\theta) d\theta$ for continuous $\theta$, and
$$f_X(x) = \sum \pi(x | \theta_i) \pi(\theta_i)$$ in the discrete case.
- Thus
$$\pi(\theta | x) \propto f_X(x | \theta) \pi(\theta) \quad (1)$$
- posterior $\propto$ likelihood $\times$ prior,
- where the constant of proportionality is chosen to make the total mass of the posterior distribution equal to one.
- In practice we use (1) and often we can recognise the family for $\pi(\theta | x)$.
- 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,
- biased 3:1 in favour of tails
- a fair coin,
- 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
- $\theta$ 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: $p(x | \theta) = \theta^x (1-\theta)^{1-x}$
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 \( r \sim \text{Binomial}(\theta, n) \), what is \( P(\theta_1 < \theta < \theta_2 | r, n) \)?

### Example 6.1

Suppose we are interested in the true mortality risk \( \theta \) 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 \( \theta \)?

- Let \( X_i = 1 \) if the \( i \)th patient dies in \( H \) (zero otherwise), \( i = 1, \ldots, n \).
- Then \( f_X(x | \theta) = \theta^x (1 - \theta)^{n-x} \).
- Suppose a prior 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

\[
\pi(\theta | x) \propto f_X(x | \theta) \pi(\theta) \\
\propto \theta^x (1 - \theta)^{n-x+b-1}, \quad 0 < \theta < 1.
\]

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

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

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 \( P(0.03 < \theta < 0.20) = 0.9 \).
- The data is \( \sum x_i = 0, n = 10 \).
- So the posterior is \( \text{Beta}(\sum x_i + a, n - \sum x_i + b) = \text{Beta}(3, 37) \).
- 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.

```r
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)
```
6. Bayesian estimation

6.2. Prior and posterior distributions

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 \( \text{Beta}(\sum x_i + 1, n - \sum x_i + 1) \), with properties.

<table>
<thead>
<tr>
<th>prior mean</th>
<th>mode</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{2} )</td>
<td>non-unique</td>
<td>( \frac{1}{12} )</td>
</tr>
</tbody>
</table>

- Notice that the mode of the posterior is the mle.
- The posterior mean estimator, \( \sum x_i + 1 \over n + 2 \) is discussed in Lecture 2, where we showed that this estimator had smaller mse than the mle for non-extreme values of \( \theta \). 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 \( \theta \).
- 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 \( h(a) = \int L(\theta, a) \pi(\theta | x) d\theta \).
- The **Bayes estimator** \( \hat{\theta} \) minimises the expected posterior loss.
- For quadratic loss
  \[
  h(a) = \int (a - \theta)^2 \pi(\theta | x) d\theta.
  \]
- \( h'(a) = 0 \) if
  \[
  a \int \pi(\theta | x) d\theta = \int \theta \pi(\theta | x) d\theta.
  \]
- So \( \hat{\theta} = \int \theta \pi(\theta | x) d\theta \), the **posterior mean**, minimises \( h(a) \).

- For absolute error loss,
  \[
  h(a) = \int |\theta - a| \pi(\theta | x) d\theta = \int_{-\infty}^{a} (a - \theta) \pi(\theta | x) d\theta + \int_{a}^{\infty} (\theta - a) \pi(\theta | x) d\theta
  \]
  \[
  = a \int_{-\infty}^{a} \pi(\theta | x) d\theta - \int_{-\infty}^{\infty} \theta \pi(\theta | x) d\theta
  \]
  \[
  + \int_{a}^{\infty} \theta \pi(\theta | x) d\theta - a \int_{a}^{\infty} \pi(\theta | x) d\theta
  \]
  Now \( h'(a) = 0 \) if
  \[
  \int_{-\infty}^{a} \pi(\theta | x) d\theta = \int_{a}^{\infty} \pi(\theta | x) d\theta.
  \]
- 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 \( N(\mu, 1) \), and that a priori \( \mu \sim N(0, \tau^{-2}) \) for known \( \tau^{-2} \).

- The posterior is given by
  \[
  \pi(\mu|x) \propto f_X(x|\mu)\pi(\mu) \\
  \propto \exp[-\frac{1}{2} \sum (x_i - \mu)^2] \exp[-\frac{\mu^2}{2\tau^2}] \\
  \propto \exp[-\frac{1}{2} (n + \tau^2) \left\{ \mu - \frac{\sum x_i}{n + \tau^2} \right\}^2] \text{ (check)}. 
  \]

- So the posterior distribution of \( \mu \) given \( x \) 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 \( \mu \) under both quadratic and absolute error loss.

Example 6.3
Suppose that \( X_1, \ldots, X_n \) are iid Poisson(\( \lambda \)) rv’s and that \( \lambda \) has an exponential distribution with mean 1, so that \( \pi(\lambda) = e^{-\lambda}, \lambda > 0 \).

- The posterior distribution is given by
  \[
  \pi(\lambda|x) \propto e^{-n\lambda} \lambda^\sum x_i e^{-\lambda} = \lambda^\sum x_i e^{-(n+1)\lambda}, \lambda > 0, 
  \]
  ie Gamma(\( \sum x_i + 1, n + 1 \)).

- Hence, under quadratic loss, \( \lambda = (\sum x_i + 1)/(n + 1) \), the posterior mean.

- Under absolute error loss, \( \lambda \) solves
  \[
  \int_0^\lambda \frac{(n + 1)^{\sum x_i + 1} \lambda^{\sum x_i} e^{-(n+1)\lambda}}{(\sum x_i)!} d\lambda = \frac{1}{2}. 
  \]
7. Simple hypotheses: 7.1. Introduction

Introduction

Let \( X_1, \ldots, X_n \) be iid, each taking values in \( X \), 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 \( X = x \), 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 \( f(x | \theta) \), 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.

Testing a simple hypothesis against a simple alternative

When \( H_0 \) and \( H_1 \) are both simple, let

\[
\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}(X \notin C | H_0 \text{ is true})
\]

\[
\beta = \mathbb{P}(\text{Type II error}) = \mathbb{P}(X \notin C | H_1 \text{ is true})
\]

We define the size of the test to be \( \alpha \).

\( 1 - \beta \) 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 \( \alpha \) and \( \beta \) arbitrarily small.

Definition 7.1

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

N.B. When Neyman and Pearson developed the theory in the 1930s, they spoke of ‘accepting’ \( H_0 \). Now we generally refer to ‘not rejecting \( H_0 \).’
Let $C$ be the critical region of any other test with size less than or equal to $\alpha$. Then among all tests of size less than or equal to $\alpha$, the test with smallest probability of a Type II error is given by $C = \{x : f_1(x)/f_0(x) > k\}$ where $k$ is chosen such that

$$\alpha = \mathbb{P}(\text{reject } H_0 | H_0) = \mathbb{P}(X \in C | H_0) = \int_C f_0(x)dx.$$ 

**Proof**

The given $C$ specifies a likelihood ratio test with size $\alpha$.

Let $\beta = \mathbb{P}(X \not\in C | f_1) = \int_{\overline{C}} f_1(x)dx$.

Let $\overline{C}^*$ be the critical region of any other test with size less than or equal to $\alpha$.

Let $\overline{C} = \{x : f_1(x) > k \overline{f}_0(x)\}$, while on $\overline{C}$ we have $f_1(x) \leq k \overline{f}_0(x)$.

Thus

$$\int_{\overline{C} \cap \overline{C}} f_1(x)dx \geq k \int_{\overline{C} \cap \overline{C}} f_0(x)dx,$$

$$\int_{\overline{C} \cap \overline{C}} f_1(x)dx \leq k \int_{\overline{C} \cap \overline{C}} f_0(x)dx. \quad (1)$$

Therefore, $\overline{C}$ is the critical region of the Neyman–Pearson Lemma.

**Theorem 7.2**

(The Neyman–Pearson Lemma) Suppose $H_0 : \mu = \mu_0, H_1 : \mu = \mu_1$, 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 $\alpha$, the test with smallest probability of a Type II error is given by $C = \{x : f_1(x)/f_0(x) > k\}$ where $k$ is chosen such that $\alpha = \mathbb{P}(\text{reject } H_0 | H_0) = \mathbb{P}(X \in C | H_0) = \int_C f_0(x)dx$.

**Example 7.3**

Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2_0)$, where $\sigma^2_0$ is known. We want to find the best size $\alpha$ test of $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, where $\mu_0$ and $\mu_1$ are known fixed values with $\mu_1 > \mu_0$.

$$\Lambda_{\alpha}(H_0; H_1) = \frac{(2\pi\sigma^2_0)^{-n/2} \exp \left(-\frac{1}{2\sigma^2_0} \sum (x_i - \mu_1)^2 \right)}{(2\pi\sigma^2_0)^{-n/2} \exp \left(-\frac{1}{2\sigma^2_0} \sum (x_i - \mu_0)^2 \right)} \exp \left(\frac{(\mu_1 - \mu_0)}{\sigma^2_0} n \bar{x} + \frac{n(\bar{x} - \mu_0^2)}{2\sigma^2_0} \right) \quad (\text{check}).$$

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

$$\Lambda_{\alpha} > k \iff \bar{x} > c \text{ for some } c.$$

Hence we reject $H_0$ if $\bar{x} > c$ where $c$ is chosen such that $\mathbb{P}(\bar{X} > c | H_0) = \alpha$.

Under $H_0$, $\bar{X} \sim N(\mu_0, \sigma^2_0/n)$, so $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 \sim N(0,1)$.

Since $\alpha > c \iff z > c'$ for some $c'$, the size $\alpha$ test rejects $H_0$ if $z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 > z_{\alpha}$.
7. Simple hypotheses

7.2. Testing a simple hypothesis against a simple alternative

Suppose \( \mu_0 = 5, \mu_1 = 6, \sigma_0 = 1, \sigma_\theta = 0.05, n = 4 \) and \( x = (5.1, 5.5, 4.9, 5.3) \), so that \( \bar{x} = 5.2. \)

- From tables, \( z_{0.05} = 1.645. \)
- We have \( z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma_\theta} = 0.4 \) and this is less than 1.645, so \( x \) 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 \( \alpha = P(Z > k \mid H_0) = 1 - \Phi(k) \), and is decreasing as \( k \) increases.
- Our observed value \( z \) will be in the rejection region
  \( \Leftrightarrow z > k \Leftrightarrow \alpha > p^* = P(Z > z \mid H_0). \)
- 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 against \( H_0. \)

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

\[
P(p^* < p \mid H_0) = P \left( [1 - \Phi(Z)] < p \mid H_0 \right) = P(Z > \Phi^{-1} (1 - p) \mid H_0)
= 1 - \Phi (\Phi^{-1} (1 - p)) = 1 - (1 - p) = p.
\]
Lecture 8. Composite hypotheses

Definition 8.1

A test specified by a critical region \( C \) is **uniformly most powerful** (UMP) size \( \alpha \) test for testing \( H_0 : \mu = \mu_0 \) against \( H_1 : \mu > \mu_0 \) if

1. \[ \sup_{\theta \in \Theta_0} W(\theta) = \alpha; \]
2. for any other test \( C^* \) with size \( \leq \alpha \) and with power function \( W^*(\theta) \) 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.

- Hence this \( C \) defines the most powerful size \( \alpha \) test of \( H_0 : \mu = \mu_0 \) against any \( \mu_1 \) that is larger than \( \mu_0 \).
- This test is then uniformly most powerful size \( \alpha \) for testing \( H_0 : \mu = \mu_0 \) against \( H_1 : \mu > \mu_0 \).

*For composite hypotheses like \( H: \theta \geq 0 \), the error probabilities do not have a single value.*

- Define the **power function** \( W(\theta) = \mathbb{P}(X \in C | \theta) = \mathbb{P}(\text{reject } H_0 | \theta) \).
- 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 \( \alpha \) test of \( H_0 : \mu = \mu_0 \) versus \( H_1 : \mu = \mu_1 \) (where \( \mu_1 > \mu_0 \)) is given by \( C = \{ x : \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 > z_{\alpha} \} \).
- This critical region depends on \( \mu_0, n, \sigma_0, \alpha \), on the fact that \( \mu_1 > \mu_0 \), but not on the particular value of \( \mu_1 \).

Example 8.2

Suppose \( X_1, \ldots, X_n \) are iid \( N(\mu_0, \sigma_0^2) \) where \( \sigma_0 \) 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 \( \alpha \) of \( H_0' \) against \( H_1' \) has
  \[ C = \{ x : \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 > z_{\alpha} \}. \]
- 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

\[
W(\mu) = \mathbb{P}_\mu(\text{reject } H_0) = \mathbb{P}_\mu \left( \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma_0} > z_{\alpha} \right)
\]

\[
= \mathbb{P}_\mu \left( \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} > z_{\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} \right)
\]

\[
= 1 - \Phi \left( z_{\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} \right).
\]
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 \( x \) to be
  \[
  L_x(H) = \sup_{\theta \in \Theta} f(x | \theta).
  \]
- So far we have considered disjoint hypotheses \( \Theta_0, \Theta_1 \), 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
  \[
  L_x(H_0, H_1) = \frac{L_x(H_1)}{L_x(H_0)} = \frac{\sup_{\theta \in \Theta_1} f(x | \theta)}{\sup_{\theta \in \Theta_0} f(x | \theta)} \geq 1, \tag{1}
  \]
  with large values of \( L_x \) indicating departure from \( H_0 \).
- Notice that if \( L_x^* = \sup_{\theta \in \Theta_1, \theta_0} f(x | \theta) / \sup_{\theta \in \Theta_0} f(x | \theta) \), then
  \[
  L_x = \max\{1, L_x^*\}.
  \]
- We know \( W(\mu_0) = \alpha \). (just plug in)
- \( W(\mu) \) is an increasing function of \( \mu \).
- 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 \( \alpha \) 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 \( \mu_1 \).)
- Let \( C^* \) and \( W^* \) belong to any other test of \( H_0 \) vs \( H_1 \) of size \( \alpha \).
- Then \( C^* \) can be regarded as a test of \( H_0 \) vs \( H_1 \) of size \( \alpha \), 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 \( \alpha \) for \( H_0 \) vs \( H_1 \).

Example 8.3

Single sample: testing a given mean, known variance (\( z \)-test). Suppose that \( X_1, \ldots, X_n \) are \( i.i.d. \) \( N(\mu, \sigma^2_0) \), with \( \sigma^2_0 \) known, and we wish to test \( H_0 : \mu = \mu_0 \)
against \( H_1 : \mu \neq \mu_0 \) (\( \mu_0 \) is a given constant).

- Here \( \Theta_0 = \{\mu_0\} \) and \( \Theta = \mathbb{R} \).
- For the denominator in (1) we have \( \sup_{\Theta_0} f(x | \mu) = f(x | \mu_0) \).
- For the numerator, we have \( \sup_{\Theta_1} f(x | \mu) = f(x | \hat{\mu}) \), where \( \hat{\mu} \) is the mle, so
  \[
  \hat{\mu} = \bar{x} \quad \text{(check)}.
  \]
- Hence
  \[
  L_x(H_0, H_1) = \frac{(2\pi\sigma^2_0)^{-n/2} \exp\left(-\frac{1}{2\sigma^2_0} \sum (x_i - \bar{x})^2\right)}{(2\pi\sigma^2_0)^{-n/2} \exp\left(-\frac{1}{2\sigma^2_0} \sum (x_i - \mu_0)^2\right)},
  \]
  and we reject \( H_0 \) if \( L_x \) is ‘large.’
- We find that
  \[
  2 \log L_x(H_0, H_1) = \frac{n}{\sigma_0^2} \left[ \sum (x_i - \mu_0)^2 - \sum (x_i - \bar{x})^2 \right] = \frac{n}{\sigma_0^2} (\bar{x} - \mu_0)^2 \quad \text{(check)}
  \]
- Thus an equivalent test is to reject \( H_0 \) if \( \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 \) is large.
8. Composite hypotheses

8.2. Generalised likelihood ratio tests

Under $H_0$, $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 \sim N(0, 1)$ so the size $\alpha$ generalised likelihood test rejects $H_0$ if $|\sqrt{n}(\bar{X} - \mu_0)/\sigma_0| > z_{\alpha/2}$.

- Since $n(\bar{X} - \mu_0)^2/\sigma_0^2 \sim \chi^2_1$ if $H_0$ is true, this is equivalent to rejecting $H_0$ if $n(\bar{X} - \mu_0)^2/\sigma_0^2 > \chi^2_1(\alpha)$ (check that $z^2_{\alpha/2} = \chi^2_1(\alpha)$). $\square$

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 $|\sqrt{n}(\bar{X} - \mu_0)/\sigma_0| > z_{\alpha/2}$. A symmetric 100$(1 - \alpha)$% confidence interval for $\mu$ is $\bar{X} \pm z_{\alpha/2} \sigma_0/\sqrt{n}$. Therefore we reject $H_0$ iff $\mu_0$ is not in this confidence interval (check).
- In later lectures the close connection between confidence intervals and hypothesis tests is explored further.

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 $X = (X_1, \ldots, X_n)$, $X_i$'s iid, we have, if $H_0$ is true,

$$2 \log \Lambda(H_0; H_1) \sim \chi^2_p.$$ 

If $H_0$ is not true, then $2 \log \Lambda$ tends to be larger. We reject $H_0$ if $2 \log \Lambda > c$ where $c = \chi^2_0(\alpha)$ for a test of approximately size $\alpha$.

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 $-2 \log \Lambda(H_1; H_0)$.)
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

<table>
<thead>
<tr>
<th>Month</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_i$</td>
<td>470</td>
<td>515</td>
<td>470</td>
<td>457</td>
<td>473</td>
<td>381</td>
<td>466</td>
<td>457</td>
<td>437</td>
<td>396</td>
<td>384</td>
<td>394</td>
</tr>
</tbody>
</table>

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

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’).

\[
2 \log \Lambda = 2 \sum n_i \log \left( \frac{\tilde{p}_i^n}{\hat{p}_i^n} \right) = 2 \sum n_i \log \left( \frac{n_i}{\hat{p}_i} \right) .
\]

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 $\alpha$ test.

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

<table>
<thead>
<tr>
<th>Month</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_i$</td>
<td>470</td>
<td>515</td>
<td>470</td>
<td>457</td>
<td>473</td>
<td>381</td>
<td>466</td>
<td>457</td>
<td>437</td>
<td>396</td>
<td>384</td>
<td>394</td>
</tr>
</tbody>
</table>

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

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’).

\[
2 \log \Lambda = 2 \sum n_i \log \left( \frac{\tilde{p}_i^n}{\hat{p}_i^n} \right) = 2 \sum n_i \log \left( \frac{n_i}{\hat{p}_i} \right) .
\]

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 $\alpha$ test.

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

<table>
<thead>
<tr>
<th>Month</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_i$</td>
<td>470</td>
<td>515</td>
<td>470</td>
<td>457</td>
<td>473</td>
<td>381</td>
<td>466</td>
<td>457</td>
<td>437</td>
<td>396</td>
<td>384</td>
<td>394</td>
</tr>
</tbody>
</table>

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

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’).

\[
2 \log \Lambda = 2 \sum n_i \log \left( \frac{\tilde{p}_i^n}{\hat{p}_i^n} \right) = 2 \sum n_i \log \left( \frac{n_i}{\hat{p}_i} \right) .
\]

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 $\alpha$ test.
Likelihood ratio tests

A similar common situation has $H_0 : p_i = p_i(\theta)$ for some parameter $\theta$ 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

$$2 \log \Lambda = 2 \log \left( \frac{\hat{p}_1^{n_1} \cdots \hat{p}_k^{n_k}}{p_1(\hat{\theta})^{n_1} \cdots p_k(\hat{\theta})^{n_k}} \right) = 2 \sum n_i \log \left( \frac{n_i}{np(\hat{\theta})} \right). \quad (2)$$

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

---

**Example 9.1 continued using R:**

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

---

### Pearson’s Chi-squared tests

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

Let $o_i = n_i$ (the observed number in $i$th set) and let $e_i$ be $np_i$ in (???) or $np_i(\hat{\theta})$ in (??). Let $\delta_i = o_i - e_i$. Then

$$2 \log \Lambda = 2 \sum \frac{o_i \log (\frac{o_i}{e_i})}{e_i}$$

$$= 2 \sum (e_i + \delta_i) \log \left( 1 + \frac{\delta_i}{e_i} \right)$$

$$\approx 2 \sum \left( \frac{\delta_i^2}{e_i} - \frac{\delta_i^3}{2e_i^2} \right)$$

$$= \sum \frac{\delta_i^2}{e_i} - \sum \frac{\delta_i^3}{2e_i^2},$$

where we have assumed $\log \left( 1 + \frac{\delta_i}{e_i} \right) \approx \frac{\delta_i}{e_i} - \frac{\delta_i^2}{2e_i}$, ignored terms in $\delta_i^3$, 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 9.2**

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 $\chi^2_3$.

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 $P(\chi^2 > 0.6) \approx 0.96$. □

**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 \( \chi^2_2 \). □

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.

<table>
<thead>
<tr>
<th>Previous</th>
<th>New car</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>Medium</td>
</tr>
<tr>
<td>Large</td>
<td>56</td>
</tr>
<tr>
<td>Medium</td>
<td>50</td>
</tr>
<tr>
<td>Small</td>
<td>18</td>
</tr>
</tbody>
</table>

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
\[
N_{+i} = \sum_j n_{ij} \quad \text{and} \quad N_{-j} = \sum_i n_{ij} \quad \text{so} \quad n_{++} = n.
\]
- 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
  \[ H_0 : p_{ij} = p_{ij}(\theta), \quad i = 1, \ldots, r; \quad j = 1, \ldots, c \quad (\text{with } \sum_i p_{ij} = 1 = \sum_j p_{ij}, \ p_{ij} > 0). \]
- Under \( H_0 \), the mle’s are \( \hat{p}_{ij} = n_{ij}/n \).
- Under \( H_0 \), using Lagrangian methods, the mle’s are \( \hat{p}_{ij} = n_{ij}/n \) and \( \hat{p}_{-j} = n_{-j}/n \).
- Write \( o_{ij} \) for \( n_{ij} \) and let \( e_{ij} = n \hat{p}_{ij} \hat{p}_{-j} = n_{ij} n_{-j}/n \).
- Then
\[
2\log A = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} (o_{ij} - e_{ij}) \log \left( \frac{o_{ij}}{e_{ij}} \right) \approx \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}
\]

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, \( |\Theta_2| = (r - 1) + (c - 1) \), because \( p_{1}, \ldots, p_{r} \) must satisfy \( \sum_j p_{-j} = 1 \) and \( p_{1}, \ldots, p_{c} \) must satisfy \( \sum_i p_{ij} = 1 \).
- 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:

<table>
<thead>
<tr>
<th>$o_{ij}$</th>
<th>New car</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large</td>
<td>Medium</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>Previous</td>
<td>Large</td>
<td>56</td>
<td>52</td>
<td>42</td>
</tr>
<tr>
<td>car</td>
<td>Medium</td>
<td>50</td>
<td>83</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>18</td>
<td>51</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>124</td>
<td>186</td>
<td>190</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$e_{ij}$</th>
<th>New car</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large</td>
<td>37.2</td>
<td>55.8</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>49.6</td>
<td>74.4</td>
<td>76.0</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>37.2</td>
<td>55.8</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>124</td>
<td>186</td>
<td>190</td>
</tr>
</tbody>
</table>

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

So our observed value of 36.20 is significant at the 1% level, i.e., 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:

<table>
<thead>
<tr>
<th></th>
<th>New car</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large</td>
<td>Medium</td>
<td>Small</td>
<td></td>
</tr>
<tr>
<td>Previous</td>
<td>Large</td>
<td>9.50</td>
<td>0.26</td>
<td>3.95</td>
</tr>
<tr>
<td>car</td>
<td>Medium</td>
<td>0.00</td>
<td>0.99</td>
<td>1.07</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>9.91</td>
<td>0.41</td>
<td>10.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>124</td>
<td>186</td>
<td>190</td>
</tr>
</tbody>
</table>

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. □
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.

<table>
<thead>
<tr>
<th></th>
<th>Improved</th>
<th>No difference</th>
<th>Worse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>18</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>Half dose</td>
<td>20</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Full dose</td>
<td>25</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

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,” i.e $H_0$ says that we have homogeneity down the rows. □

Lecture 10. Tests of homogeneity, and connections to confidence intervals
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 $f_X(x|\theta)$, $\theta \in \Theta$.

**Theorem 10.3**

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

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

In words,

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

(ii) says that given a confidence set, we define the test by rejecting $\theta_0$ 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 $\mu$.

- 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 $I(X) = \{\mu : \bar{X} \in A(\mu)\} = \{\mu : |\sqrt{n}(\bar{X} - \mu)| < 1.96\}$ so a 95% confidence set for $\mu$ 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:

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th></th>
<th></th>
<th>Men</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applied</td>
<td>Accepted</td>
<td>%</td>
<td>Applied</td>
<td>Accepted</td>
</tr>
<tr>
<td>Total</td>
<td>1184</td>
<td>274</td>
<td>23%</td>
<td>2470</td>
<td>584</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th></th>
<th></th>
<th>Men</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applied</td>
<td>Accepted</td>
<td>%</td>
<td>Applied</td>
<td>Accepted</td>
</tr>
<tr>
<td>Computer Science</td>
<td>26</td>
<td>7</td>
<td>27%</td>
<td>228</td>
<td>58</td>
</tr>
<tr>
<td>Economics</td>
<td>240</td>
<td>63</td>
<td>26%</td>
<td>512</td>
<td>112</td>
</tr>
<tr>
<td>Engineering</td>
<td>164</td>
<td>52</td>
<td>32%</td>
<td>972</td>
<td>252</td>
</tr>
<tr>
<td>Medicine</td>
<td>416</td>
<td>99</td>
<td>24%</td>
<td>578</td>
<td>140</td>
</tr>
<tr>
<td>Veterinary medicine</td>
<td>338</td>
<td>53</td>
<td>16%</td>
<td>180</td>
<td>22</td>
</tr>
</tbody>
</table>

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

Multivariate normal distribution

- Recall that a univariate normal $X \sim N(\mu, \sigma^2)$ has density
  \[ f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right), \quad x \in \mathbb{R}, \]
  and mgf
  \[ M_X(s) = \mathbb{E}[e^{sX}] = \exp(\mu s + \frac{1}{2} s^2 \sigma^2). \]
- $X$ has a **multivariate normal distribution** if, for every $t \in \mathbb{R}^n$, the rv $t^T X$ has a normal distribution.
- If $\mathbb{E}(X) = \mu$ and $\text{cov}(X) = \Sigma$, we write $X \sim N_n(\mu, \Sigma)$.
- Note $\Sigma$ is symmetric and is non-negative definite because by (1),
  \[ t^T \Sigma t = \text{var}(t^T X) \geq 0. \]
- By (1), $t^T X \sim N(t^T \mu, t^T \Sigma t)$ and so has mgf
  \[ M_{t^T X}(s) = \mathbb{E}[e^{s t^T X}] = \exp\left(t^T \mu s + \frac{1}{2} t^T \Sigma ts^2\right). \]
- Hence $X$ has mgf
  \[ M_X(t) = \mathbb{E}[e^{t^T X}] = M_{t^T X}(1) = \exp\left(t^T \mu + \frac{1}{2} t^T \Sigma t\right). \quad (2) \]

Properties of means and covariances of vectors

- A random (column) vector $X = (X_1, ..., X_n)^T$ has mean
  \[ \mu = \mathbb{E}(X) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_n))^T = (\mu_1, ..., \mu_n)^T \]
  and covariance matrix
  \[ \text{cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)^T] = (\text{cov}(X_i, X_j))_{i,j}, \]
  provided the relevant expectations exist.
- For $m \times n$ $A$,
  \[ \mathbb{E}[AX] = A\mu, \]
  and
  \[ \text{cov}(AX) = A \text{cov}(X) A^T, \quad (1) \]
  since
  \[ \text{cov}(AX) = \mathbb{E} [(AX - \mathbb{E}(AX))(AX - \mathbb{E}(AX))^T] = \mathbb{E} [A(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T A^T]. \]
- Define $\text{cov}(V, W)$ to be a matrix with $(i, j)$ element $\text{cov}(V_i, W_j)$.
  Then $\text{cov}(AX, B^TX) = A \text{cov}(X) B^T$. (check. Important for later)

Proposition 11.1

(i) If $X \sim N_n(\mu, \Sigma)$ and $A$ is $m \times n$, then $AX \sim N_m(A\mu, A\Sigma A^T)$

(ii) If $X \sim N_n(0, \sigma^2 I)$ then
  \[ \frac{|X|^2}{\sigma^2} = X^T X \frac{1}{\sigma^2} = \sum \frac{X_i^2}{\sigma^2} \sim \chi_n^2. \]

Proof:

(i) from exercise sheet 3.

(ii) Immediate from definition of $\chi_n^2$. □

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

**Proposition 11.2**

Let \( \mathbf{X} \sim N_n(\mu, \Sigma) \), \( \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \), where \( \mathbf{X}_i \) is a \( n_i \times 1 \) column vector, and \( n_1 + n_2 = n \). Write similarly \( \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \), and \( \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \), where \( \Sigma_{ij} \) is \( n_i \times n_j \). Then

(i) \( \mathbf{X}_i \sim N_{n_i}(\mu_i, \Sigma_{ii}) \),

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

**Proof:**

(i) See Example sheet 3.

(ii) From (2), \( M_k(t) = \exp(t^T \mu + \frac{1}{2} t^T \Sigma t) \), \( t \in \mathbb{R}^n \). Write

\[
M_k(t) = \exp(t_1^T \mu_1 + t_2^T \mu_2 + \frac{1}{2} t_1^T \Sigma_{11} t_1 + \frac{1}{2} t_2^T \Sigma_{22} t_2 + \frac{1}{2} t_1^T \Sigma_{12} t_2 + \frac{1}{2} t_2^T \Sigma_{21} t_1).
\]

From (i), \( M_k(t_i) = \exp(t_i^T \mu_i + \frac{1}{2} t_i^T \Sigma_{ii} t_i) \) so \( M_k(t) = M_k(t_1)M_k(t_2) \), for all \( t = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \) iff \( \Sigma_{12} = 0 \).

---

**Lecture 11. Multivariate Normal theory**

**11.4. Normal random samples**

**Density for a multivariate normal**

When \( \Sigma \) is positive definite, then \( \mathbf{X} \) has pdf

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

---

**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 \( \mathbf{X} \sim N_n(\mu, \sigma^2 I) \), where \( \mu = \mu \mathbf{1} \) ( \( \mathbf{1} \) is a \( n \times 1 \) column vector of 1’s).

Let \( A \) be the \( n \times n \) orthogonal matrix

\[
A = \begin{bmatrix}
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\
\frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{-1}{\sqrt{n(n-1)}} \\
\end{bmatrix}
\]

So \( A^TA = AA^T = I \). (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.)
Lecture 11. Multivariate Normal theory

- By Proposition 11.1(i), \( Y = AX \sim N(\mu, \sigma^2 A^T A) \sim N(\mu, \sigma^2 I) \), since \( AA^T = I \).
- We have \( A\mu = \left( \sqrt{n}\mu \right)^T 0 \cdot \), so \( Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i = \sqrt{n}\bar{X} \sim N(\sqrt{n}\mu, \sigma^2) \) (Prop 11.1 (ii)) and \( Y_i \sim N(0, \sigma^2), i = 2, \ldots, n \) and \( Y_1, \ldots, Y_n \) are independent.
- Note also that \( Y_2^2 + \ldots + Y_n^2 = Y^T Y - Y_1^2 = X^T A^T A X - Y_1^2 = X^T X - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = S_{XX} \).
- To prove (ii), note that \( S_{XX} = Y_2^2 + \ldots + Y_n^2 \sim \sigma^2 \chi^2_{n-1} \) (from definition of \( \chi^2_{n-1} \)).
- Finally, for (iii), since \( Y_1 \) and \( Y_2, \ldots, Y_n \) 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 \( T = \frac{Y}{\sqrt{Z/k}} \) 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
  \[
  f_T(t) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \frac{1}{\sqrt{k} \Gamma(k/2)} (1 + \frac{t^2}{k})^{-(k+1)/2}, \quad t \in \mathbb{R}.
  \]
- 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 \( k \to \infty \).
- \( E_t(T) = 0 \) for \( k > 1 \), otherwise undefined.
- \( \text{var}_t(T) = \frac{k}{k-2} \) for \( k > 2 \), \( = \infty \) if \( k = 2 \), otherwise undefined.
- \( k = 1 \) is known as the Cauchy distribution, and has an undefined mean and variance.

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
  \[
  \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{S_{XX}/((n-1)\sigma^2)}} \sim t_{n-1}, \quad \text{i.e.} \quad \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1}.
  \]

- Let \( \hat{\sigma}^2 = \frac{S_{XX}}{n-1} \). Note this is an unbiased estimator, as \( E(S_{XX}) = (n-1)\sigma^2 \).
- Then a 100(1 - \( \alpha \))% CI for \( \mu \) is found from
  \[
  1 - \alpha = \mathbb{P} \left( -t_{n-1}(\frac{\alpha}{2}) \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \leq t_{n-1}(\frac{\alpha}{2}) \right)
  \]
  and has endpoints
  \[
  \bar{X} \pm \hat{\sigma}/\sqrt{n} t_{n-1}(\frac{\alpha}{2}).
  \]
- See example sheet 3 for use of \( t \)-distributions in hypothesis tests.

Lecture 11. Multivariate Normal theory

- Application of Student’s \( t \)-distribution to normal random samples

- Student’s \( t \)-distribution

- 11.5. Student’s \( t \)-distribution

- 11.6. Application of Student’s \( t \)-distribution to normal random samples
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 \( Y_1, \ldots, Y_n \) are modelled as

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

(1)

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

From (1),
- \( \mathbb{E}(Y_i) = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} \)
- \( \text{var}(Y_i) = \text{var}(\varepsilon_i) = \sigma^2 \)
- \( Y_1, \ldots, Y_n \) are independent (or uncorrelated).

Note that (1) is linear in the parameters \( \beta_1, \ldots, \beta_p \) (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.

```r
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, 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, $i = 1, \ldots, 24$.

A possible model is

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

where $\varepsilon_i$ are independent random variables with variance $\sigma^2$, and $a$ and $b$ are constants.

Example 12.1 continued

Recall $Y_i = a + bx_i + \varepsilon_i, \quad i = 1, \ldots, 24$.

In matrix form:

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

Then

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

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

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

So

$$\frac{\partial S}{\partial \beta_k} \bigg|_{\beta = \hat{\beta}} = 0, \quad k = 1, \ldots, p.$$

So

$$-2\sum_{i=1}^{n}x_i(Y_i - \sum_{j=1}^{p}x_{ij}\hat{\beta_j}) = 0, \quad k = 1, \ldots, p.$$

i.e.

$$\sum_{i=1}^{n}x_i\sum_{j=1}^{p}x_{ij}\hat{\beta_j} = \sum_{i=1}^{n}x_iY_i, \quad k = 1, \ldots, p.$$

In matrix form,

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

the least squares equation.
Recall we assume $X$ is of full rank $p$.

This implies
\[ t^TX^Xt = (Xt)^T(Xt) = ||Xt||^2 > 0 \]
for $t \neq 0$ in $\mathbb{R}^p$.

i.e. $X^TX$ is positive definite, and hence has an inverse.

Hence
\[ \hat{\beta} = (X^TX)^{-1}X^TY \]
which is linear in the $Y_i$’s.

We also have that
\[ \mathbb{E} (\hat{\beta}) = (X^TX)^{-1}X^T\mathbb{E}(Y) = (X^TX)^{-1}X^T\beta = \beta \]
so $\hat{\beta}$ is unbiased for $\beta$.

And
\[ \text{cov}(\hat{\beta}) = (X^TX)^{-1}X^T\text{cov}(Y)X(X^TX)^{-1} = (X^TX)^{-1}\sigma^2 \]

since $\text{cov}(Y) = \sigma^2 I$.

In matrix form, $X = \begin{pmatrix} 1 & (x_1 - \bar{x}) \\ \vdots & \vdots \\ 1 & (x_n - \bar{x}) \end{pmatrix}$, so that $X^TX = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$,

where $S_{xx} = \sum_i (x_i - \bar{x})^2$.

Hence
\[ (X^TX)^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix}, \]
so that
\[ \hat{\beta} = (X^TX)^{-1}X^TY = \begin{pmatrix} \bar{Y} & 0 \\ 0 & \frac{1}{S_{xx}} \end{pmatrix}, \]

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

\[ \text{Simple linear regression using standardised } x \text{'s} \]

The model
\[ Y_i = a + bx_i + \epsilon_i, \quad i = 1, \ldots, n, \]
can be reparametrised to
\[ Y_i = a' + b(x_i - \bar{x}) + \epsilon_i, \quad i = 1, \ldots, n, \]
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.

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), $\text{cov}(\hat{\beta}) = (X^TX)^{-1}\sigma^2$, and so
\[ \text{var}(\hat{a'}) = \text{var}(\bar{y}) = \frac{\sigma^2}{n}; \quad \text{var}(\hat{b}) = \frac{\sigma^2}{S_{xx}}; \]

These estimators are uncorrelated.

All these results are obtained without any explicit distributional assumptions.
Example 12.1 continued

\( n = 24, \hat{\beta}' = \bar{y} = 826.5 \).
\( S_{xx} = 783.5 = 28.0^2, S_{xy} = -10077, S_{yy} = 444^2, r = -0.81, \hat{b} = -12.9. \)

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

\begin{align*}
S_{xx} &= 783.5 = 28.0^2, S_{xy} = -10077, S_{yy} = 444^2, r = -0.81, \hat{b} = -12.9.
\end{align*}

\[ \text{Example 12.1 continued} \]

\[ n = 24, \hat{\beta}' = \bar{y} = 826.5. \]
\[ S_{xx} = 783.5 = 28.0^2, S_{xy} = -10077, S_{yy} = 444^2, r = -0.81, \hat{b} = -12.9. \]

\[ \text{Line goes through } (\bar{x}, \bar{y}). \]

\[ \begin{align*}
\text{Example 12.1 continued} \quad n &= 24, \hat{\beta}' = \bar{y} = 826.5. \\
S_{xx} &= 783.5 = 28.0^2, S_{xy} = -10077, S_{yy} = 444^2, r = -0.81, \hat{b} = -12.9. \\
\text{Line goes through } (\bar{x}, \bar{y}). \\
\end{align*} \]

\[ \text{Example 12.1 continued} \]

\[ n = 24, \hat{\beta}' = \bar{y} = 826.5. \]
\[ S_{xx} = 783.5 = 28.0^2, S_{xy} = -10077, S_{yy} = 444^2, r = -0.81, \hat{b} = -12.9. \]

\[ \text{Line goes through } (\bar{x}, \bar{y}). \]
Example 13.1

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

Let \( Y_{i,j} \) be the resistivity of the \( j \)th wafer measured by instrument \( i \), where \( i,j = 1, \ldots, 5 \).

A possible model is, for \( i,j = 1, \ldots, 5 \).

\[
Y_{i,j} = \mu_i + \varepsilon_{i,j},
\]

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

This can be written in matrix form: Let

\[
\begin{align*}
Y_{25 \times 1} & = \begin{pmatrix}
Y_{1,1} \\
Y_{1,5} \\
Y_{2,1} \\
\vdots \\
Y_{6,5}
\end{pmatrix},
X_{25 \times 5} & = \begin{pmatrix}
1 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{pmatrix}, \\
\beta_{5 \times 1} & = \begin{pmatrix}
\mu_1 \\
\mu_2 \\
\mu_3 \\
\mu_4 \\
\mu_5
\end{pmatrix}, \\
\varepsilon_{25 \times 1} & = \begin{pmatrix}
\varepsilon_{1,1} \\
\varepsilon_{1,5} \\
\varepsilon_{2,1} \\
\vdots \\
\varepsilon_{5,5}
\end{pmatrix},
\end{align*}
\]

Then

\[
Y = X\beta + \varepsilon.
\]
For these data, \( \hat{\text{RSS}} = ( \hat{Y}_1 - \bar{Y}_1 )^2 \)
so that
\[
\text{Hence } (X^TX)^{-1} = \begin{pmatrix}
\frac{1}{1^2} & 0 & \cdots & 0 \\
0 & \frac{1}{2^2} & \cdots & 0 \\
\vdots & \cdots & \vdots & \vdots \\
0 & 0 & \cdots & \frac{1}{5^2}
\end{pmatrix},
\]
so that
\[
\hat{\mu} = (X^TX)^{-1}X^TY = \begin{pmatrix}
\frac{\bar{Y}_1}{5} \\
\vdots \\
\frac{\bar{Y}_5}{5}
\end{pmatrix}
\]
\[
\text{RSS} = \sum_{i=1}^5 \sum_{j=1}^5 (Y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^5 \sum_{j=1}^5 (Y_{ij} - \bar{Y}_i)^2 \text{ on } n-p = 25 - 5 = 20\text{ degrees of freedom.}
\]
For these data, \( \hat{\sigma} = \sqrt{\text{RSS}/(n-p)} = \sqrt{2170/20} = 10.4. \)

**Assuming normality**

- We now make a Normal assumption
  \[
  Y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \sigma^2 I), \quad \text{rank } (X) = p(< n).
  \]
- This is a special case of the linear model of Lecture 12, so all results hold.
- Since \( Y \sim N_n(X\beta, \sigma^2 I) \), the log-likelihood is
  \[
  \ell(\beta, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} S(\beta),
  \]
  where \( S(\beta) = (Y - X\beta)^T(Y - X\beta) \).
- Maximising \( \ell \) wrt \( \beta \) is equivalent to minimising \( S(\beta) \), so MLE is
  \[
  \hat{\beta} = (X^TX)^{-1}X^TY,
  \]
  the same as for least squares.

**Lemma 13.2**

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

(ii) For a symmetric idempotent matrix \( A \), \( \text{rank } (A) = \text{trace } (A) \).

**Proof:**

- (i) \( A^2 = A \) since idempotent, and so eigenvalues of \( A \) are \( \lambda_i \in \{0, 1\}, i = 1, \ldots, n \) \( \lambda_i x = Ax = A^2 x = \lambda_i^2 x \).
- \( A \) is also symmetric, and so there exists an orthogonal \( Q \) such that
  \[
  Q^T A Q = \text{diag } (\lambda_1, \ldots, \lambda_n) = \text{diag } (1, 1, \ldots, 0) = \Lambda \text{ (say)}.
  \]
- Let \( \mathbf{W} = Q^T Z \), and so \( Z = Q \mathbf{W} \). Then \( \mathbf{W} \sim \mathcal{N}_n(0, \sigma^2 I) \) by Proposition 11.1(i) \( \text{ (since } \text{cov} (\mathbf{W}) = Q^T \sigma^2 I Q = \sigma^2 I) \).
- Then
  \[
  Z^T A Z = \mathbf{W}^T Q^T A Q \mathbf{W} = \mathbf{W}^T \Lambda \mathbf{W} = \sum_{i=1}^r w_i^2 \sim \chi^2_r,
  \]
  from the definition of \( \chi^2 \).
Lecture 13. Linear models with normal assumptions

(ii) We can apply Lemma 13.2(i) with \( Z = Y - X\hat{\beta} \sim N_p(0, \sigma^2I_n) \) and \( A = (I_n - P) \), where \( P = X(X'X)^{-1}X' \) is the projection matrix covered after Definition 12.3.

(P is also known as the 'hat' matrix since it projects from the observation \( Y \) onto the fitted values \( \hat{Y} \).)

\( P \) is symmetric and idempotent, so \( I_n - P \) is also symmetric and idempotent (check).

By Lemma 13.2(ii),

\[
\text{rank}(P) = \text{trace}(P) = \text{trace}(X(X'X)^{-1}X') = \text{trace}((X'X)^{-1}X'X) = p.
\]

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

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

\[
Z'AZ = (Y - X\beta)'(I_n - P)(Y - X\beta) = \hat{Y}'(I_n - P)\hat{Y} \text{ since } (I_n - P)X = 0.
\]

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

\[
\text{RSS} = \hat{R}'\hat{R} = \hat{Y}'(I_n - P)\hat{Y} = (Z'AZ).
\]

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

\[\Box\]

Theorem 13.3

For the normal linear model \( Y \sim N_p(X\beta, \sigma^2I) \),

(i) \( \beta \sim N_p(\beta, \sigma^2(X'X)^{-1}) \).

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

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

Proof:

(i) \( \hat{\beta} = (X'X)^{-1}X'y \), say \( CY \).

Then from Proposition 11.1(i), \( \hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1}) \).

(ii) Let \( V = \left( \begin{array}{cc} \hat{\beta} \\
R \end{array} \right) = DY \), where \( D = \left( \begin{array}{c} C \\
I_n - P \end{array} \right) \) is a \( (p+n) \times n \) matrix.

By Proposition 11.1(i), \( V \) is multivariate normal with

\[
\text{cov}(V) = \sigma^2DD' = \sigma^2\left( \begin{array}{cc} CC' & C(I_n - P)' \\
(I_n - P)C & (I_n - P)(I_n - P)' \end{array} \right).
\]

We have \( C(I_n - P) = 0 \) (check) \( [X'X]^{-1}X'(I_n - P) = 0 \) because \( (I_n - P)X = 0 \).

Hence \( \hat{\beta} \) and \( R \) are independent by Proposition 11.2(ii).

Hence \( \hat{\beta} \) and \( \text{RSS} = R'R \) are independent, and so \( \hat{\beta} \) and \( \hat{\sigma}^2 \) are independent.

\[\Box\]

From (ii), \( \mathbb{E}(\text{RSS}) = \sigma^2(n-p) \), and so \( \hat{\sigma}^2 = \frac{\text{RSS}}{n-p} \) is an unbiased estimator of \( \sigma^2 \).

\( \hat{\sigma} \) 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

\[
\hat{\sigma}^2 = \frac{\text{RSS}}{n - p} = \frac{67968}{24 - 2} = 3089.
\]

Residual standard error is \(\hat{\sigma} = \sqrt{3089} = 55.6\) 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 \(\chi^2_1\) and \(\chi^2_n\) variables.
Lecture 14. Applications of the distribution theory

14. Applications of the distribution theory

Inference for $\beta$

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

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

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

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

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

Then $\hat{\beta}_j - \beta_j / \text{s.e.}(\hat{\beta}_j) \sim t_{n-p}$.

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

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 $\varepsilon_i, i = 1, \ldots, n$ are iid $N(0, \sigma^2)$.

Then from Lecture 12 and Theorem 13.3 we have that

$$\hat{a}' \sim N \left( a', \frac{\sigma^2}{n} \right), \quad \hat{b} = \frac{S_{xy}}{S_{xx}} \sim N \left( b, \frac{\sigma^2}{S_{xx}} \right),$$

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

and $(\hat{a}', \hat{b})$ and $\hat{\sigma}^2 = \text{RSS}/n$ are independent.
Example 12.1 continued

- We have seen that \( \hat{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{6968}{24} = 3089 = 55.6^2. \)
- So the standard error of \( \hat{b} \) is \( \text{s.e.}(\hat{b}) = \sqrt{\hat{\sigma}^2 (X^TX)^{-1}} = \sqrt{\frac{3089}{5280}} = 55.6 \times 28.0 = 1.99. \)
- So a 95\% interval for \( \hat{b} \) has endpoints \( \hat{b} \pm 1.96 \times 1.99 = -12.9 \pm 1.99 + 2.07 = (-12.9, -8.8), \) 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 \( \frac{\hat{b}}{\text{s.e.}(\hat{b})} = -12.9 -1.99 = -6.48, \) and we would reject \( H_0 \) since this is less than \( -1.96 \times 2.07 = -2.07. \)

```
(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 \( x^* \)**

- Let \( x^* \) be a new vector of values for the explanatory variables
- The expected response at \( x^* \) is \( \mathbb{E}(Y|x^*) = x^T \beta. \)
- We estimate this by \( x^T \hat{\beta}. \)
- By Theorem 13.3 and Proposition 11.1(i),
  \[
  x^T(\hat{\beta} - \beta) \sim N(0, \sigma^2 x^T(X^TX)^{-1}x^*).
  \]
- Let \( \tau^2 = x^T(X^TX)^{-1}x^*. \)
- Then
  \[
  \frac{x^T(\hat{\beta} - \beta)}{\sigma \tau} \sim t_{n-p}.
  \]
- A 100(1-\(\alpha\))\% confidence interval for the expected response \( x^T \beta \) has endpoints
  \[
  x^T \hat{\beta} \pm \sigma \tau t_{p-n}(\frac{\alpha}{2}).
  \]

```
 Lines(oxy, pred[, "upr"], lty = "dotted")
 Lines(oxy, pred[, "lwr"], lty = "dotted")
 Lines(oxy, pred[, "fit"])
```

---

**95\% CI for fitted line**

![Graph showing 95% CI for fitted line](image)

- Time for 2 miles: 700 to 1000
- Oxygen uptake: 40 to 60
- CI for \( x^T \beta \): 738.6 to 832.2
Predicted response at \( x^* \)

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

We therefore find that
  \[
  \frac{\hat{Y}^* - Y^*}{\hat{\sigma} \sqrt{\tau^2 + 1}} \sim t_{n-p}.
  \]

So the interval with endpoints
  \[
  x^T \hat{\beta} \pm \hat{\sigma} \sqrt{(\tau^2 + 1)} \ t_{n-p}(\frac{\alpha}{2})
  \]

is a 95% prediction interval for \( Y^* \).

Example 12.1 continued

A 95% prediction interval for \( Y^* \) at \( x^T = (1, (50 - \bar{x})) \) is
  \[
  x^T \hat{\beta} \pm \hat{\sigma} \sqrt{(\tau^2 + 1)} \ t_{n-p}(\frac{\alpha}{2}) = 808.5 \pm 55.6 \times 1.02 \times 2.07 = (691.1, 925.8).
  \]

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 \( x^T = (1, 0, \ldots, 0) \).
- The estimated expected response at \( x^T \) is
  \[
  x^T \hat{\mu} = \hat{\mu}_1 = \bar{Y}_1 = 124.3
  \]
- We find \( \tau^2 = x^T (X^T X)^{-1} x^* = \frac{1}{5} \).
- So a 95% CI for \( \mathbb{E}(Y_{1j}) \) is \( x^T \hat{\mu} \pm \hat{\sigma} \tau \ t_{n-p}(\frac{\alpha}{2}) \)
  \[
  = 124.3 \pm 10.4/\sqrt{5} \times 2.09 = 124.3 \pm 4.66 \times 2.09 = (114.6, 134.0).
  \]
- Note that we are using an estimate of \( \sigma \) obtained from all five instruments. If we had only used the data from the first instrument, \( \sigma \) would be estimated as \( \hat{\sigma}_1 = \sqrt{\sum_{j=1}^5 (Y_{1j} - \bar{Y}_1)^2/(5 - 1)} = 8.74 \).
- The observed 95% confidence interval for \( \mu_1 \) would have been
  \[
  \bar{Y}_1 \pm \frac{\hat{\sigma}}{\sqrt{5}} \ t_{n-p}(\frac{\alpha}{2}) = 124.3 \pm 3.91 \times 2.78 = (113.5, 135.1).
  \]
- The ‘pooled’ analysis gives a slightly narrower interval.
A 95% prediction interval for $Y_1 \ast$ at $x^T = (1, 0, \ldots, 0)$ is

$$x^T \hat{\mu} \pm \hat{\sigma} \sqrt{\left( \hat{\tau}^2 + 1 \right) t_{n-p} \left( \frac{\hat{\tau}}{2} \right)} = 124.3 \pm 10.42 \times 1.1 \times 2.07 = (100.5, 148.1).$$
Lecture 15. Hypothesis testing in the linear model

15. Hypothesis testing in the linear model

15.1. Preliminary lemma

Lemma 15.1
Suppose $Z \sim N_n(0, \sigma^2 I_n)$ and $A_1$ and $A_2$ and symmetric, idempotent $n \times n$ matrices with $A_1 A_2 = 0$. Then $Z^T A_1 Z$ and $Z^T A_2 Z$ are independent.

Proof:
Let $W_i = A_i Z$, $i = 1, 2$ and $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$, where $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$. By Proposition 11.1(i), $W \sim N_{2n} \begin{pmatrix} 0 \\ A \end{pmatrix}$ check.

So $W_1$ and $W_2$ are independent, which implies $W_1^T W_1 = Z^T A_1 Z$ and $W_2^T W_2 = Z^T A_2 Z$ are independent.

15.2. Hypothesis testing

Hypothesis testing

Suppose $X_{n \times p} = (X_0, X_1)$ 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 $H_0$, $Y = X_0 \beta_0 + \epsilon$.

Under $H_0$, MLEs of $\beta_0$ and $\sigma^2$ are

$\hat{\beta}_0 = (X_0^T X_0)^{-1} X_0^T Y$

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

and these are independent, by Theorem 13.3.

So fitted values under $H_0$ are

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

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

15.3. Geometric interpretation

Geometric interpretation

The figure illustrates the geometric interpretation of hypothesis testing in the linear model. The space spanned by the columns of $X$ is projected onto the unbiased estimator $\hat{Y} = X \hat{\beta} = P_Y$. The fitted values $\hat{Y}$ are then compared to the residual sum of squares (RSS).
Lecture 15. Hypothesis testing in the linear model

Generalised likelihood ratio test

- The generalised likelihood ratio test of $H_0$ against $H_1$ is

$$
\Lambda(Y) = \frac{1}{\sqrt{2\pi \sigma^2}} \frac{\rho^2}{\sigma^2} \exp \left( -\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) \right) \frac{1}{\sqrt{2\pi \sigma^2}} \frac{\rho^2}{\sigma^2} \exp \left( -\frac{1}{2\sigma^2}(Y - X\beta_0)^T(Y - X\beta_0) \right)
$$

$$
= \left( \frac{\rho}{\sigma} \right)^2 \left( \frac{\text{RSS}_0}{\text{RSS}} \right)^2 \left( 1 + \frac{\text{RSS}_0 - \text{RSS}}{\text{RSS}} \right)^2
$$

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

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

is approximately a $\chi^2_{p_0 - p_0}$ rv.

- But we can get an exact null distribution.

Null distribution of test statistic

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

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

- Now $I_n - P$ and $P - P_0$ are symmetric and idempotent, and therefore

$$
\text{rank}(I_n - P) = n - p, \quad \text{and} \quad \text{rank}(P - P_0) = \text{tr}(P - P_0) = \text{tr}(P) - \text{tr}(P_0) = \text{rank}(P) - \text{rank}(P_0) = p - p_0.
$$

- Also

$$
(Y^T(I_n - P)Y)^2 = (Y - X_0\beta_0)^T(I_n - P)(Y - X_0\beta_0) \text{ since } (I_n - P)X_0 = 0,
$$

$$
(Y^T(P - P_0)Y)^2 = (Y - X_0\beta_0)^T(P - P_0)(Y - X_0\beta_0) \text{ since } (P - P_0)X_0 = 0.
$$

Arrangement as an 'analysis of variance' table

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>degrees of freedom (df)</th>
<th>sum of squares</th>
<th>mean square</th>
<th>F statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted model</td>
<td>$p - p_0$</td>
<td>$\text{RSS}_0 - \text{RSS}$</td>
<td>$(\text{RSS}_0 - \text{RSS})/(p - p_0)$</td>
<td>$\text{RSS}_0/(p - p_0)$</td>
</tr>
<tr>
<td>Residual</td>
<td>$n - p$</td>
<td>$\text{RSS}$</td>
<td>$\text{RSS}/(n - p)$</td>
<td>$\text{RSS}$</td>
</tr>
</tbody>
</table>
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 \( \varepsilon_i, i = 1, \ldots, n \) 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 \( a' = \bar{Y} \), and the fitted values are \( \hat{Y}_i = \bar{Y} \).
- The observed RSS is therefore
  \[ RSS_0 = \sum_i (Y_i - \bar{Y})^2 = S_{yy}. \]
- The fitted sum of squares is therefore
  \[ RSS_0 - RSS = \sum_i \left( (Y_i - \bar{Y})^2 - (Y_i - \bar{Y} - b(x_i - \bar{x}))^2 \right) = b^2(x_i - \bar{x})^2 = b^2 S_{xx}. \]

Example 12.1 continued

As R code

```r
> 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 N(0, \( \sigma^2 \)) random variables, and the \( \mu_i \)'s are unknown constants.
- Fitting this model gives
  \[ RSS = \sum_{i=1}^I \sum_{j=1}^J (Y_{i,j} - \hat{\mu}_i)^2 = \sum_{i=1}^I \sum_{j=1}^J (Y_{i,j} - \bar{Y}_i)^2 \]
  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 \( Y_{i,j} \sim N(\mu, \sigma^2) \), and so \( \hat{\mu} = \bar{Y}_i \), and the fitted values are \( \hat{Y}_{i,j} = \bar{Y}_i \).
- The observed RSS is therefore
  \[ RSS_0 = \sum_j \sum_{y_{i,j} - \bar{Y}_i}^2. \]
The fitted sum of squares is therefore
\[
\text{RSS}_0 - \text{RSS} = \sum_i \sum_j ((y_{ij} - \bar{y}_{..})^2 - (y_{ij} - \bar{y}_{..})^2) = J \sum_j (\bar{y}_j - \bar{y}_{..})^2.
\]

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>sum of squares</th>
<th>mean square</th>
<th>F statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitted model</td>
<td>$I - 1$</td>
<td>$J \sum_j (\bar{y}<em>j - \bar{y}</em>{..})^2$</td>
<td>$\frac{J \sum_j (\bar{y}<em>j - \bar{y}</em>{..})^2}{(I-1)\sigma^2}$</td>
<td>$F = \frac{J \sum_j (\bar{y}<em>j - \bar{y}</em>{..})^2}{(I-1)\sigma^2}$</td>
</tr>
<tr>
<td>Residual</td>
<td>$n - (I-1)$</td>
<td>$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$</td>
<td>$\sigma^2$</td>
<td></td>
</tr>
</tbody>
</table>

Example 13.1
As R code

```r
> 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 $\sigma^2$ unknown.
- We wish to test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X \neq \mu_Y$.

Using the generalised likelihood ratio test
- $L_{X,Y}(H_0) = \sup_{\mu_X, \mu_Y} f_X(x | \mu_X, \sigma^2)f_Y(y | \mu_Y, \sigma^2)$.
- Under $H_0$ the mle's are
  \[
  \hat{\mu} = \frac{(mx + ny)}{(m + n)} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{m+n} \left( \sum(x_i - \hat{\mu})^2 + \sum(y_i - \hat{\mu})^2 \right) = \frac{1}{m+n} \left( S_{XX} + S_{YY} + \frac{mn}{m+n} (\bar{x} - \bar{y})^2 \right) \]

so
\[
L_{X,Y}(H_0) = \frac{(2\pi\hat{\sigma}^2)^{-\frac{m+n}{2}} e^{-\frac{1}{2\hat{\sigma}^2} \left( \sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2 \right)}}{(2\pi\hat{\sigma}^2)^{-\frac{m+n}{2}} e^{-\frac{1}{2\hat{\sigma}^2} \left( \sum(x_i - \bar{x})^2 \right)}}.
\]

- Under $H_0$, $\bar{X} \sim N(\mu, \frac{\sigma^2}{m})$, $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$ and so
  \[
  \frac{(\bar{X} - \bar{Y})}{\left( \sigma \sqrt{\frac{1}{m} + \frac{1}{n}} \right)} \sim N(0,1).
  \]
- 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 $\chi^2$ 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 $\alpha$ test is to reject $H_0$ if $|t| > t_{n+m-2}(\alpha/2)$.

- Similarly
  \[
  L_{x,y}(H_1) = \sup_{\mu_X, \mu_Y, \sigma^2} f_X(x | \mu_X, \sigma^2)f_Y(y | \mu_Y, \sigma^2) = (2\pi\hat{\sigma}^2)^{-\frac{m+n}{2}} e^{-\frac{1}{2\hat{\sigma}^2} \left( \sum(x_i - \bar{x})^2 \right)},
  \]

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

- Hence
  \[
  L_{x,y}(H_0; H_1) = \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} \right)^{\frac{m+n}{2}} = \left( 1 + \frac{mn}{m+n}(\bar{x} - \bar{y})^2 \right)^{\frac{m+n}{2}}.
  \]

- We reject $H_0$ if $mn(\bar{x} - \bar{y})^2/(S_{XX} + S_{YY})(m + n)$ is large, or equivalently if
  \[
  |t| = \frac{\bar{x} - \bar{y}}{\sqrt{S_{XX} + S_{YY} \left( \frac{1}{m} + \frac{1}{n} \right)}}
  \]
  is large.
16. Linear model examples, and ‘rules of thumb’

Lecture 16. Linear model examples, and ‘rules of thumb’

Paired observations

- Suppose the observations were paired: say because pairs of plants were randomly assigned.
  - We can introduce a parameter $\gamma_i$ for the ith pair, where $\sum_i \gamma_i = 0$, so that we assume
    $$X_i \sim N(\mu_X + \gamma_i, \sigma^2), \quad Y_i \sim N(\mu_Y + \gamma_i, \sigma^2), \quad i = 1, \ldots, n,$$
    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 $D_i = X_i - Y_i, \quad i = 1, \ldots, n, \quad D_i \sim N(\mu_X - \mu_Y, \phi^2)$, where $\phi^2 = 2\sigma^2$.
  - Thus $\overline{D} \sim N(\mu_X - \mu_Y, \phi^2/\overline{n})$, and we test $H_0: \mu_X - \mu_Y = 0$ by the t statistic
    $$t = \frac{\overline{D}}{\phi/\sqrt{n}},$$
    where $\overline{\phi^2} = S_{DD}/(n-1) = \sum_i (D_i - \overline{D})^2/(n-1)$, and $t \sim t_{n-1}$ distribution under $H_0$.

Example 16.1

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.

<table>
<thead>
<tr>
<th>Control</th>
<th>4.17</th>
<th>5.58</th>
<th>5.18</th>
<th>6.11</th>
<th>4.50</th>
<th>4.61</th>
<th>5.17</th>
<th>4.53</th>
<th>5.33</th>
<th>5.14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>4.81</td>
<td>4.17</td>
<td>4.41</td>
<td>3.59</td>
<td>5.87</td>
<td>3.83</td>
<td>6.03</td>
<td>4.89</td>
<td>4.32</td>
<td>4.69</td>
</tr>
</tbody>
</table>

- 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 $\bar{\phi}^2 = (S_{xx} + S_{yy})/(m + n - 2) = 0.485$.
- Then $|t| = (\bar{x} - \bar{y})/\sqrt{\bar{\phi}^2/(m + n)} = 1.19$.
- From tables $t_{0.025}(0.025) = 2.101$, 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.

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).

<table>
<thead>
<tr>
<th>Pair</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>4.17</td>
<td>5.58</td>
<td>5.18</td>
<td>6.11</td>
<td>4.50</td>
<td>4.61</td>
<td>5.17</td>
<td>4.53</td>
<td>5.33</td>
<td>5.14</td>
</tr>
<tr>
<td>Treatment</td>
<td>4.81</td>
<td>4.17</td>
<td>4.41</td>
<td>3.59</td>
<td>5.87</td>
<td>3.83</td>
<td>6.03</td>
<td>4.89</td>
<td>4.32</td>
<td>4.69</td>
</tr>
</tbody>
</table>

- Observed statistics are $\overline{D} = 0.37, S_{dd} = 12.54, n = 10$, so that $\overline{\phi} = \sqrt{S_{dd}/(n-1)} = \sqrt{2.33/9} = 1.18$.
- Thus $t = \overline{D}/\overline{\phi}/\sqrt{n} = 0.37/1.18\sqrt{10} = 0.99$.
- This can be compared to $t_{0.025}(0.025) = 2.262$ to show that we cannot reject $H_0: \overline{(D)} = 0$, 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.
  $$P(|t| > |t|/H_0) = 2P(t > |t|) = 2 \times 0.17 = 0.34.$$
16. Linear model examples

16.2. Paired observations

In R code:
```r
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
```

16.3. 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

$0.05 = P(X = 0 | p') = (1 - p')^n$.

• 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 $1 - e^{\log(0.05)/100} = 0.0295$.

16.4. 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 $E(X | H_0) = np_0$.
• Then the probability of the difference between observed and expected exceeding $\sqrt{n}$, given $H_0$ is true, is

$P(|X - np_0| > \sqrt{n} | H_0) = P \left( \frac{|X - np_0|}{\sqrt{np_0(1 - p_0)}} > \frac{1}{\sqrt{p_0(1 - p_0)}} | H_0 \right)$

$< P \left( \frac{|X - np_0|}{\sqrt{np_0(1 - p_0)}} > 2 | H_0 \right)$ since $\frac{1}{\sqrt{p_0(1 - p_0)}} > 2$

$\approx P(|Z| > 2)$

$\approx 0.05$
Lecture 16. Linear model examples, and ‘rules of thumb’

Rules of thumb: the 'rule of root n'*

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 \( 2 \times \Pr( X \geq 550 ) = 2 \times ( 1 - \Pr( X \leq 549 ) ) \), where \( X \sim \text{Binom}(1000, 0.5) \), which according to R is \( > 2 \times ( 1 - \text{pbinom}(549,1000,0.5) ) \)

\[ 0.001730536 \]

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 \( \sqrt{4 \times \text{expected}} \), reject \( H_0 \).

- We assume \( X \sim \text{Binom}(n, p) \), and \( H_0 : p = p_0 \), so the expected number of events is \( E(X|H_0) = np_0 \).
- 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.

Rules of thumb: non-overlapping confidence intervals*

Suppose we have 95% confidence intervals for \( y_1 \) and \( y_2 \), and 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 \( \overline{Y}_1 \sim N(\mu_1, s_1^2) \), \( \overline{Y}_2 \sim N(\mu_2, s_2^2) \), where \( s_1 \) and \( s_2 \) are known standard errors.
- Suppose wlg that \( \overline{Y}_1 > \overline{Y}_2 \). Then since \( \overline{Y}_1 - \overline{Y}_2 \sim N(\mu_1 - \mu_2, s_1^2 + s_2^2) \), 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

\[ \overline{Y}_1 - 1.96s_1 > \overline{Y}_2 + 1.96s_2 \] 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 $|\bar{y}_1 - \bar{y}_2| = 1.96 \times 2s = 3.92 \times s$.

So $p$-value =

$$P(|\bar{Y}_1 - \bar{Y}_2| > 3.92s) = P\left(\frac{|\bar{Y}_1 - \bar{Y}_2|}{\sqrt{2s}} > \frac{3.92}{\sqrt{2}}\right)$$

$$= P(|Z| > 2.77) = 2 \times P(Z > 2.77) = 0.0055.$$

And if 'just not touching' $100(1 - \alpha)\%$ CIs were to be equivalent to 'just rejecting $H_0$', then we would need to set $\alpha$ so that the critical difference between $\bar{y}_1 - \bar{y}_2$ was exactly the width of each of the CIs, and so

$$1.96 \times \sqrt{2} \times s = s \times \Phi^{-1}(1 - \frac{\alpha}{2}).$$

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.