# Prelims Statistics and Data Analysis, Lectures 1–10

Christl Donnelly

Trinity Term 2023 (version of 12-03-2022)

In some places these notes will contain more than the lectures and in other places they will contain less. I strongly encourage you to attend all of the lectures.

These notes are largely based on previous versions of the course, and I would like to make grateful acknowledgement to Neil Laws, Peter Donnelly and Bernard Silverman. Please send any comments or corrections to `christl.donnelly@stats.ox.ac.uk`.

TT 2023 updates:

None so far. If you spot any errors please let me know.

# Introduction

**Probability:** in probability we start with a probability model $P$, and we deduce properties of $P$.

E.g. Imagine flipping a coin 5 times. Assuming that flips are *fair* and *independent*, what is the probability of getting 2 heads?

**Statistics:** in statistics we have *data*, which we regard as having been generated from some *unknown* probability model $P$. We want to be able to say some useful things about $P$.

E.g. We flip a coin 1000 times and observe 530 heads. Is the coin fair? i.e. is the probability of heads greater than $\frac{1}{2}$? or could it be equal to $\frac{1}{2}$?

### Precision of estimation

Question: What is the average body temperature in humans?

Collect data: let $x_1, \ldots, x_n$ be the body temperatures of $n$ randomly chosen individuals, i.e. our *model* is that individuals are chosen independently from the general population.

Then estimate: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ could be our *estimate*, i.e. the *sample* average.

- How precise is $\overline{x}$ as an estimate of the population mean?

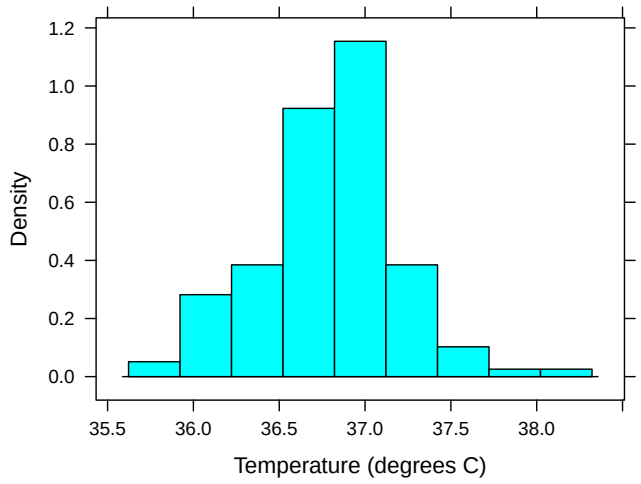


**Figure.** Histogram of body temperatures of 130 individuals. A description of the dataset is available here: https://rdrr.io/cran/UsingR/man/normtemp.html

### Relationships between observations

Let $y_i$ = price of house in month $x_i$, for $i = 1, \ldots, n$.

- Is it reasonable that $y_i = \alpha + \beta x_i +$ "error"?
- Is $\beta > 0$?
- How accurately can we estimate $\alpha$ and $\beta$, and how do we estimate them?

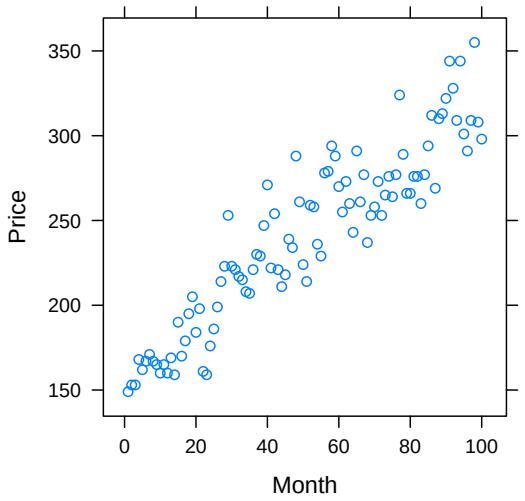


**Figure.** Scatterplot of some Oxford house price data.

### Notation and conventions

Lower case letters $x_1, \ldots, x_n$ denote *observations*: we regard these as the observed values of random variables $X_1, \ldots, X_n$.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$.

It is convenient to think of $x_i$ as

- the observed value of $X_i$
- or sometimes as a possible value that $X_i$ can take.

Since $x_i$ is a possible value for $X_i$ we can calculate probabilities, e.g. if $X_i \sim \text{Poisson}(\lambda)$ then

$$P(X_i = x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \quad x_i = 0, 1, \ldots.$$

# 1 Random Samples

**Definition.** A *random sample* of size $n$ is a set of random variables $X_1, \ldots, X_n$ which are independent and identically distributed (i.i.d.).

**Example.** Let $X_1, \ldots, X_n$ be a random sample from a Poisson($\lambda$) distribution.

E.g. $X_i$ = # traffic accidents on St Giles' in year $i$.

We'll write $f(\mathbf{x})$ to denote the joint probability mass function (p.m.f.) of $X_1, \ldots, X_n$. Then

$$\begin{aligned} f(\mathbf{x}) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) && \text{definition of joint p.m.f.} \\ &= P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n) && \text{by independence} \\ &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} && \text{since } X_i \text{ Poisson} \\ &= \frac{e^{-n\lambda}\ \lambda^{(\sum_{i=1}^n x_i)}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

**Example.** Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with probability density function (p.d.f.) given by

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geqslant 0.$$

E.g. $X_i$ = time until the first breakdown of machine $i$ in a factory.

We'll write $f(\mathbf{x})$ to denote the joint p.d.f. of $X_1, \ldots, X_n$. Then

$$\begin{aligned} f(\mathbf{x}) &= f(x_1)\ldots f(x_n) && \text{since the } X_i \text{ are independent} \\ &= \prod_{i=1}^n \frac{1}{\mu} e^{-x_i/\mu} \\ &= \frac{1}{\mu^n} \exp\left( -\frac{1}{\mu} \sum_{i=1}^n x_i \right). \end{aligned}$$

**Note.**

1. We use $f$ to denote a p.m.f. in the first example and to denote a p.d.f. in the second example. It is convenient to use the same letter (i.e. $f$) in both the discrete and continuous cases. (In introductory probability you may often see $p$ for p.m.f. and $f$ for p.d.f.)

   We could write $f_{X_i}(x_i)$ for the p.m.f./p.d.f. of $X_i$, and $f_{X_1,\ldots,X_n}(\mathbf{x})$ for the joint p.m.f./p.d.f. of $X_, \ldots, X_n$. However it is convenient to keep things simpler by omitting subscripts on $f$.

2. In the second example $E(X_i) = \mu$ and we say "$X_i$ has an exponential distribution with mean $\mu$" (i.e. expectation $\mu$).

Sometimes, and often in probability, we work with "an exponential distribution with parameter $\lambda$" where $\lambda = 1/\mu$. To change the parameter from $\mu$ to $\lambda$ all we do is replace $\mu$ by $1/\lambda$ to get

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

Sometimes (often in statistics) we parametrise the distribution using $\mu$, sometimes (often in probability) we parametrise it using $\lambda$.

In *probability* we assume that the parameters $\lambda$ and $\mu$ in our two examples are known. In *statistics* we wish to estimate $\lambda$ and $\mu$ from data.

- What is the best way to estimate them? And what does "best" mean?
- For a given method, how precise is the estimation?

## 2 Summary Statistics

The expected value of $X$, $\mathrm{E}(X)$, is also called its mean. This is often denoted $\mu$.

The variance of $X$, $\mathrm{var}(X)$, is $\mathrm{var}(X) = E[(X-\mu)^2]$. This is often denoted $\sigma^2$. The standard deviation of $X$ is $\sigma$.

**Definition.** Let $X_1, \ldots, X_n$ be a random sample. The *sample mean* $\overline{X}$ is defined by

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

The *sample variance* $S^2$ is defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The *sample standard deviation* is $S = \sqrt{\text{sample variance}}$.

**Note.**

1. The denominator in the definition of $S^2$ is $n-1$, not $n$.
2. $\overline{X}$ and $S^2$ are random variables. Their distributions are called the *sampling distributions* of $\overline{X}$ and $S^2$.
3. Given observations $x_1, , \ldots, x_n$ we can compute the *observed values* $\overline{x}$ and $s^2$.
4. The sample mean $\overline{x}$ is a summary of the *location* of the sample.
5. The sample variance $s^2$ (or the sample standard deviation $s$) is a summary of the *spread* of the sample about $\overline{x}$.

### The Normal Distribution

The random variable $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$, written $X \sim N(\mu, \sigma^2)$, if the p.d.f. of $X$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

Properties: $E(X) = \mu$, $\mathrm{var}(X) = \sigma^2$.

### Standard Normal Distribution

If $\mu = 0$ and $\sigma^2 = 1$, then $X \sim N(0,1)$ is said to have a *standard normal distribution*. Properties:

- If $Z \sim N(0,1)$ and $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$ and $Z = (X-\mu)/\sigma$, then $Z \sim N(0,1)$.

- The cumulative distribution function (c.d.f.) of the *standard* normal distribution is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du.$$

This cannot be written in a closed form, but can be found by numerical integration to an arbitrary degree of accuracy.

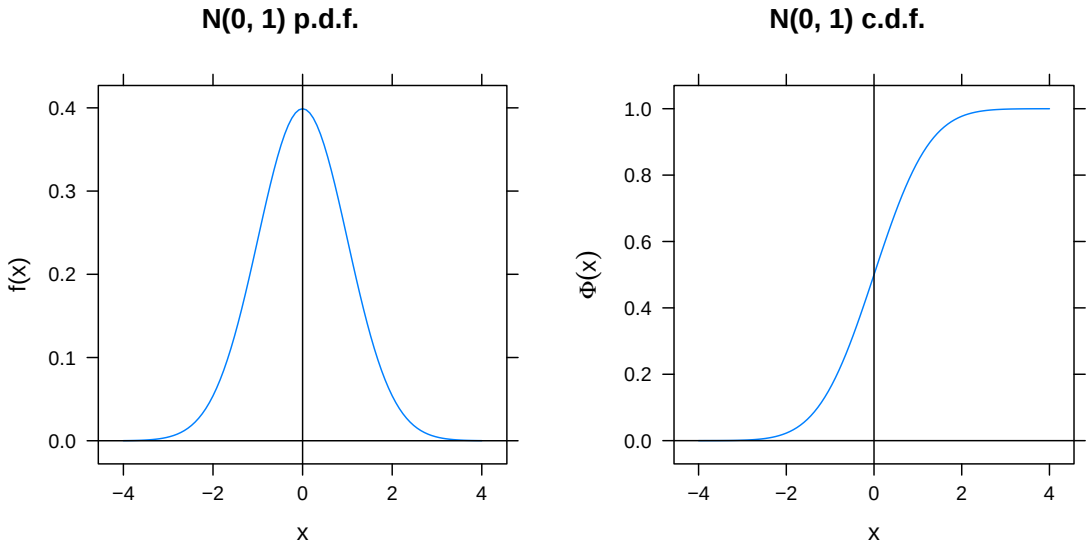


**Figure 2.1.** Standard normal p.d.f. and c.d.f.

# 3 Maximum Likelihood Estimation

Maximum likelihood estimation is a general method for estimating unknown parameters from data. This turns out to be the method of choice in many contexts, though this isn't obvious at this stage.

**Example.** Suppose e.g. that $x_1, \ldots, x_n$ are time intervals between major earthquakes. Assume these are observations of $X_1, \ldots, X_n$ independently drawn from an exponential distribution with mean $\mu$, so that each $X_i$ has p.d.f.

$$f(x; \mu) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geqslant 0.$$

We have written $f(x; \mu)$ to indicate that the p.d.f. $f$ depends on $\mu$.

- How do we estimate $\mu$?
- Is $\overline{X}$ a good estimator for $\mu$?
- Is there a general principle we can use?

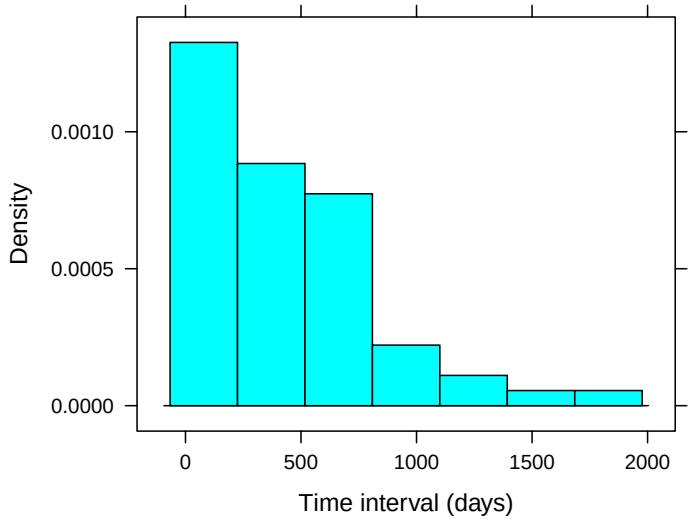


**Figure 3.1.** Histogram of time intervals between 62 major earthquakes 1902–77: an exponential density looks plausible. (There are more effective ways than a histogram to check if a particular distribution is appropriate, i.e. to check the exponential assumption in this case – see Part A Statistics.)

In general we write $f(x; \theta)$ to indicate that the p.d.f./p.m.f. $f$, which is a function of $x$, depends on a parameter $\theta$. Similarly, $f(\mathbf{x}; \theta)$ denotes the joint p.d.f./p.m.f. of $\mathbf{X}$. (Sometimes $f(x; \theta)$ is written $f(x \mid \theta)$.)

The parameter $\theta$ may be a vector, e.g. $\theta = (\mu, \sigma^2)$ in the earlier $N(\mu, \sigma^2)$ example.

If we regard $\theta$ as unknown, then we need to estimate it using $x_1, \ldots, x_n$.

**Definition.** Let $X_1, \ldots, X_n$ have joint p.d.f./p.m.f. $f(\mathbf{x}; \theta)$. Given observed values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$, the *likelihood* of $\theta$ is the function

$$L(\theta) = f(\mathbf{x}; \theta). \tag{3.1}$$

The *log-likelihood* is $\ell(\theta) = \log L(\theta)$.

So $L(\theta)$ is the joint p.d.f./p.m.f. of the observed data *regarded as a function of* $\theta$, for fixed $\mathbf{x}$.

In the definition of $\ell(\theta)$, log means log to base $e$, i.e. log = $\log_e$ = ln.

Often we assume that $X_1, \ldots, X_n$ are a random sample from $f(x; \theta)$, so that

$$\begin{aligned} L(\theta) &= f(\mathbf{x}; \theta) \\ &= \prod_{i=1}^{n} f(x_i; \theta) \qquad \text{since the } X_i \text{ are independent.} \end{aligned}$$

Sometimes we have independent $X_i$ whose distributions differ, say $X_i$ is from $f_i(x; \theta)$. Then the likelihood is

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i; \theta).$$

**Definition.** The *maximum likelihood estimate* (MLE) $\widehat{\theta}(\mathbf{x})$ is the value of $\theta$ that maximises $L(\theta)$ for the given $\mathbf{x}$.

The idea of maximum likelihood is to estimate the parameter by the value of $\theta$ that gives the greatest likelihood to observations $x_1, \ldots, x_n$. That is, the $\theta$ for which the probability or probability density (3.1) is maximised.

Since taking logs is monotone, $\widehat{\theta}(\mathbf{x})$ also maximises $\ell(\theta)$. Finding the MLE by maximising $\ell(\theta)$ is often more convenient.

**Example** (continued)**.** In our exponential mean $\mu$ example, the parameter is $\mu$ and

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu} = \mu^{-n} \exp\left( -\frac{1}{\mu} \sum_{i=1}^{n} x_i \right)$$

$$\ell(\mu) = \log L(\mu) = -n \log \mu - \frac{1}{\mu} \sum_{i=1}^{n} x_i.$$

To find the maximum

$$\frac{d\ell}{d\mu} = -\frac{n}{\mu} + \frac{\sum_{i=1}^{n} x_i}{\mu^2}.$$

So

$$\frac{d\ell}{d\mu} = 0 \iff \frac{n}{\mu} = \frac{\sum_{i=1}^{n} x_i}{\mu^2} \iff \mu = \overline{x}.$$

This is a maximum since

$$\left.\frac{d^2\ell}{d\mu^2}\right|_{\mu=\overline{x}} = \frac{n}{\overline{x}^2} - \frac{2\sum_{i=1}^{n} x_i}{\overline{x}^3} = -\frac{n}{\overline{x}^2} < 0.$$

So the MLE is $\widehat{\mu}(\mathbf{x}) = \overline{x}$. Often we'll just write $\widehat{\mu} = \overline{x}$.

In this case the maximum likelihood *estimator* of $\mu$ is $\widehat{\mu}(\mathbf{X}) = \overline{X}$, which is a random variable. (More on the difference between estimates and estimators later.)

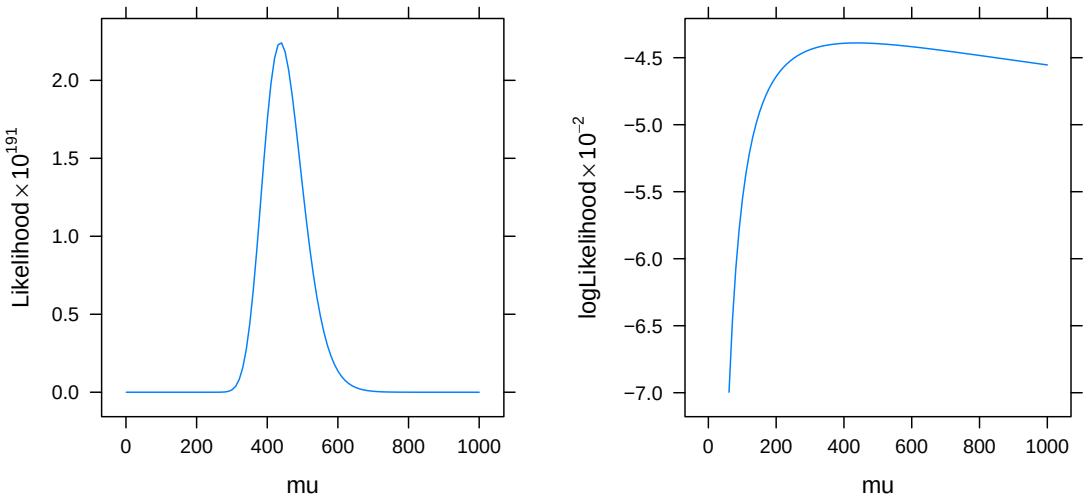


**Figure 3.2.** Likelihood $L(\mu)$ and log-likelihood $\ell(\mu)$ for exponential (earthquake) example.

**Definition.** If $\widehat{\theta}(\mathbf{x})$ is the maximum likelihood estimate of $\theta$, then the *maximum likelihood estimator* (MLE) is defined by $\widehat{\theta}(\mathbf{X})$.

Note: both maximum likelihood *estimate* and maximum likelihood *estimator* are often abbreviated to MLE.

**Example** (Opinion poll)**.** Suppose $n$ individuals are drawn independently from a large population. Let

$$X_i = \begin{cases} 1 & \text{if individual } i \text{ is a Labour voter} \\ 0 & \text{otherwise.} \end{cases}$$

Let $p$ be the proportion of Labour voters, so that

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p.$$

This is a Bernoulli distribution, for which the p.m.f. can be written

$$f(x; p) = P(X = x) = p^x (1-p)^{1-x}, \quad x = 0, 1.$$

The likelihood is

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \\ &= p^r (1-p)^{n-r} \end{aligned}$$

where $r = \sum_{i=1}^{n} x_i$. So the log-likelihood is

$$\begin{aligned} \ell(p) &= \log L(p) \\ &= r \log p + (n-r) \log(1-p) \end{aligned}$$

For a maximum, differentiate and set to zero:

$$\frac{r}{p} - \frac{n-r}{1-p} = 0 \iff \frac{r}{p} = \frac{n-r}{1-p} \iff r - rp = np - rp$$

and so $p = r/n$. This is a maximum since $\ell''(p) < 0$.

So $\widehat{p} = r/n$, i.e. the MLE is $\widehat{p} = \sum_{i=1}^{n} X_i/n$ which is the proportion of Labour voters in the sample.

**Example** (Genetics)**.** Suppose we test randomly chosen individuals at a particular locus on the genome. Each chromosome can be type $A$ or $a$. Every individual has two chromosomes (one from each parent), so the genotype can be $AA$, $Aa$ or $aa$. (Note that order is not relevant, there is no distinction between $Aa$ and $aA$.)

Hardy-Weinberg law: under plausible assumptions,

$$P(AA) = p_1 = \theta^2, \quad P(Aa) = p_2 = 2\theta(1-\theta), \quad P(aa) = p_3 = (1-\theta)^2$$

for some $\theta$ with $0 \leqslant \theta \leqslant 1$.

Now suppose the random sample of $n$ individuals contains:

$$x_1 \text{ of type } AA, \quad x_2 \text{ of type } Aa, \quad x_3 \text{ of type } aa$$

where $x_1 + x_2 + x_3 = n$ and these are observations of $X_1, X_2, X_3$. Then

$$\begin{aligned} L(\theta) &= P(X_1 = x_1, X_2 = x_2, X_3 = x_3) \\ &= \frac{n!}{x_1!\, x_2!\, x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3} \\ &= \frac{n!}{x_1!\, x_2!\, x_3!} (\theta^2)^{x_1} [2\theta(1-\theta)]^{x_2} [(1-\theta)^2]^{x_3}. \end{aligned}$$

This is a *multinomial distribution*.

So

$$\begin{aligned} \ell(\theta) &= \text{constant} + 2x_1 \log\theta + x_2[\log 2 + \log\theta + \log(1-\theta)] + 2x_3 \log(1-\theta) \\ &= \text{constant} + (2x_1 + x_2)\log\theta + (x_2 + 2x_3)\log(1-\theta) \end{aligned}$$

(where the constants depend on $x_1, x_2, x_3$ but not $\theta$).

Then $\ell'(\widehat{\theta}) = 0$ gives

$$\frac{2x_1 + x_2}{\widehat{\theta}} - \frac{x_2 + 2x_3}{1 - \widehat{\theta}} = 0$$

or

$$(2x_1 + x_2)(1 - \widehat{\theta}) = (x_2 + 2x_3)\widehat{\theta}.$$

So

$$\widehat{\theta} = \frac{2x_1 + x_2}{2(x_1 + x_2 + x_3)} = \frac{2x_1 + x_2}{2n}.$$

Steps:

- Write down the (log) likelihood

- Find the maximum (usually by differentiation, but not quite always)
- Rearrange to give the parameter estimate in terms of the data.

**Example** (Estimating multiple parameters)**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown.

[Here $\overset{\text{iid}}{\sim}$ means "are independent and identically distributed as."]

The likelihood is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

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

with log-likelihood

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

We maximise $\ell$ jointly over $\mu$ and $\sigma^2$:

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

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

and solving $\frac{\partial \ell}{\partial \mu} = 0$ and $\frac{\partial \ell}{\partial (\sigma^2)} = 0$ simultaneously we obtain

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Hence: the MLE of $\mu$ is the sample mean, but the MLE of $\sigma^2$ is $(n-1)s^2/n$. (More later.)

**"Estimate" and "estimator"**

Estimator:

- A rule for constructing an estimate.
- A function of the random variables $\mathbf{X}$ involved in the random sample.
- Itself a random variable.

Estimate:

- The numerical value of the estimator for the particular data set.
- The value of the function evaluated at the data $x_1, \ldots, x_n$.

# 4 Parameter Estimation

Recall the earthquake example:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential distribution, mean } \mu.$$

Possible estimators of $\mu$:

- $\overline{X}$
- $\frac{1}{3}X_1 + \frac{2}{3}X_2$
- $X_1 + X_2 - X_3$
- $\frac{2}{n(n+1)}(X_1 + 2X_2 + \cdots + nX_n)$.

How should we choose?

In general, suppose $X_1, \ldots, X_n$ is a random sample from a distribution with p.d.f./p.m.f. $f(x;\theta)$. We want to estimate $\theta$ from observations $x_1, \ldots, x_n$.

**Definition.** A *statistic* is any function $T(\mathbf{X})$ of $X_1, \ldots, X_n$ that does not depend on $\theta$.

An *estimator* of $\theta$ is any statistic $T(\mathbf{X})$ that we might use to estimate $\theta$.

$T(\mathbf{x})$ is the *estimate* of $\theta$ obtained via $T$ from observed values $\mathbf{x}$.

**Note.** $T(\mathbf{X})$ is a random variable, e.g. $\overline{X}$.

$T(\mathbf{x})$ is a fixed number, based on data, e.g. $\overline{x}$.

We can choose between estimators by studying their properties. A good estimator should take values close to $\theta$.

**Definition.** The estimator $T = T(\mathbf{X})$ is said to be *unbiased* for $\theta$ if, whatever the true value of $\theta$, we have $E(T) = \theta$.

This means that "on average" $T$ is correct.

**Example** (Earthquakes)**.** Possible estimators $\overline{X}$, $\frac{1}{3}X_1 + \frac{2}{3}X_2$, etc.

Since $E(X_i) = \mu$, we have

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$E(\tfrac{1}{3}X_1 + \tfrac{2}{3}X_2) = \tfrac{1}{3}\mu + \tfrac{2}{3}\mu = \mu.$$

Similar calculations show that $X_1 + X_2 - X_3$ and $\frac{2}{n(n+1)}\sum_{j=1}^{n} jX_j$ are also unbiased.

**Example** (Normal variance)**.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, with $\mu$ and $\sigma^2$ unknown, and let $T = \frac{1}{n}\sum(X_i - \overline{X})^2$. Then $T$ is the MLE of $\sigma^2$. Is $T$ unbiased?

Let $Z_i = (X_i - \mu)/\sigma$. So the $Z_i$ are independent and $N(0,1)$, $E(Z_i) = 0$, $\text{var}(Z_i) = E(Z_i^2) = 1$.

$$
\begin{aligned}
&E[(X_i - \overline{X})^2] \\
&\quad = E[\sigma^2 (Z_i - \overline{Z})^2] \\
&\quad = \sigma^2 \,\text{var}(Z_i - \overline{Z}) \qquad \text{since } E(Z_i - \overline{Z}) = 0 \\
&\quad = \sigma^2 \,\text{var}\left( -\frac{1}{n} Z_1 - \frac{1}{n} Z_2 - \cdots + \frac{n-1}{n} Z_i + \cdots - \frac{1}{n} Z_n \right) \\
&\quad = \sigma^2 \left( \frac{1}{n^2} \,\text{var}(Z_1) + \frac{1}{n^2} \,\text{var}(Z_2) + \cdots + \frac{(n-1)^2}{n^2} \,\text{var}(Z_i) + \cdots + \frac{1}{n^2} \,\text{var}(Z_n) \right) \\
&\qquad \text{since } \text{var}(\sum a_i U_i) = \sum a_i^2 \,\text{var}(U_i) \text{ for indep } U_i \\
&\quad = \sigma^2 \left( (n-1) \times \frac{1}{n^2} + \frac{(n-1)^2}{n^2} \right) \\
&\quad = \frac{(n-1)}{n} \sigma^2.
\end{aligned}
$$

So

$$E(T) = \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \overline{X})^2] = \frac{(n-1)}{n} \sigma^2 < \sigma^2.$$

Hence $T$ is *not* unbiased, $T$ will *underestimate* $\sigma^2$ on average.

But

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} T$$

so the sample variance satisfies

$$E(S^2) = \frac{n}{n-1} E(T) = \sigma^2.$$

So $S^2$ is unbiased for $\sigma^2$.

**Example** (Uniform distribution – some unusual features!)**.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0, \theta]$, where $\theta > 0$, i.e.

$$f(x; \theta) = \begin{cases} \dfrac{1}{\theta} & \text{if } 0 \leqslant x \leqslant \theta \\ 0 & \text{otherwise.} \end{cases}$$

What is the MLE for $\theta$? Is the MLE unbiased?

Calculate the likelihood:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(x; \theta) \\
&= \begin{cases} \dfrac{1}{\theta^n} & \text{if } 0 \leqslant x_i \leqslant \theta \text{ for all } i \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} 0 & \text{if } 0 < \theta < \max x_i \\ \dfrac{1}{\theta^n} & \text{if } \theta \geqslant \max x_i. \end{cases}
\end{aligned}
$$

Note: $\theta \geqslant x_i$ for all $i \iff \theta \geqslant \max x_i$. (And $\max x_i$ means $\max_{1 \leqslant i \leqslant n} x_i$).

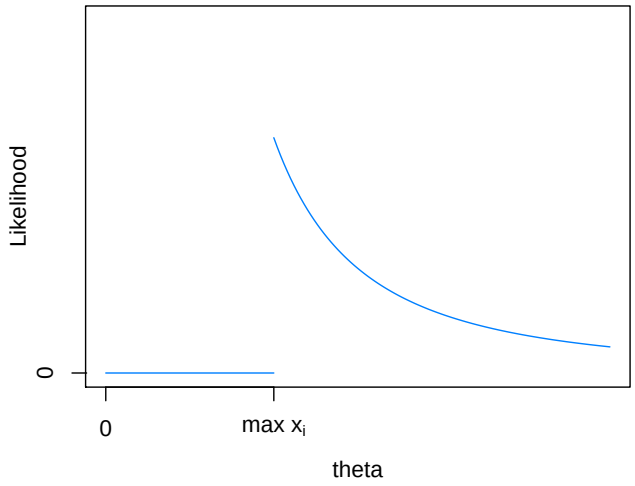


**Figure 4.1.** Likelihood $L(\theta)$ for Uniform$[0, \theta]$ example.

From the diagram:

- the max occurs at $\widehat{\theta} = \max x_i$
- this is *not* a point where $\ell'(\theta) = 0$
- taking logs doesn't help.

Consider the range of values of $x$ for which $f(x; \theta) > 0$, i.e. $0 \leqslant x \leqslant \theta$. The thing that makes this example different to our previous ones is that this range *depends* on $\theta$ (and we must take this into account because the likelihood is a *function* of $\theta$).

The MLE of $\theta$ is $\widehat{\theta} = \max X_i$. What is $E(\widehat{\theta})$?

Find the c.d.f. of $\widehat{\theta}$:

$$\begin{aligned} F(y) &= P(\widehat{\theta} \leqslant y) \\ &= P(\max X_i \leqslant y) \\ &= P(X_1 \leqslant y, X_2 \leqslant y, \ldots, X_n \leqslant y) \\ &= P(X_1 \leqslant y) P(X_2 \leqslant y) \ldots P(X_n \leqslant y) \qquad \text{since } X_i \text{ independent} \\ &= \begin{cases} (y/\theta)^n & \text{if } 0 \leqslant y \leqslant \theta \\ 1 & \text{if } y > \theta. \end{cases} \end{aligned}$$

So, differentiating the c.d.f., the p.d.f. is

$$f(y) = \frac{n y^{n-1}}{\theta^n}, \quad 0 \leqslant y \leqslant \theta.$$

So

$$E(\widehat{\theta}) = \int_0^\theta y \cdot \frac{n y^{n-1}}{\theta^n}\, dy$$
$$= \frac{n}{\theta^n} \int_0^\theta y^n\, dy$$
$$= \frac{n\theta}{n+1}.$$

So $\widehat{\theta}$ is *not* unbiased. But note that it *is* asymptotically unbiased: $E(\widehat{\theta}) \to \theta$ as $n \to \infty$.

In fact under mild assumptions MLEs are always asymptotically unbiased.

## 4.1 Further Properties of Estimators

**Definition.** The *mean squared error* (MSE) of an estimator $T$ is defined by

$$\text{MSE}(T) = E[(T-\theta)^2].$$

The *bias* $b(T)$ of $T$ is defined by

$$b(T) = E(T) - \theta.$$

**Note.**

1. Both MSE$(T)$ and $b(T)$ may depend on $\theta$.
2. MSE is a measure of the "distance" between $T$ and $\theta$, so is a good overall measure of performance.
3. $T$ is unbiased if $b(T) = 0$ for all $\theta$.

**Theorem 4.1.** $\text{MSE}(T) = \text{var}(T) + [b(T)]^2$.

**Proof.** Let $\mu = E(T)$. Then

$$\begin{aligned}
\text{MSE}(T) &= E[\{(T-\mu) + (\mu-\theta)\}^2] \\
&= E[(T-\mu)^2 + 2(\mu-\theta)(T-\mu) + (\mu-\theta)^2] \\
&= E[(T-\mu)^2] + 2(\mu-\theta)E[T-\mu] + (\mu-\theta)^2 \\
&= \text{var}(T) + 2(\mu-\theta) \times 0 + (\mu-\theta)^2 \\
&= \text{var}(T) + [b(T)]^2.
\end{aligned}$$

□

So an estimator with small MSE needs to have small variance *and* small bias. Unbiasedness alone is not particularly desirable – it is the combination of small variance and small bias which is important.

**Reminder**

Suppose $a_1, \ldots, a_n$ are constants. It is always the case that

$$E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + \cdots + a_nE(X_n).$$

If $X_1, \ldots, X_n$ are independent then

$$\operatorname{var}(a_1X_1 + \cdots + a_nX_n) = a_1^2 \operatorname{var}(X_1) + \cdots + a_n^2 \operatorname{var}(X_n).$$

In particular, if $X_1, \ldots, X_n$ is a random sample with $E(X_i) = \mu$ and $\operatorname{var}(X_i) = \sigma^2$, then

$$E(\overline{X}) = \mu \quad \text{and} \quad \operatorname{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

**Example** (Uniform distribution)**.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0, \theta]$, i.e.

$$f(x; \theta) = \begin{cases} \dfrac{1}{\theta} & \text{if } 0 \leqslant x \leqslant \theta \\ 0 & \text{otherwise.} \end{cases}$$

We will consider two estimators of $\theta$:

- $T = 2\overline{X}$, the natural estimator based on the sample mean (because the mean of the distribution is $\theta/2$)
- $\widehat{\theta} = \max X_i$, the MLE.

Now $E(T) = 2E(\overline{X}) = \theta$, so $T$ is unbiased. Hence

$$\begin{aligned} \text{MSE}(T) &= \operatorname{var}(T) \\ &= 4 \operatorname{var}(\overline{X}) \\ &= \frac{4 \operatorname{var}(X_1)}{n}. \end{aligned}$$

We have $E(X_1) = \theta/2$ and

$$E(X_1^2) = \int_0^\theta x^2 \cdot \frac{1}{\theta}\, dx = \frac{\theta^2}{3}$$

so

$$\operatorname{var}(X_1) = \frac{\theta^2}{3} - \left(\frac{\theta}{2}\right)^2 = \frac{\theta^2}{12}$$

hence

$$\text{MSE}(T) = \frac{4 \operatorname{var}(X_1)}{n} = \frac{\theta^2}{3n}.$$

Previously we showed that $\widehat{\theta}$ has p.d.f.

$$f(y) = \frac{ny^{n-1}}{\theta^n}, \quad 0 \leqslant y \leqslant \theta$$

and $E(\widehat{\theta}) = n\theta/(n+1)$. So $b(\widehat{\theta}) = n\theta/(n+1) - \theta = -\theta/(n+1)$. Also,

$$E(\widehat{\theta^2}) = \int_0^\theta y^2 \cdot \frac{ny^{n-1}}{\theta^n}\, dy = \frac{n\theta^2}{n+2}$$

so

$$\text{var}(\widehat{\theta}) = \theta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) = \frac{n\theta^2}{(n+1)^2(n+2)}$$

hence

$$\begin{aligned}
\text{MSE}(\widehat{\theta}) &= \text{var}(\widehat{\theta}) + [b(\widehat{\theta})]^2 \\
&= \frac{2\theta^2}{(n+1)(n+2)} \\
&< \frac{\theta^2}{3n} \qquad \text{for } n \geqslant 3 \\
&= \text{MSE}(T).
\end{aligned}$$

- $\text{MSE}(\widehat{\theta}) \ll \text{MSE}(T)$ for large $n$, so $\widehat{\theta}$ is much better – its MSE decreases like $1/n^2$ rather than $1/n$.
- Note that $(\frac{n+1}{n})\widehat{\theta}$ is unbiased and

$$\begin{aligned}
\text{MSE}\left(\frac{n+1}{n}\widehat{\theta}\right) &= \text{var}\left(\frac{n+1}{n}\widehat{\theta}\right) \\
&= \frac{(n+1)^2}{n^2}\,\text{var}(\widehat{\theta}) \\
&= \frac{\theta^2}{n(n+2)} \\
&< \text{MSE}(\widehat{\theta}) \qquad \text{for } n \geqslant 2.
\end{aligned}$$

  However, among all estimators of the form $\lambda\widehat{\theta}$, the MSE is minimized by $(\frac{n+2}{n+1})\widehat{\theta}$.

  [To show this: note $\text{var}(\lambda\widehat{\theta}) = \lambda^2\,\text{var}(\widehat{\theta})$ and $b(\lambda\widehat{\theta}) = \frac{\lambda n\theta}{n+1} - \theta$. Now plug in formulae and minimise over $\lambda$.]

**Estimation so far**

So far we've considered getting an estimate which is a single number – a *point* estimate – for the parameter of interest: e.g. $\overline{x}$, $\max x_i$, $s^2$, ....

Maximum likelihood is a good way (usually) of producing an estimate (but we did better when the range of the distribution depends on $\theta$ – fairly unusual).

MLEs are usually asymptotically unbiased, and have MSE decreasing like $1/n$ for large $n$.

MLEs can be found in quite general situations.

# 5 Accuracy of estimation: Confidence Intervals

A crucial aspect of statistics is not just to estimate a quantity of interest, but to assess how *accurate* or *precise* that estimate is. One approach is to find an interval, called a *confidence interval* (CI) within which we think the true parameter falls.

**Theorem 5.1.** *Suppose $a_1, \ldots, a_n$ are constants and that $X_1, \ldots, X_n$ are independent with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $Y = \sum_{i=1}^n a_i X_i$. Then*

$$Y \sim N\Big( \sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2 \Big).$$

**Proof.** See Sheet 1.

We know from Prelims Probability how to calculate the mean and variance of $Y$. The additional information here is that $Y$ has a *normal* distribution, i.e. "a linear combination of normals is itself normal." □

**Example.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ where $\mu$ is unknown and $\sigma_0^2$ is known. What can we say about $\mu$?

By Theorem 5.1,

$$\sum_{i=1}^n X_i \sim N(n\mu, n\sigma_0^2)$$

$$\overline{X} \sim N\Big(\mu, \frac{\sigma_0^2}{n}\Big).$$

So, standardising $\overline{X}$,

$$\overline{X} - \mu \sim N\Big(0, \frac{\sigma_0^2}{n}\Big)$$

$$\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0,1).$$

Now if $Z \sim N(0,1)$, then $P(-1.96 < Z < 1.96) = 0.95$.

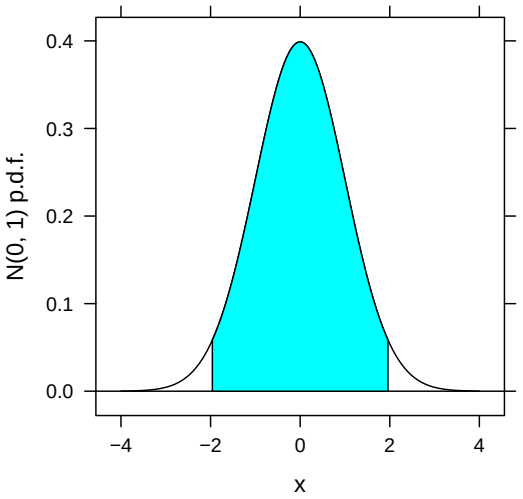


**Figure 5.1.** Standard normal p.d.f.: the shaded area, i.e. the area under the curve from $x = -1.96$ to $x = 1.96$, is 0.95.

So

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} < 1.96\right) = 0.95$$

$$P\left(-1.96\frac{\sigma_0}{\sqrt{n}} < \overline{X} - \mu < 1.96\frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

$$P\left(\overline{X} - 1.96\frac{\sigma_0}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma_0}{\sqrt{n}}\right) = 0.95$$

$$P\left(\text{the interval } \left(\overline{X} \pm 1.96\frac{\sigma_0}{\sqrt{n}}\right) \text{ contains } \mu\right) = 0.95.$$

Note that we write $\left(\overline{X} \pm 1.96\frac{\sigma_0}{\sqrt{n}}\right)$ to mean the interval

$$\left(\overline{X} - 1.96\frac{\sigma_0}{\sqrt{n}},\ \overline{X} + 1.96\frac{\sigma_0}{\sqrt{n}}\right).$$

This interval is a *random* interval since its endpoints involve $\overline{X}$ (and $\overline{X}$ is a random variable). It is an example of a *confidence interval*.

### Confidence Intervals: general definition

**Definition.** If $a(\mathbf{X})$ and $b(\mathbf{X})$ are two statistics, and $0 < \alpha < 1$, the interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is called a *confidence interval* for $\theta$ with confidence level $1 - \alpha$ if, for all $\theta$,

$$P\big(a(\mathbf{X}) < \theta < b(\mathbf{X})\big) = 1 - \alpha.$$

The interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is also called a $100(1-\alpha)\%$ CI, e.g. a "95% confidence interval" if $\alpha = 0.05$.

Usually we are interested in small values of $\alpha$: the most commonly used values are 0.05 and 0.01 (i.e. confidence levels of 95% and 99%) but there is nothing special about any confidence level.

The interval $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ is called an *interval estimate* and the random interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is called an *interval estimator*.

Note: $a(\mathbf{X})$ and $b(\mathbf{X})$ do *not* depend on $\theta$.

We would like to construct $a(\mathbf{X})$ and $b(\mathbf{X})$ so that:

- the width of the interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ is small
- the probability $P\big(a(\mathbf{X}) < \theta < b(\mathbf{X})\big)$ is large.

**Percentage points of normal distribution**

For any $\alpha$ with $0 < \alpha < 1$, let $z_\alpha$ be the constant such that $\Phi(z_\alpha) = 1 - \alpha$, where $\Phi$ is the $N(0,1)$ c.d.f. (i.e. if $Z \sim N(0,1)$ then $P(Z > z_\alpha) = \alpha$).

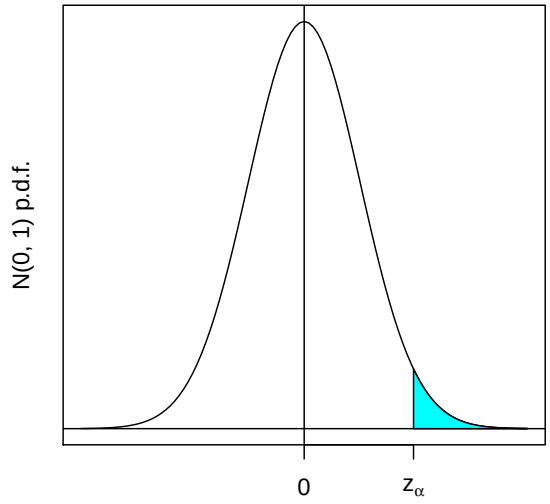


**Figure 5.2.** Standard normal p.d.f.: the shaded area, i.e. the area under the curve to the right of $z_\alpha$, is $\alpha$.

We call $z_\alpha$ the "$1 - \alpha$ quantile of $N(0,1)$."

| $\alpha$ | 0.1 | 0.05 | 0.025 | 0.005 |
|---|---|---|---|---|
| $z_\alpha$ | 1.28 | 1.64 | 1.96 | 2.58 |

By the same argument as before, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ with $\sigma_0^2$ known, then a level $1 - \alpha$ confidence interval for $\mu$ is

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

**Example** (See slides on rainfall example)**.** Measure annual rainfall amounts for 20 years (i.e. $n = 20$). Suppose $\overline{x} = 688.4$ and $\sigma_0 = 130$.

Then the endpoints of a 95% CI for $\mu$ are

$$\overline{x} \pm 1.96\frac{\sigma_0}{\sqrt{n}} = 688.4 \pm 57.0$$

So a 95% CI for $\mu$ is $(631.4, 745.4)$.

The endpoints of a 99% CI are $\overline{x} \pm 2.58\frac{\sigma_0}{\sqrt{n}}$ giving a 99% CI for $\mu$ of $(613.4, 763.4)$.

The symmetric confidence interval for $\mu$ above is called a *central* confidence interval for $\mu$.

Suppose now that $c$ and $d$ are any constants such that

$$P(-c < Z < d) = 1 - \alpha$$

for $Z \sim N(0, 1)$.

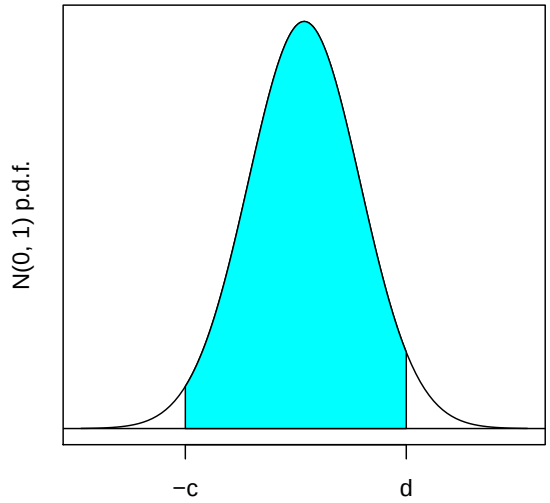


**Figure 5.3.** Standard normal p.d.f.: the shaded area, i.e. the area under the curve from $-c$ to $d$, is $1 - \alpha$.

Then

$$P\left(\overline{X} - \frac{d\sigma_0}{\sqrt{n}} < \mu < \overline{X} + \frac{c\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

The choice $c = d = z_{\alpha/2}$ gives the *shortest* such interval.

**One-sided confidence limits**

Continuing our normal example we have

$$P\left(\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$$

so

$$P\left(\mu < \overline{X} + \frac{z_\alpha \sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$

and so $(-\infty, \overline{X} + \frac{z_\alpha \sigma_0}{\sqrt{n}})$ is a "one-sided" confidence interval. We call $\overline{X} + \frac{z_\alpha \sigma_0}{\sqrt{n}}$ an *upper* $1-\alpha$ *confidence limit* for $\mu$.

Similarly

$$P\left(\mu > \overline{X} - \frac{z_\alpha \sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$

and $\overline{X} - \frac{z_\alpha \sigma_0}{\sqrt{n}}$ is a *lower* $1-\alpha$ *confidence limit* for $\mu$.

**Interpretation of a Confidence Interval**

- The parameter $\theta$ is fixed but unknown.
- If we imagine repeating our experiment then we'd get new data, $\mathbf{x}' = (x_1', \ldots, x_n')$ say, and hence we'd get a new confidence interval $\big(a(\mathbf{x}'), b(\mathbf{x}')\big)$. If we did this repeatedly we would "catch" the true parameter value about 95% of the time, for a 95% confidence interval: i.e. about 95% of our intervals would contain $\theta$.
- The confidence level is a *coverage probability*, the probability that the *random* confidence interval $\big(a(\mathbf{X}), b(\mathbf{X})\big)$ covers the true $\theta$. (It's a random interval because the endpoints $a(\mathbf{X}), b(\mathbf{X})$ are random variables.)

But note that the interval $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ is *not* a random interval, e.g. $\big(a(\mathbf{x}), b(\mathbf{x})\big) = (631.4, 745.4)$ in the rainfall example. So it is wrong to say that $\big(a(\mathbf{x}), b(\mathbf{x})\big)$ contains $\theta$ with probability $1-\alpha$: this interval, e.g. $(631.4, 745.4)$, *either* definitely does *or* definitely does not contain $\theta$, but we can't say which of these two possibilities is true as $\theta$ is unknown.

## 5.1 Confidence Intervals using the CLT

**The Central Limit Theorem (CLT)**

We know that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ then

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

**Theorem 5.2.** *Let $X_1, \ldots, X_n$ be i.i.d. from any distribution with mean $\mu$ and variance $\sigma^2 \in (0, \infty)$. Then, for all $x$,*

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leqslant x\right) \to \Phi(x) \quad \text{as } n \to \infty.$$

Here, as usual, $\Phi$ is the $N(0,1)$ c.d.f. So for large $n$, *whatever the distribution* of the $X_i$,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

where $\approx$ means "has approximately the same distribution as." Usually $n > 30$ is ok for a reasonable approximation.

With $X_1, X_2, \ldots$ i.i.d. from *any* distribution with $E(X_i) = \mu$, $\text{var}(X_i) = \sigma^2$:

- the weak law of large numbers (Prelims Probability) tells us that the distribution of $\overline{X}$ concentrates around $\mu$ as $n$ becomes large,

  i.e. for $\epsilon > 0$, we have $P(|\overline{X} - \mu| > \epsilon) \to 0$ as $n \to \infty$

- the CLT adds to this
  - the fluctuations of $\overline{X}$ around $\mu$ are of order $1/\sqrt{n}$
  - the asymptotic distribution of these fluctuations is normal.

**Example.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential mean $\mu$, e.g. $X_i$ = survival time of patient $i$. So

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu}, \quad x \geqslant 0$$

and $E(X_i) = \mu$, $\text{var}(X_i) = \mu^2$.

For large $n$, by CLT,

$$\frac{\overline{X} - \mu}{\mu/\sqrt{n}} \approx N(0,1).$$

So

$$P\left( -z_{\alpha/2} < \frac{\overline{X} - \mu}{\mu/\sqrt{n}} < z_{\alpha/2} \right) \approx 1 - \alpha$$

$$P\left( \mu\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right) < \overline{X} < \mu\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right) \right) \approx 1 - \alpha$$

$$P\left( \frac{\overline{X}}{1 + \frac{z_{\alpha/2}}{\sqrt{n}}} < \mu < \frac{\overline{X}}{1 - \frac{z_{\alpha/2}}{\sqrt{n}}} \right) \approx 1 - \alpha.$$

Hence an approximate $1 - \alpha$ CI for $\mu$ is

$$\left( \frac{\overline{X}}{1 + \frac{z_{\alpha/2}}{\sqrt{n}}}, \frac{\overline{X}}{1 - \frac{z_{\alpha/2}}{\sqrt{n}}} \right).$$

Numerically, if we have $n = 100$ patients and $\alpha = 0.05$ (so $z_{\alpha/2} = 1.96$), then

$$(0.84\overline{x}, 1.24\overline{x})$$

is an approximate 95% CI for $\mu$.

**Example** (Opinion poll)**.** In a opinion poll, suppose 321 of 1003 voters said they would vote for the Party X. What's the underlying level of support for Party X?

Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli($p$) distribution, i.e.

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p.$$

The MLE of $p$ is $\overline{X}$. Also $E(X_i) = p$ and $\text{var}(X_i) = p(1-p) = \sigma^2(p)$ say.

For large $n$, by CLT,

$$\frac{\overline{X} - p}{\sigma(p)/\sqrt{n}} \approx N(0,1).$$

So

$$\begin{aligned} 1 - \alpha &\approx P\left( -z_{\alpha/2} < \frac{\overline{X} - p}{\sigma(p)/\sqrt{n}} < z_{\alpha/2} \right) \\ &= P\left( \overline{X} - z_{\alpha/2}\frac{\sigma(p)}{\sqrt{n}} < p < \overline{X} + z_{\alpha/2}\frac{\sigma(p)}{\sqrt{n}} \right). \end{aligned}$$

The interval $\left(\overline{X} \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sigma(p)\right)$ has approximate probability $1-\alpha$ of containing the true $p$, but it is *not* a confidence interval since its endpoints depend on $p$ via $\sigma(p)$.

To get an approximate confidence interval:

- either, solve the inequality to get $P\big(a(\mathbf{X}) < p < b(\mathbf{X})\big) \approx 1 - \alpha$ where $a(\mathbf{X})$, $b(\mathbf{X})$ *don't* depend on $p$
- or, estimate $\sigma(p)$ by $\sigma(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})}$ where $\widehat{p} = \overline{x}$ the MLE. This gives endpoints

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

  For $n = 1003$ and $\widehat{p} = \overline{x} = 321/1003$, this gives a 95% CI for $p$ of $(0.29, 0.35)$ using the two approximations: (i) CLT and (ii) $\sigma(p)$ approximated by $\sigma(\widehat{p})$.

**Example** (Opinion polls continued)**.** Opinion polls often mention "±3% error."

Note that for any $p$,

$$\sigma^2(p) = p(1-p) \leqslant \frac{1}{4}$$

since $p(1-p)$ is maximised at $p = \frac{1}{2}$. Then we have

$$\begin{aligned} P(\widehat{p} - 0.03 < p < \widehat{p} + 0.03) &= P\left( \frac{-0.03}{\sigma(p)/\sqrt{n}} < \frac{\widehat{p} - p}{\sigma(p)/\sqrt{n}} < \frac{0.03}{\sigma(p)/\sqrt{n}} \right) \\ &\approx \Phi\left( \frac{0.03}{\sigma(p)/\sqrt{n}} \right) - \Phi\left( \frac{-0.03}{\sigma(p)/\sqrt{n}} \right) \\ &\geqslant \Phi(0.03\sqrt{4n}) - \Phi(-0.03\sqrt{4n}). \end{aligned}$$

For this probability to be at least 0.95 we need $0.03\sqrt{4n} \geqslant 1.96$, or $n \geqslant 1068$. Opinion polls typically use $n \approx 1100$.

### Standard errors

**Definition.** Let $\widehat{\theta}$ be an estimator of $\theta$ based on $\mathbf{X}$. The *standard error* $\mathrm{SE}(\widehat{\theta})$ of $\widehat{\theta}$ is defined by

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\mathrm{var}(\widehat{\theta})}.$$

**Example.**

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Then $\widehat{\mu} = \overline{X}$ and $\text{var}(\widehat{\mu}) = \sigma^2/n$. So $\text{SE}(\widehat{\mu}) = \sigma/\sqrt{n}$.
- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. Then $\widehat{p} = \overline{X}$ and $\text{var}(\widehat{p}) = p(1-p)/n$. So $\text{SE}(\widehat{p}) = \sqrt{p(1-p)/n}$.

Sometimes $\text{SE}(\widehat{\theta})$ depends on $\theta$ itself, meaning that $\text{SE}(\widehat{\theta})$ is unknown. In such cases we have to plug-in parameter estimates to get the *estimated* standard error. E.g. plug-in to get estimated standard errors $\text{SE}(\overline{X}) = \widehat{\sigma}/\sqrt{n}$ and $\text{SE}(\widehat{p}) = \sqrt{\widehat{p}(1-\widehat{p})/n}$.

The values plugged-in ($\widehat{\sigma}$ and $\widehat{p}$ above) could be maximum likelihood, or other, estimates.

(We could write $\widehat{\text{SE}}(\widehat{p})$, i.e. with a hat on the SE, to denote that $\widehat{p}$ has been plugged-in, but this is ugly so we won't, we'll just write $\text{SE}(\widehat{p})$.)

If $\widehat{\theta}$ is unbiased, then $\text{MSE}(\widehat{\theta}) = \text{var}(\widehat{\theta}) = [\text{SE}(\widehat{\theta})]^2$. So the standard error (or estimated standard error) gives some quantification of the accuracy of estimation.

If in addition $\widehat{\theta}$ is approximately $N(\theta, \text{SE}(\widehat{\theta})^2)$ then, by the arguments used above, an approximate $1-\alpha$ CI for $\theta$ is given by $\left(\widehat{\theta} \pm z_{\alpha/2}\,\text{SE}(\widehat{\theta})\right)$ where again we might need to plug-in to obtain the estimated standard error. Since, roughly, $z_{0.025} = 2$ and $z_{0.001} = 3$,

| | |
|---|---|
| (estimate ± 2 estimated std errors) | is an approximate 95% CI |
| (estimate ± 3 estimated std errors) | is an approximate 99.8% CI. |

The CLT justifies the normal approximation for $\widehat{\theta} = \overline{X}$, but $\widehat{\theta} \approx N(\theta, \text{SE}(\widehat{\theta})^2)$ is also appropriate for more general MLEs by other theory (see Part A).

**Example.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with $\mu$ and $\sigma^2$ unknown.

The MLEs are $\widehat{\mu} = \overline{X}$, $\widehat{\sigma}^2 = \frac{1}{n}\sum(X_i - \overline{X})^2$, and $\text{SE}(\widehat{\mu}) = \sigma/\sqrt{n}$ is unknown because $\sigma$ is unknown. So to use $\widehat{\mu} \pm z_{\alpha/2}\,\text{SE}(\widehat{\mu})$ as the basis for a confidence interval we need to estimate $\sigma$. One possibility is to use $\widehat{\sigma}$ and so get the interval

$$\left(\widehat{\mu} \pm z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right).$$

However we can improve on this:

(i) use $s$ (the sample standard deviation) instead of $\widehat{\sigma}$ (better as $E(S^2) = \sigma^2$ whereas $E(\widehat{\sigma}^2) < \sigma^2$)

(ii) use a critical value from a $t$-distribution instead of $z_{\alpha/2}$ – see Part A Statistics (better as $(\overline{X} - \mu)/(S/\sqrt{n})$ has a $t$-distribution, exactly, whereas using the CLT is an approximation).

# 6 Linear Regression

Suppose we measure two variables in the same population:

$x$ the *explanatory* variable

$y$ the *response* variable.

Other possible names for $x$ are the *predictor* or *feature* or *input variable* or *independent variable*. Other possible names for $y$ are the *output variable* or *dependent variable*.

**Example.** Let $x$ measure the GDP per head, and $y$ the $CO_2$ emissions per head, for various countries.

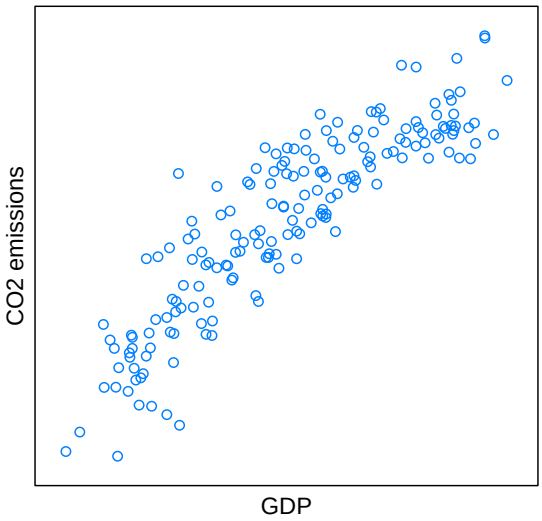


**Figure 6.1.** Scatterplot of measures of $CO_2$ emissions per head against GDP per head, for 178 countries.

Questions of interest:

For fixed $x$, what is the average value of $y$?

How does that average value change with $x$?

A simple model for the dependence of $y$ on $x$ is

$$y = \alpha + \beta x + \text{"error"}.$$

Note: a linear relationship like this does not necessarily imply that $x$ *causes* $y$.

### More precise model

We regard the values of $x$ as being fixed and known, and we regard the values of $y$ as being the observed values of random variables.

We suppose that

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n \tag{6.1}$$

where

$$\begin{aligned} x_1, \ldots, x_n &\quad \text{are known constants} \\ \epsilon_1, \ldots, \epsilon_n &\quad \text{are i.i.d. } N(0, \sigma^2) \text{ "random errors"} \\ \alpha, \beta &\quad \text{are unknown parameters.} \end{aligned}$$

The "random errors" $\epsilon_1, \ldots, \epsilon_n$ represent random scatter of the points $(x_i, y_i)$ about the line $y = \alpha + \beta x$, we do not expect these points to lie on a perfect straight line.

Sometimes we will refer to the above model as being $Y = \alpha + \beta x + \epsilon$.

**Note.**

1. The values $y_1, \ldots, y_n$ (e.g. the $CO_2$ emissions per head in the various countries) are the observed values of random variables $Y_1, \ldots, Y_n$.
2. The values $x_1, \ldots, x_n$ (e.g. the GDP per head in the various countries) do *not* correspond to random variables. They are *fixed* and *known* constants.

Questions:

- How do we estimate $\alpha$ and $\beta$?
- Does the mean of $Y$ actually depend on the value of $x$? i.e. is $\beta \neq 0$?

We now find the MLEs of $\alpha$ and $\beta$, and we regard $\sigma^2$ as being known. The MLEs of $\alpha$ and $\beta$ are the same if $\sigma^2$ is unknown. If $\sigma^2$ is unknown, then we simply maximise over $\sigma^2$ as well to obtain its MLE – this is no harder than what we do here (try it!). However, working out all of the properties of this MLE *is* harder and beyond what we can do in this course.

From (6.1) we have $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$. So $Y_i$ has p.d.f.

$$f_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2 \right), \quad -\infty < y_i < \infty.$$

So the likelihood $L(\alpha, \beta)$ is

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2 \right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right) \end{aligned}$$

with log-likelihood

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

So *maximising* $\ell(\alpha, \beta)$ over $\alpha$ and $\beta$ is equivalent to *minimising* the sum of squares

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

For this reason the MLEs of $\alpha$ and $\beta$ are also called *least squares estimators*.

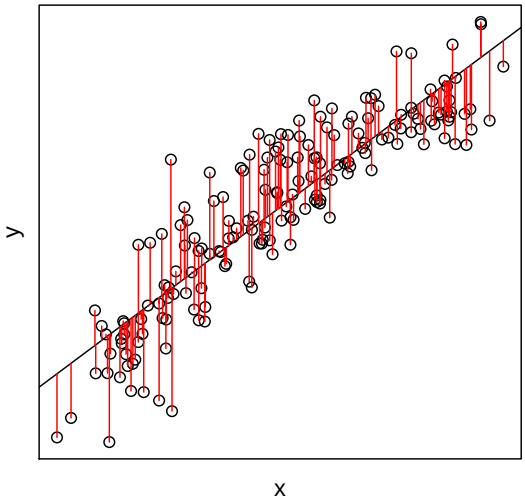


**Figure 6.2.** Vertical distances from the points $(x_i, y_i)$ to the line $y = \alpha + \beta x$. $S(\alpha, \beta)$ is the sum of these squared distances. The MLEs/least squares estimates of $\alpha$ and $\beta$ minimise this sum.

**Theorem 6.1.** *The MLEs (or, equivalently, the least squares estimates) of $\alpha$ and $\beta$ are given by*

$$\widehat{\alpha} = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\widehat{\beta} = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}.$$

The sums in Theorem 6.1, and similar sums below, are from $i = 1$ to $n$.

**Proof.** To find $\widehat{\alpha}$ and $\widehat{\beta}$ we calculate

$$\frac{\partial S}{\partial \alpha} = -2 \sum (y_i - \alpha - \beta x_i)$$

$$\frac{\partial S}{\partial \beta} = -2 \sum x_i (y_i - \alpha - \beta x_i).$$

Putting these partial derivatives equal to zero, the minimisers $\widehat{\alpha}$ and $\widehat{\beta}$ satisfy

$$n\widehat{\alpha} + \widehat{\beta} \sum x_i = \sum y_i$$

$$\widehat{\alpha} \sum x_i + \widehat{\beta} \sum x_i^2 = \sum x_i y_i.$$

Solving this pair of simultaneous equations for $\widehat{\alpha}$ and $\widehat{\beta}$ gives the required MLEs. □

**Note.** Sometimes we consider the model

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

and find the MLEs of $a$ and $b$ by minimising $\sum(y_i - a - b(x_i - \overline{x}))^2$. This model is just an alternative parametrisation of our original model: the first model is

$$Y = \alpha + \beta x + \epsilon$$

and the second is

$$\begin{aligned} Y &= a + b(x - \overline{x}) + \epsilon \\ &= (a - b\overline{x}) + bx + \epsilon. \end{aligned}$$

Here $Y, x$ denote general values of $Y, x$ (and $\overline{x} = \frac{1}{n}\sum x_i$ is the mean of the $n$ data values of $x$). Comparing the two model equations, $b = \beta$ and $a - b\overline{x} = \alpha$.

The interpretation of the parameters is that $\beta = b$ is the increase in $E(Y)$ when $x$ increases by 1. The parameter $\alpha$ is the value of $E(Y)$ when $x$ is 0; whereas $a$ is the value of $E(Y)$ when $x$ is $\overline{x}$.

- Alternative expressions for $\widehat{\alpha}$ and $\widehat{\beta}$ are

$$\widehat{\beta} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \tag{6.2}$$

$$= \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2} \tag{6.3}$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}.$$

  The above alternative for $\widehat{\alpha}$ follows directly from $\frac{\partial S}{\partial \alpha} = 0$.

  To obtain the alternatives for $\widehat{\beta}$: Theorem 6.1 gives

$$\begin{aligned} \widehat{\beta} &= \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \\ &= \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2}. \end{aligned} \tag{6.4}$$

  Now check that the numerators and denominators in (6.2) and (6.4) are the same. Then observe that the numerators of (6.2) and (6.3) differ by $\sum(x_i - \overline{x})\overline{y}$, which is 0.

- The *fitted* regression line is the line $y = \widehat{\alpha} + \widehat{\beta}x$. The point $(\overline{x}, \overline{y})$ always lies on this line.

## 6.1 Properties of regression parameter estimators

**Bias**

Let $w_i = x_i - \overline{x}$ and note $\sum w_i = 0$.

From (6.3) the maximum likelihood *estimator* of $\beta$ is

$$\widehat{\beta} = \frac{1}{\sum w_i^2}\Big(\sum w_i Y_i\Big)$$

so

$$E(\widehat{\beta}) = \frac{1}{\sum w_i^2} E\Big(\sum w_i Y_i\Big)$$
$$= \frac{1}{\sum w_i^2} \sum w_i E(Y_i).$$

Note $E(Y_i) = \alpha + \beta x_i = \alpha + \beta\overline{x} + \beta w_i$ (using $x_i = w_i + \overline{x}$), and so

$$E(\widehat{\beta}) = \frac{1}{\sum w_i^2} \sum w_i(\alpha + \beta\overline{x} + \beta w_i)$$
$$= \frac{1}{\sum w_i^2} \Big[(\alpha + \beta\overline{x}) \sum w_i + \beta \sum w_i^2\Big]$$
$$= \beta \qquad \text{since } \sum w_i = 0.$$

Also $\widehat{\alpha} = \overline{Y} - \widehat{\beta}\,\overline{x}$ and so

$$E(\widehat{\alpha}) = E(\overline{Y}) - \overline{x}E(\widehat{\beta})$$
$$= \frac{1}{n} \sum E(Y_i) - \beta\overline{x} \qquad \text{since } E(\widehat{\beta}) = \beta$$
$$= \frac{1}{n} \sum (\alpha + \beta x_i) - \beta\overline{x}$$
$$= \frac{1}{n} \cdot n(\alpha + \beta\overline{x}) - \beta\overline{x}$$
$$= \alpha.$$

So $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased.

Note the unbiasedness of $\widehat{\alpha}$, $\widehat{\beta}$ does not depend on the assumptions that the $\epsilon_i$ are independent, normal and have the same variance, only on the assumptions that the errors are additive and $E(\epsilon_i) = 0$.

**Variance**

We are usually only interested in the variance of $\widehat{\beta}$:

$$\text{var}(\widehat{\beta}) = \text{var}\left(\frac{\sum w_i Y_i}{\sum w_i^2}\right)$$
$$= \frac{1}{(\sum w_i^2)^2} \text{var}\Big(\sum w_i Y_i\Big)$$
$$= \frac{1}{(\sum w_i^2)^2} \sum w_i^2 \,\text{var}(Y_i)$$
$$= \frac{1}{(\sum w_i^2)^2} \sum w_i^2 \sigma^2$$
$$= \frac{\sigma^2}{\sum w_i^2}.$$

Since $\widehat{\beta}$ is a linear combination of the independent normal random variables $Y_i$, the estimator $\widehat{\beta}$ is itself normal: $\widehat{\beta} \sim N(\beta, \sigma^2_{\beta})$ where $\sigma^2_{\beta} = \sigma^2 / \sum w_i^2$.

So the standard error of $\widehat{\beta}$ is $\sigma_{\beta}$ and if $\sigma^2$ is known, then a 95% CI for $\beta$ is

$$(\widehat{\beta} \pm 1.96\sigma_{\beta}).$$

However, this is only a valid CI when $\sigma^2$ is known and, in practice, $\sigma^2$ is rarely known.

For $\sigma^2$ unknown we need to plug-in an estimate of $\sigma^2$, i.e. use $\widehat{\sigma}^2_{\beta} = \widehat{\sigma}^2 / \sum w_i^2$ where $\widehat{\sigma}^2$ is some estimate of $\sigma^2$. For example we could use the MLE which is $\widehat{\sigma}^2 = \frac{1}{n} \sum (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$. Using the $\widehat{\theta} \pm 2\,\mathrm{SE}(\widehat{\theta})$ approximation for a 95% confidence interval, we have that $(\widehat{\beta} \pm 2\widehat{\sigma}_{\beta})$ is an approximate 95% confidence interval for $\beta$.

[A better approach here, but beyond the scope of this course, is to estimate $\sigma^2$ using

$$s^2 = \frac{1}{n-2} \sum (y_i - \widehat{\alpha} - \widehat{\beta} x_i)^2$$

and to base the confidence interval on a $t$-distribution rather than a normal distribution. This estimator $S^2$ is unbiased for $\sigma^2$ (see Sheet 5), but details about its distribution and the $t$-distribution are beyond this course – see Parts A/B.]

# 7 Multiple Linear Regression

For the material in the last two sections of these notes (this section and the next), Chapter 3 of the book *An Introduction to Statistical Learning* by James, Witten, Hastie and Tibshirani (2013) is useful – we cover some, but not all, of the material in that chapter. The whole book is freely available online at:

`http://www-bcf.usc.edu/~gareth/ISL/`

We'll refer to this book as JWHT (2013).

**Example** (Hill races)**.** Let $Y$ = time, $x_1$ = distance, $x_2$ = climb.

We can consider a model in which $Y$ depends on both $x_1$ and $x_2$, of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

This model has one response variable $Y$ (as usual), but now we have *two* explanatory variables $x_1$ and $x_2$, and *three* regression parameters $\beta_0, \beta_1, \beta_2$.

Let the $i$th race have time $y_i$, distance $x_{i1}$ and climb $x_{i2}$. Then in more detail our model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n$$

where

$$\begin{aligned} &x_{i1}, x_{i2}, \text{ for } i = 1, \ldots, n, \text{ are known constants} \\ &\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2) \\ &\beta_0, \beta_1, \beta_2 \text{ are unknown parameters} \end{aligned}$$

and (as usual) $y_i$ denotes the observed value of the random variable $Y_i$.

As for simple linear regression we obtain the MLEs/least squares estimates of $\beta_0, \beta_1, \beta_2$ by minimising

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

with respect to $\beta_0, \beta_1, \beta_2$, i.e. by solving $\frac{\partial S}{\partial \beta_k} = 0$ for $k = 0, 1, 2$.

As before, the only property of the $\epsilon_i$ needed to define the least squares estimates is $E(\epsilon_i) = 0$.

A general multiple regression model has $p$ explanatory variables $(x_1, \ldots, x_p)$,

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

and the MLEs/least squares estimates are obtained by minimising

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

with respect to $\beta_0, \beta_1, \ldots, \beta_p$. In this course we will focus on $p = 1$ or 2.

**Example** (Quadratic regression)**.** The relationship between $Y$ and $x$ may be approximately quadratic in which case we can consider the model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. This is the case $p = 2$ with $x_1 = x$ and $x_2 = x^2$.

**Example.** For convenience write the two explanatory variables as $x$ and $z$. So suppose

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i, \quad i = 1, \ldots, n$$

where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ and assume $\sum x_i = \sum z_i = 0$.

Then minimising $S(\beta)$ gives

$$\widehat{\beta}_0 = \frac{\sum y_i}{n}$$

$$\widehat{\beta}_1 = \frac{1}{\Delta}\left(\sum z_i^2 \sum x_i y_i - \sum x_i z_i \sum z_i y_i\right)$$

$$\widehat{\beta}_2 = \frac{1}{\Delta}\left(\sum x_i^2 \sum z_i y_i - \sum x_i z_i \sum x_i y_i\right)$$

where

$$\Delta = \sum x_i^2 \sum z_i^2 - \left(\sum x_i z_i\right)^2.$$

The method (solving $\frac{\partial S}{\partial \beta_k} = 0$ for $k = 0, 1, 2$, i.e. 3 equations in 3 unknowns) is straightforward, the algebra less so – the are more elegant ways to do some of this (using matrices, in 3rd year).

**Interpretation of regression coefficients**

Consider the model with $p = 2$, so two regressors $x_1$ and $x_2$:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

We interpret $\beta_1$ as the average effect on $Y$ of a one unit increase in $x_1$, *holding the other regressor $x_2$ fixed*.

Similarly we interpret $\beta_2$ as the average effect on $Y$ of a one unit increase in $x_2$, *holding the other regressor $x_1$ fixed*.

In these interpretations "average" means we are talking about the change in $E(Y)$ when changing $x_1$ (or $x_2$).

One important thing to note here is that $x_1$ and $x_2$ often change together. E.g. in the hill races example, a race whose distance is one mile greater will usually to have an increased value of climb as well. This makes interpretation more difficult.

# 8 Assessing the fit of a model

Having fitted a model, we should consider how well it fits the data. A model is normally an approximation to reality: is the approximation sufficiently good that the model is useful? This question applies to mathematical models in general. In this course we will approach the question by considering the fit of a simple linear regression (generalisations are possible).

For the model $Y = \alpha + \beta x + \epsilon$ let $\widehat{\alpha}, \widehat{\beta}$ be the usual estimates of $\alpha, \beta$ based on the observation pairs $(x_1, y_1), \ldots, (x_n, y_n)$.

From now on we consider this model, with the usual assumptions about $\epsilon$, unless otherwise stated.

**Definition.** The $i$th *fitted value* $\widehat{y}_i$ of $Y$ is defined by $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$, for $i = 1, \ldots, n$.

The $i$th *residual* $e_i$ is defined by $e_i = y_i - \widehat{y}_i$, for $i = 1, \ldots, n$.

The *residual sum of squares* RSS is defined by $\text{RSS} = \sum e_i^2$.

The *residual standard error* RSE is defined by $\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}}$.

The RSE is an estimate of the standard deviation $\sigma$. If the fitted values are close to the observed values, i.e. $\widehat{y}_i \approx y_i$ for all $i$ (so that the $e_i$ are small), then the RSE will be small. Alternatively if one or more of the $e_i$ is large then the RSE will be higher.

We have $E(e_i) = 0$. In taking this expectation, we treat $y_i$ as the random variable $Y_i$, and we treat $\widehat{y}_i$ as the random variable $\widehat{\alpha} + \widehat{\beta} x_i$ (in particular, $\widehat{\alpha}$ and $\widehat{\beta}$ are estimators, not estimates). Hence

$$\begin{aligned} E(e_i) &= E(Y_i - \widehat{\alpha} - \widehat{\beta} x_i) \\ &= E(Y_i) - E(\widehat{\alpha}) - E(\widehat{\beta}) x_i \\ &= E(\alpha + \beta x_i + \epsilon_i) - \alpha - \beta x_i \quad \text{since } \widehat{\alpha}, \widehat{\beta} \text{ are unbiased} \\ &= \alpha + \beta x_i + E(\epsilon_i) - \alpha - \beta x_i \\ &= 0 \quad \text{since } E(\epsilon_i) = 0. \end{aligned}$$

## 8.1 Potential problem: non-linearity

The model $Y = \alpha + \beta x + \epsilon$ assumes a straight-line relationship between $Y$ (the response) and $x$ (the predictor). If the true relationship is far from linear then any conclusions (e.g. predictions) we draw from the fit will be suspect.

A *residual plot* is a useful graphical tool for identifying non-linearity: for simple linear regression we can plot the residuals $e_i$ against the fitted values $\widehat{y}_i$. Ideally the plot will show no pattern. The existence of a pattern may indicate a problem with some aspect of the linear model.

Note that for simple linear regression (i.e. the case $p = 1$) plotting $e_i$ against $x_i$ gives an equivalent plot, just with a different horizontal scale, since there is an exact linear relation between the $x_i$ and $\widehat{y}_i$ (i.e. $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$).

[Plotting $e_i$ against $\widehat{y}_i$ generalises better to multiple regression.]

## 8.2 Potential problem: non-constant variance of errors

We have assumed that the errors have a constant variance, i.e. $\text{var}(Y_i) = \text{var}(\epsilon_i) = \sigma^2$. That is, the *same* variance *for all* $i$. Unfortunately, this is often not true. E.g. the variance of the error may increase as $Y$ increases.

Non-constant variance is also called *heteroscedasticity*. We can identify this from the presence of a funnel-type shape in the residual plot.

How might we deal with non-constant variance of the errors?

(i) One possibility is to transform the response $Y$ using a transformation such as $\log Y$ or $\sqrt{Y}$ (which shrinks larger responses more), leading to a reduction in heteroscedasticity.

(ii) Sometimes we might have a good idea about of the variance of $Y_i$: we might think $\text{var}(Y_i) = \text{var}(\epsilon_i) = \frac{\sigma^2}{w_i}$ where $\sigma^2$ is unknown but where the $w_i$ are known. E.g. if $Y_i$ is actually the mean of $n_i$ observations, where each of these $n_i$ observations are made at $x = x_i$, then $\text{var}(Y_i) = \frac{\sigma^2}{n_i}$. So $w_i = n_i$ in this case.

It is straightforward to show (exercise) that the MLEs of $\alpha, \beta$ are obtained by minimising

$$\sum_{i=1}^{n} w_i(y_i - \alpha - \beta x_i)^2. \tag{8.1}$$

The form of (8.1) is intuitively correct: if $w_i$ is small then $\text{var}(Y_i)$ is large, so there is a lot of uncertainty about observation $i$, so this observation shouldn't affect the fit too much – this is achieved in (8.1) by observation $i$ being weighted by the small value of $w_i$. Hence this approach is called *weighted least squares*.

## 8.3 Potential problem: outliers

An *outlier* is a point for which $y_i$ is far from the value $\widehat{y}_i$ predicted by the model. Outliers can arise for a variety of reasons, e.g. incorrect recording of an observation during data collection.

Residual plots can be used to identify outliers. Recall that $E(e_i) = 0$. But in practice it can be difficult to decide how large (i.e. how far from the expected value of zero) a residual needs to be before we consider it a possible outlier. To address this we can plot *studentized residuals* instead of residuals, where

$$i\text{th studentized residual} = \frac{e_i}{\text{SE}(e_i)}.$$

**Theorem 8.1.** $\text{var}(e_i) = \sigma^2(1 - h_i)$ *where*

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

**Definition.** The $i$th *studentized residual* $r_i$ is defined by

$$r_i = \frac{e_i}{s\sqrt{1 - h_i}}$$

where $s = \text{RSE}$ is the residual standard error.

[Here we follow the terminology in the synopses and JWHT (2013) and call the $r_i$ "studentized" residuals. Some authors call the $r_i$ "standardized" residuals and save the word "studentized" to mean something similar but different.]

So the $r_i$ are all on a comparable scale, each having a standard deviation of about 1. Following JWHT (2013) we will say that observations with $|r_i| > 3$ are possible outliers.

If we believe an outlier is due to an error in data collection, then one solution is to simply remove the observation from the data and re-fit the model. However, an outlier may instead indicate a problem with the model, e.g. a nonlinear relationship between $Y$ and $x$, so care must be taken.

Similarly, this kind of problem could arise if we have a missing regressor, i.e. we could be using $Y = \beta_0 + \beta_1 x_1 + \epsilon$ when we should really be using $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

**Proof of Theorem 8.1**

Idea of the proof: write $e_j = \sum_j a_j Y_j$ and use

$$\text{var}\left(\sum_j a_j Y_j\right) = \sum_j a_j^2 \,\text{var}(Y_j) = \sigma^2 \sum_j a_j^2 \tag{8.2}$$

since the $Y_j$ are independent with $\text{var}(Y_j) = \sigma^2$ for all $j$. Here and below, all sums are from 1 to n.

First recall

$$\widehat{\beta} = \frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}} \quad \text{where } S_{xx} = \sum_k (x_k - \overline{x})^2$$

and

$$\begin{aligned}
\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\,\overline{x} \\
&= \frac{1}{n}\left(\sum_j Y_j\right) - \overline{x}\left(\frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}}\right) \\
&= \sum_j \left(\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{S_{xx}}\right) Y_j.
\end{aligned}$$

So

$$\begin{aligned}
\widehat{y_i} &= \widehat{\alpha} + \widehat{\beta} x_i \\
&= \sum_j \left(\frac{1}{n} - \frac{\overline{x}(x_j - \overline{x})}{S_{xx}}\right) Y_j + x_i \left(\frac{\sum_j (x_j - \overline{x}) Y_j}{S_{xx}}\right) \\
&= \sum_j \left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}}\right) Y_j.
\end{aligned} \tag{8.3}$$

We can write

$$Y_i = \sum_j \delta_{ij} Y_j \tag{8.4}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Note $\delta_{ij}^2 = \delta_{ij}$ and $\sum_j \delta_{ij}^2 = \sum_j \delta_{ij} = 1$. So

$$e_i = Y_i - \widehat{\alpha} - \widehat{\beta} x_i$$

and using (8.3) and (8.4)

$$= \sum_j \left( \delta_{ij} - \frac{1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right) Y_j.$$

As the $Y_j$ are independent, as at (8.2),

$$\begin{aligned}
\text{var}(e_i) &= \sum_j \left( \delta_{ij} - \frac{1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right)^2 \text{var}(Y_j) \\
&= \sigma^2 \sum_j \left( \delta_{ij}^2 + \frac{1}{n^2} + \frac{(x_i - \overline{x})^2 (x_j - \overline{x})^2}{S_{xx}^2} - 2\frac{1}{n}\delta_{ij} \right. \\
&\qquad \left. - 2\frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}}\delta_{ij} + 2\frac{1}{n}\frac{(x_i - \overline{x})(x_j - \overline{x})}{S_{xx}} \right) \\
&= \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{S_{xx}} - \frac{2}{n} - 2\frac{(x_i - \overline{x})^2}{S_{xx}} + \frac{2}{n}\frac{(x_i - \overline{x})}{S_{xx}} \sum_j (x_j - \overline{x}) \right) \\
&= \sigma^2 \left( 1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}} \right) \\
&= \sigma^2 (1 - h_i). \quad \square
\end{aligned}$$

## 8.4 Potential problem: high leverage points

Outliers are observations for which the response $y_i$ is unusual given the value of $x_i$. On the other hand, observations with *high leverage* have an unusual value of $x_i$.

**Definition.** The *leverage* of the $i$th observation is $h_i$, where

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^n (x_j - \overline{x})^2}.$$

Clearly $h_i$ depends only on the values of the $x_i$ (it doesn't depend on the $y_i$ values at all). We see that $h_i$ increases with the distance of $x_i$ from $\overline{x}$.

High leverage points tend to have a sizeable impact on the regression line. Since $\text{var}(e_i) = \sigma^2(1 - h_i)$, a large leverage $h_i$ will make $\text{var}(e_i)$ small. And then since $E(e_i) = 0$, this means that the regression line will be "pulled" close to the point $(x_i, y_i)$.

Note that $\sum h_i = 2$, so the average leverage is $\overline{h} = \frac{2}{n}$. One rule-of-thumb is to regard points with a leverage more than double this to be high leverage points, i.e. points with $h_i > 4/n$.

Why does this matter? We should be concerned if the regression line is heavily affected by just a couple of points, because any problems with these points might invalidate the entire fit. Hence it is important to identify high leverage observations.

**SLIDES.** Slide with example involving points $A$ and $B$ goes here.

Removing point $B$ – which has high leverage and a high residual – has a much more substantial impact on the regression line than removing the outlier had at the end of Section 8.3.

In the plot of studentized residuals against leverage:

- point B has high leverage and a high residual, it is having a substantial effect on the regression line yet it still has a high residual
- in contrast point A has a high residual, the model isn't fitting well at A, but we saw in Section 8.3 that A isn't affecting the fit much at all – the reason that A has little effect on the fit is that it has low leverage.