MODS STATISTICS

HT 2012

Simon Myers,

Department of Statistics (and The Wellcome Trust Centre for Human Genetics)

myers@stats.ox.ac.uk

Lecture notes and problem sheets will be available from the Mathematical Institute's website, but more directly:

www.stats.ox.ac.uk/~myers/modsstats.html

Introduction.

We will be concerned with the mathematical framework for making inferences from data. The tools of probability provide the backdrop, allowing us to quantify the uncertainties involved.

Examples

1. Question: How tall is the average five year old girl?

Data: $x_1, x_2, \ldots, x_n$, the heights of n randomly chosen girls.

An obvious estimate is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

How precise is our estimate?

2. Measure two variables, for example:

$$\begin{aligned} x_i &= \text{Height of father} \\ y_i &= \text{Height of son}\,, \end{aligned}$$

where $i = 1, \ldots, n$.

Is it reasonable that

$$y_i = \alpha + \beta x_i + \text{“random error”}\,?$$

Is $\beta > 0$?

Which α and β?

Notation

We usually denote observations by lower case letters: $x_1, x_2, \ldots, x_n$.

Regard these as observed values of random variables (rv's) (for which we usually use upper case) $X_1, X_2, \ldots, X_n$.

We often write x (respectively X) for the collection $x_1, x_2, \ldots, x_n$ (respectively $X_1, X_2, \ldots, X_n$).

In different settings, it is convenient to think of x_i as the observed value of X_i, or as a possible value that X_i can take.

For example, if X_i is a Poisson random variable with mean λ,

$$\mathbf{P}(X_i = x_i) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!},$$

for $x_i = 0, 1, 2, \ldots$.

1. Random Samples.

Definition 1 A *random sample* of size n is a set of random variables $X_1, X_2, \ldots, X_n$ which are independent and identically distributed (i.i.d.).

Examples

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean λ. (e.g. $X_i = \#$ of accidents on Parks Road in year i.) Then,

$$\begin{aligned} f(x) &= \mathrm{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= \mathrm{P}(X_1 = x_1)\mathrm{P}(X_2 = x_2) \cdots \mathrm{P}(X_n = x_n) \\ &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdot \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\ \lambda^{(\sum_{i=1}^n x_i)}}{\prod_{i=1}^n x_i!}, \end{aligned}$$

where the second equality follows from the independence of the X_i.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from an exponential distribution with probability density function (p.d.f.)

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

(e.g. X_i might be the time until the ith of a collection of pedestrians is able to cross Parks Road on the way to a lecture.)

Again, since the X_i are independent, their joint distribution is

$$\begin{aligned} f(x) &= f(x_1) \cdot f(x_2) \cdots f(x_n) \\ &= \prod_{i=1}^{n} \frac{1}{\mu} e^{-\frac{x_i}{\mu}} \\ &= \frac{1}{\mu^n} e^{\left(-\frac{1}{\mu} \sum_{i=1}^{n} x_i\right)} . \end{aligned}$$

In probability questions we would usually assume that the parameters λ and μ from our previous examples are known.

In many settings they will not be known, and we wish to estimate them from data. Two key questions of interest are:

1. What is the best way to estimate them? (And what does "best" mean here?)

2. For a given method of estimation, how precise is a particular estimator?

2. Summary Statistics.

Definition 2 Let $X_1, X_2, \ldots, X_n$ be a random sample. The *sample mean* is defined as

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

The *sample variance* is defined as

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

The *sample standard deviation* is S ($= \sqrt{S^2}$).

Notes

1. The denominator in the definition of S^2 is $n-1$, not n.

2. $\bar{X}$ and S^2 are random variables, so they have distributions (called the *sampling distributions* of $\bar{X}$ and S^2.)

3. Given observations $x_1, x_2, \ldots, x_n$, we can compute the observed values of $\bar{x}$ and s^2.

The sample mean $\bar{x}$ is a summary of the location of the sample.

The sample standard deviation S (or the sample variance S^2) is a summary of the spread of the sample about $\bar{x}$.

The Normal Distribution.

Definition 3 Recall that X has a normal distribution with mean μ and variance σ^2, written $X \sim \mathrm{N}(\mu, \sigma^2)$, if the p.d.f. of X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\; e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

for $-\infty < x < \infty$.

Recall also that $\mathbf{E}(X) = \mu$ and $\mathrm{var}(X) = \sigma^2$.

If $\mu = 0$ and $\sigma = 1$, then X is said to have a *standard normal distribution*, and we write $X \sim \mathsf{N}(0, 1)$.

Important Result

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

The cumulative distribution function (c.d.f.) of a standard normal random variable is:

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

3. Maximum Likelihood Estimation.

We now describe one method for estimating unknown parameters from data, called the method of maximum likelihood. Although this shouldn't be obvious at this stage, it turns out to be the method of choice in many contexts.

Example 1. Suppose X has an exponential distribution with mean μ. We indicate the dependence on μ by writing the p.d.f. as

$$f(x;\mu) = \begin{cases} \frac{1}{\mu}e^{-\frac{x}{\mu}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

In general we write $f(x;\theta)$ to indicate that the p.d.f. (or p.m.f.) f, which is a function of x, depends on the parameter θ (sometimes this is written $f(x|\theta)$).

Example 1. continued

Suppose $n = 62$ and $x_1, x_2, \ldots, x_n$ are 62 time intervals between major earthquakes. Assume $X_1, X_2, \ldots, X_n$ are exponential random variables with mean μ.

How does one estimate the unknown μ? Intuition suggests using $\mu = \bar{x}$. But is this a good idea? Are there general principles we can use to choose estimators?

In general, suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with p.d.f. (or p.m.f.) $f(x; \theta)$. If we regard the parameter θ as unknown, we need to estimate it using $x_1, x_2, \ldots, x_n$.

Definition 4 Given observations $x_1, x_2, \ldots, x_n$ and unknown parameter θ, the *likelihood* of θ is the function

$$\begin{aligned} L(\theta) &= f(x;\theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta)\,. \end{aligned} \tag{1}$$

That is, L is the joint density (or mass) function, but regarded as a function of θ, for a fixed $x_1, x_2, \ldots, x_n$. The likelihood $L(\theta)$ is the probability (or probability density) of observing $x = x_1, x_2, \ldots, x_n$ if the unknown parameter is θ.

The *log-likelihood* is $l(\theta) = \log L(\theta)$ (The logarithm is to the base e).

The *maximum likelihood estimate* $\widehat{\theta}(x)$, is the value of θ that maximizes $L(\theta)$.

$\widehat{\theta}(X)$ is the *maximum likelihood estimator* (m.l.e.).

The idea of maximum likelihood is to estimate the parameter by the value of θ that gives the greatest likelihood to observations $x_1, x_2, \ldots, x_n$. That is, the θ for which the probability or probability density (1), is maximized.

In practice it is usually easiest to maximize $l(\theta)$, and since the taking of logarithms is a monotone function, this is equivalent to maximizing L.

Example 1 again

In this case the parameter of interest is μ.

$$
\begin{aligned}
L(\mu) &= \prod_{i=1}^{n} \frac{1}{\mu} e^{-\frac{x_i}{\mu}} \\
&= \frac{1}{\mu^n} e^{\left(-\frac{1}{\mu}\sum_{i=1}^{n} x_i\right)},
\end{aligned}
$$

and so

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

Then

$$\frac{\mathrm{d}l}{\mathrm{d}\mu} = -\frac{n}{\mu} + \frac{\sum_{i=1}^{n} x_i}{\mu^2},$$

and

$$\frac{\mathrm{d}l}{\mathrm{d}\mu} = 0 \Rightarrow \mu = \bar{x},$$

(which is a maximum).

Therefore, the maximum likelihood estimate of μ is $\bar{x}$.

The maximum likelihood estimator is $\bar{X}$.

Example

Consider a random variable X with a Bernoulli distribution with parameter p (this is the same as a Binomial$(1, p)$).

$$\begin{array}{l} \mathbf{P}(X = 1) = p\,, \\ \mathbf{P}(X = 0) = 1 - p\,. \end{array}$$

The probability mass function of X is

$$\begin{aligned} f(x;p) &= \mathbf{P}(X = x) \\ &= \begin{cases} p^x(1-p)^{1-x} & x = 0, 1\,. \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Suppose $X_1, X_2, \ldots, X_n$ is a random sample. Then, 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 $r = \sum_{i=1}^{n} x_i$.

The log-likelihood is

$$l(p) = r \log p + (n - r) \log(1 - p)$$

so,

$$l'(p) = \frac{r}{p} - \frac{n - r}{1 - p}\,.$$

Setting $l'(p)$ to zero gives $\widehat{p} = r/n$ (which is a maximum).

Therefore, the maximum likelihood estimator is

$$\widehat{p} = \frac{\sum_{i=1}^{n} X_i}{n}\,.$$

Example

Suppose we take a random sample of individuals from a population, and test their genetic type at a particular chromosomal location (called a "locus" in genetics). At this particular position, each chromosome in the population will have one of two possible variants, which we denote by A and a. Since each individual has two chromosomes (we receive one from each of our parents), then the type of a particular individual could be one of three so-called genotypes, AA, Aa, or aa, depending on whether they have 2, 1, or 0 copies of the A variant. (Note that order is not relevant, so there is no distinction between Aa and aA.)

There is a simple result, called the Hardy-Weinberg law, which states that under plausible assumptions, the genotypes AA, Aa and aa will occur with probabilities $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta)$ and $p_3 = (1-\theta)^2$ respectively, for some $0 \le \theta \le 1$.

Now suppose the random sample of n individuals contains:

$$\begin{array}{ll} x_1 & \text{of type AA;} \\ x_2 & \text{of type Aa;} \\ x_3 & \text{of type aa;} \end{array}$$

where $\sum_{i=1}^{3} x_i = n$.

Then the likelihood $L(\theta)$ is the probability that we observe (x_1, x_2, x_3) if we assign individuals to genotypes with probabilities (p_1, p_2, p_3). That is,

$$L(\theta) = \frac{n!}{x_1!x_2!x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}.$$

This is a multinomial distribution (the generalization of the binomial distribution in the setting when there are more than two possible outcomes).

Hence,

$$L(\theta) \propto \theta^{2x_1}\{\theta(1-\theta)\}^{x_2}(1-\theta)^{2x_3}\,,$$

and thus

$$\begin{aligned} l(\theta) = {} & (2x_1 + x_2)\log\theta \\ & +(x_2 + 2x_3)\log(1-\theta) + \text{const}\,, \end{aligned}$$

and

$$\begin{aligned} \frac{\mathrm{d}l}{\mathrm{d}\theta} = 0 \;\; & \Rightarrow \;\; \frac{2x_1 + x_2}{\theta} = \frac{x_2 + 2x_3}{1-\theta} \\ & \Rightarrow \;\; \theta = \frac{2x_1 + x_2}{2n}\,. \end{aligned}$$

[Do Sheet 1, Question 3 like this.]

Example

What if there is more than one parameter we wish to estimate?

Let $X_1, X_2, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. The likelihood is

$$
\begin{aligned}
L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{-1}{2\sigma^2}(x_i-\mu)^2)} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2)\,,
\end{aligned}
$$

and so

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

We just maximize l jointly over both μ and σ^2:

$$\begin{aligned}\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\\ \frac{\partial l}{\partial(\sigma^2)} &= -\frac{n}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2}\cdot\frac{1}{(\sigma^2)^2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

Solving $\frac{\partial l}{\partial \mu} = \frac{\partial l}{\partial(\sigma^2)} = 0$ simultaneously we obtain

$$\begin{aligned}\widehat{\mu} &= \bar{X}\,,\\ \widehat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\,.\end{aligned}$$

Here the m.l.e. of μ is just the sample mean.

Note that the m.l.e. of σ^2 is not quite the sample variance S^2, because of the divisor of n rather than $(n-1)$. However, the two will be numerically close unless n is small.

Try to avoid confusion over the terms "estimator" and "estimate".

An estimator is a rule for constructing an estimate: it is a function of the random variables $(X_1, X_2, \ldots, X_n)$ involved in the random sample.

In contrast, the estimate is the numerical value taken by the estimator for a particular data set: it is the value of the function evaluated at the data $x_1, x_2, \ldots, x_n$.

An estimate is just a number. An estimator is a function of random variables and hence is itself a random variable.

4. Parameter Estimation.

Earthquake Example

Let $X_1, X_2, \ldots, X_n$ be a random sample from the p.d.f.:

$$f(x;\mu) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note $\mathbf{E}(X_i) = \mu$.

Maximum likelihood gave $\widehat{\mu} = \bar{X}$.

Alternative estimators are:

(i) $\frac{1}{3}X_1 + \frac{2}{3}X_2$;

(ii) $X_1 + X_2 - X_3$;

(iii) $\frac{2}{n(n+1)}(X_1 + 2X_2 + \cdots + nX_n)$.

How should we decide between different estimators?

In general, suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with p.d.f. (or p.m.f.) $f(x;\theta)$. We want to estimate the unknown parameter θ using the observations $x_1, x_2, \ldots, x_n$.

Definition 5 A *statistic* is any function $T(X)$ of $X_1, X_2, \ldots, X_n$ that does not depend on θ.

An *estimator* of θ is any statistic $T(X)$ that we might use to estimate θ.

$T(x)$ is the *estimate* of θ, obtained via the estimator T, based on observations $x_1, x_2, \ldots, x_n$.

An estimator $T(X)$, e.g. $\bar{X}$, is a random variable. (It is a function of the random variables $X_1, X_2, \ldots X_n$.)

An estimate $T(x)$, e.g. $\bar{x}$, is a fixed number, based on data. (It is a function of the numbers $x_1, x_2, \ldots x_n$.)

Properties of Estimators

A good estimator should take values close to the true value of the parameter it is trying to estimate.

Definition 6 The estimator $T = T(X)$ is said to be *unbiased for* θ if $\mathbf{E}(T) = \theta$ for all θ.

That is, T is unbiased if it is 'correct on average'.

Earthquakes

The MLE is $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and we know that $\mathbf{E}(X_i) = \mu$. So,

and hence $\widehat{\mu}$ is unbiased for μ.

Note that our alternative estimator (i) is unbiased since

$$\mathbf{E}(\frac{1}{3}X_1 + \frac{2}{3}X_2) \ =$$

Similar calculations show that alternatives (ii) and (iii) are also unbiased.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a N(μ, σ^2) distribution. Consider

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

as an estimator of σ^2. (T is the MLE of σ^2 when μ and σ are unknown.)

$$\sum_{i=1}^{n}(X_i - \bar{X})^2$$

So,

$$\mathbf{E}(T) \;=$$

Hence, T is not unbiased. On average, T will underestimate σ^2. However, $\mathbf{E}(T) \to \sigma^2$ as $n \to \infty$, i.e. T is asymptotically unbiased.

Observe that the sample variance is

$$S^2 = \frac{n}{n-1} T \,,$$

and so

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

Therefore S^2 is unbiased for σ^2.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a uniform distribution on $[0, \theta]$, i.e.

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

What is the mle for θ? Is the mle unbiased?

1. We first calculate the likelihood:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(x;\theta) \\ &= \begin{cases} \frac{1}{\theta^n} & \text{if } 0 \le x_i \le \theta \text{ for all } i \\ 0 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{1}{\theta^n} & \text{if } \max_{1\le i\le n} x_i \le \theta \\ 0 & \text{otherwise.} \end{cases} \\ &= \frac{1}{\theta^n} I_{\{\max_{1\le i\le n} x_i \le \theta\}}. \end{aligned}$$

The maximum occurs when

$$\theta = \max_{1 \le i \le n} x_i \,.$$

Therefore, the MLE is

$$\widehat{\theta} = \max_{1 \le i \le n} X_i \,.$$

2. We now find the c.d.f. of $\widehat{\theta}$:

$$F(y) \ =$$

for $0 \leq y \leq \theta$, where the second last equality follows from the independence of the X_i.

So, the p.d.f. is

$$f(y) = F'(y) = \frac{ny^{n-1}}{\theta^n},$$

for $0 \leq y \leq \theta$.

3. So,

$$\begin{aligned}\mathbf{E}(\widehat{\theta}) &= \int_0^\theta y \cdot \frac{ny^{n-1}}{\theta^n}\mathrm{d}y \\ &= \frac{n}{n+1}\,\theta\,.\end{aligned}$$

Therefore, $\widehat{\theta}$ is not unbiased.

Since each $X_i < \theta$, we must have $\widehat{\theta} < \theta$ and so we should have expected $\mathbf{E}(\widehat{\theta}) < \theta$. Note, however, that the mle $\widehat{\theta}$ is asymptotically unbiased since $\mathbf{E}(\widehat{\theta}) \to \theta$ as $n \to \infty$.

In fact, under mild assumptions, MLEs are always asymptotically unbiased (one attractive feature of MLEs).

Further Properties of Estimators.

Definition 7 The *mean squared error* (MSE) of an estimator T is defined by:

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

The *bias* of T is defined by

$$b(T) = \mathbf{E}(T) - \theta\,.$$

Note that T is unbiased iff $b(T) = 0$.

Theorem 1

$$\text{MSE}(T) = \text{var}(T) + \{b(T)\}^2\,.$$

Proof: Let $\mu = \mathbf{E}(T)$. Then,

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

$\square$

MSE(T) is a measure of the ‘distance’ between an estimator T and the parameter θ, so good estimators have small MSE.

To minimize the MSE we have to consider the bias *and* the variance.

Unbiasedness alone is not particularly desirable - it is the combination of small variance and small bias which is important.

Important Results

It is always the case that

$$\mathbf{E}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = \\ a_1\mathbf{E}(X_1) + a_2\mathbf{E}(X_2) + \cdots + a_n\mathbf{E}(X_n)\,.$$

If $X_1, X_2, \ldots, X_n$ are independent then

$$\text{var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = \\ a_1^2\text{var}(X_1) + a_2^2\text{var}(X_2) + \cdots + a_n^2\text{var}(X_n)\,.$$

In particular, if $X_1, X_2, \ldots, X_n$ is a random sample with $\mathbf{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, then

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

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a uniform distribution on $[0, \theta]$, i.e.

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

Consider the estimator

$$T = \frac{2}{n} \sum_{i=1}^{n} X_i \, .$$

Then,

$$\begin{aligned}\mathbf{E}(T) &= \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) \\ &= \frac{2}{n}\cdot n\cdot\frac{\theta}{2} \\ &= \theta\,.\end{aligned}$$

Therefore T is unbiased.

Hence, since the X_i are independent, we have

$$\begin{aligned}\text{MSE}(T) &= \text{var}(T) \\ &= \end{aligned}$$

Now consider the maximum likelihood estimator $\widehat{\theta} = \max_{1 \le i \le n} X_i$.

Previously, we found that the p.d.f. of $\widehat{\theta}$ is

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

for $0 < y < \theta$.

We find:

$$\mathbf{E}(\widehat{\theta}) = \frac{n\theta}{n+1},$$

and

$$\mathrm{var}(\widehat{\theta}) = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

So

$$b(\widehat{\theta}) = \frac{n\theta}{n+1} - \theta = \frac{-\theta}{n+1}.$$

Thus,

$$\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathrm{var}(\widehat{\theta}) + \{b(\widehat{\theta})\}^2 \\
&= \frac{2\theta^2}{(n+1)(n+2)} \\
&\leq \frac{\theta^2}{3n} \\
&= \mathrm{MSE}(T)\,,
\end{aligned}$$

with strict inequality for $n > 2$.

So, $\widehat{\theta}$ is better in terms of MSE. In fact, it is much better since MSE decreases like $1/n^2$, rather than like $1/n$.

Note that $(\frac{n+1}{n})\widehat{\theta}$ is unbiased, but among estimators of the form $\lambda\widehat{\theta}$, MSE is minimized at

$$\lambda = \frac{n+2}{n+1}\,.$$

Estimation so far:

All of the estimates we have seen so far are point estimates (i.e. single numbers), e.g. $\bar{x}$, $\max_{1 \le i \le n} x_i$, s^2, ...

When an 'obvious' estimate exists, maximum likelihood will typically produce it (e.g. $\bar{x}$).

It can be shown that maximum likelihood estimators have good properties, especially when the sample size is large.

An important additional feature of maximum likelihood as a method for finding estimators is its generality: it works well when no 'obvious' estimate exists (e.g. Sheet 2).

Accuracy of an Estimate.

Earthquakes again.

We supposed that the p.d.f. of X_i was

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

for $x > 0$. Recall that in this case $\mathbf{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \mu^2$.

Suppose our point estimate of μ is $\bar{x} = 437$ days.

Better than the point estimate of 437 would be a range of plausible or believable values of μ, for example an interval (μ_1, μ_2) containing the point 437.

The mle is $\bar{X}$, and to understand the uncertainty in the mle, we could calculate

$$\text{var}(\bar{X}) = \frac{1}{n}\text{var}(X_1) = \frac{\mu^2}{n}\,.$$

Therefore, the standard deviation is

$$\text{s.d.}(\bar{X}) = \frac{\mu}{\sqrt{n}}\,.$$

Notice that the standard deviation depends on μ, which is unknown, and therefore we need to estimate it. Our estimate of the standard deviation is called the *standard error:*

$$\text{s.e.}(\bar{x}) = \frac{\bar{x}}{\sqrt{n}}\,.$$

(To find the standard error, we "plug in" to the formula for the standard deviation of the estimator an estimate for the unknown parameter. So here, we replace μ by $\bar{x}$.)

5. Confidence Intervals.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample of heights, where X_i is the height of the ith person.

Suppose we can assume $X_i \sim \mathrm{N}(\mu, \sigma_0^2)$ where μ is unknown and σ_0 is known.

Consider the interval $[a(X), b(X)]$. We would like to construct $a(X)$ and $b(X)$ so that:

1. The width of this interval is small.

2. The probability, $\mathbf{P}(a(X) \leq \mu \leq b(X))$ is large.

Note that the interval is a *random interval* since its endpoints $a(X)$ and $b(X)$ are random variables.

Definition 8 If $a(X)$ and $b(X)$ are two statistics, the interval $[a(X), b(X)]$ is called *a confidence interval for* θ with a confidence level of $1 - \alpha$ if

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

The interval $[a(x), b(x)]$ is called an *interval estimate*.

The random interval $[a(X), b(X)]$ is called an *interval estimator*.

The interval $[a(x), b(x)]$ is also called the $100(1-\alpha)\%$ *confidence interval for* θ.

(n.b. $a(X)$ and $b(X)$ do *not* depend on θ.)

The most commonly used values of α are $0.1, 0.05, 0.01$ (i.e confidence levels of 90%, 95%, 99%), but there is nothing special about any one confidence level.

Theorem 2 If $X_1, X_2, \ldots, X_n$ are independent random variables with $X_i \sim \mathrm{N}(\mu_i, \sigma_i)$ and $Y = \sum_{i=1}^n a_i X_i$ then $Y \sim \mathrm{N}(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2)$.

Proof: The proof is omitted.

Notation Let z_α be the constant such that if $Z \sim \mathrm{N}(0,1)$, then $\mathbf{P}(Z > z_\alpha) = \alpha$.

z_α is the "upper α point of $\mathrm{N}(0,1)$".

$\Phi(z_\alpha) = 1 - \alpha$, where Φ is the c.d.f. of a $\mathrm{N}(0,1)$. For example:

α	0.1	0.05	0.025	0.005
z_α	1.28	1.64	1.96	2.58

Example Suppose $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim \mathsf{N}(\mu, \sigma_0^2)$, where μ is unknown and σ_0 is known. The mle for μ is $\widehat{\mu} = \bar{X}$.

By Theorem 2,

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

Therefore,

$$\bar{X} \sim \mathsf{N}(\mu, \sigma_0^2/n)\,.$$

So, standardizing $\bar{X}$,

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

Hence,

$$\mathbf{P}(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha\,.$$

Therefore

$$\mathbf{P}(-\frac{z_{\alpha/2}\cdot\sigma_0}{\sqrt{n}} \leq \bar{X} - \mu \leq \frac{z_{\alpha/2}\cdot\sigma_0}{\sqrt{n}}) = 1 - \alpha\,,$$

and thus

$$\mathbf{P}(\bar{X} - \frac{z_{\alpha/2}\cdot\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\cdot\sigma_0}{\sqrt{n}}) = 1 - \alpha\,.$$

i.e. we have a random interval that contains the unknown μ with probability $1-\alpha$.

Hence

$$[\bar{x} - \frac{z_{\alpha/2}\cdot\sigma_0}{\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2}\cdot\sigma_0}{\sqrt{n}}]$$

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

Example (p82 of Daly *et al.*)

Suppose we measure the heights of 351 elderly women (i.e. $n = 351$), and suppose $\bar{x} = 160$ and $\sigma_0 = 6$.

The end points of a 95% confidence interval (i.e. $\alpha = 0.05$) are

$$160 \pm 1.96(6/\sqrt{351}),$$

giving

$$[159.4, 160.6]$$

as the 95% confidence interval for μ.

Notes

1. Our symmetric confidence interval for μ is sometimes called a *central* confidence interval for μ.

If c and d are constants such that $Z \sim \mathrm{N}(0,1)$, $\mathbf{P}(-c \leq Z \leq d) = 1 - \alpha$ then

$$\mathbf{P}(\bar{X} - \frac{d \cdot \sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{c \cdot \sigma_0}{\sqrt{n}}) = 1 - \alpha\,.$$

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

2. Note that

$$\mathbf{P}(\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \geq -z_\alpha) = 1-\alpha\,.$$

Therefore,

$$\mathbf{P}(\mu \leq \bar{X} + \frac{z_\alpha \sigma_0}{\sqrt{n}}) = 1-\alpha\,,$$

and then

$$\bar{x} + \frac{z_\alpha \sigma_0}{\sqrt{n}}$$

is an *upper* $1-\alpha$ *confidence limit for* μ.

Similarly,

$$\mathbf{P}(\mu \geq \bar{X} - \frac{z_\alpha \sigma_0}{\sqrt{n}}) = 1-\alpha\,,$$

and

$$\bar{x} - \frac{z_\alpha \sigma_0}{\sqrt{n}}$$

is an *lower* $1-\alpha$ *confidence limit for* μ.

Interpretation of a Confidence Interval

Since the parameter θ is *fixed*, the interval

$$[a(x), b(x)]$$

either definitely does *or* definitely does not contain θ.

So it is wrong to say that $[a(x), b(x)]$ contains θ with probability $1 - \alpha$.

Rather, if we repeatedly obtain new data, $X^{(1)}, X^{(2)}, \ldots$ say, and construct intervals

$$[a(X^{(i)}), b(X^{(i)})],$$

for each data set, then a proportion $1 - \alpha$ of the intervals constructed will contain θ.

(That is, it is the endpoints $a(X)$ and $b(X)$ that are random variables, not the parameter θ.)

The Central Limit Theorem (CLT).

We already know that if $X_1, X_2, \ldots, X_n$ are independent random variables with distribution $\mathrm{N}(\mu, \sigma^2)$ then

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

Theorem 3 (Central Limit Theorem) Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables, each with mean μ and variance σ^2. Then, the standardized random variables

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

satisfy, as $n \to \infty$,

$$\mathbf{P}(Z_n \leq x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \mathrm{d}u\,, \qquad (2)$$

for $x \in \mathbf{R}$.

The right hand side of (2) is just $\Phi(x)$, the cumulative distribution function of a standard normal random variable, $\mathrm{N}(0, 1)$.

The CLT says $\mathbf{P}(Z_n \leq x) \to \Phi(x)$ for $x \in \mathbf{R}$.

So, for n large, $Z_n \approx \mathrm{N}(0, 1)$ where $\approx$ means "approximately equal in distribution". The important point about the result is that it holds *whatever* the distribution of the X's. In other words, whatever the distribution of the data, the sample mean will be approximately normally distributed when the sample size n is large. (Usually for $n > 30$ the distribution of the sample mean will be close to normal.)

Confidence Intervals Using the CLT.

Example Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with mean μ and p.d.f.

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

for $x > 0$. e.g. X_i = the survival time of patient i.

It is straightforward to check that $\mathbf{E}(X_i) = \mu$ and $\text{var}(X_i) = \mu^2$.

So, since $\sigma^2 = \mu^2$, by the CLT we obtain (for large n),

$$\frac{\bar{X} - \mu}{\mu/\sqrt{n}} \approx \mathrm{N}(0, 1)\,. \qquad (3)$$

For clarity, we set $z = z_{\alpha/2}$. So,

$$\mathbf{P}(-z \leq \frac{\bar{X} - \mu}{\mu/\sqrt{n}} \leq z) \approx 1 - \alpha\,.$$

Therefore

$$\mathbf{P}(\mu(1-\frac{z}{\sqrt{n}}) \le \bar{X} \le \mu(1+\frac{z}{\sqrt{n}})) \approx 1-\alpha$$

and thus,

$$\mathbf{P}(\frac{\bar{X}}{1+\frac{z}{\sqrt{n}}} \le \mu \le \frac{\bar{X}}{1-\frac{z}{\sqrt{n}}}) \approx 1-\alpha\,.$$

Hence

$$\left[\frac{\bar{x}}{1+\frac{z}{\sqrt{n}}},\ \frac{\bar{x}}{1-\frac{z}{\sqrt{n}}}\right]$$

is a confidence interval with confidence level of approximately $1-\alpha$. Note that this is not exact because (3) is an approximation.

Opinion Polls. In a poll preceding the 2005 general election, 519 of 1105 voters said they would vote Labour.

With $n = 1105$, suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli(p) distribution:

$$\mathbf{P}(X_i = 1) = p = 1 - \mathbf{P}(X_i = 0)\,.$$

The mle of p is $\widehat{p} = \bar{X}$. We can easily check that $\mathbf{E}(X_i) = p$ and $\mathrm{var}(X_i) = p(1-p) = \{\sigma(p)\}^2$ say.

Then by the CLT,

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

and so

$$\mathbf{P}(-z_{\alpha/2} \leq \frac{\bar{X} - p}{\sigma(p)/\sqrt{n}} \leq z_{\alpha/2}) \approx 1 - \alpha\,,$$

or

$$\mathbf{P}(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma(p)}{\sqrt{n}} \leq p \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma(p)}{\sqrt{n}}) \approx 1 - \alpha\,.$$

The endpoints of the interval thus obtained are unknown, since $\sigma(p)$ depends on p.

(i) We could solve the quadratic inequality to find $\mathbf{P}(a(X) \le p \le b(X)) \approx 1 - \alpha$ where a and b don't depend on p.

(ii) Our estimate of p is $\bar{x}$, so we could estimate $\sigma(p)$ by the standard error: $\sigma(\bar{x}) = \sqrt{\bar{x}(1-\bar{x})}$, giving endpoints of

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\,.$$

With $n = 1105$, and $\bar{x} = 519/1105$, an approximate 95% confidence interval is $[0.44, 0.50]$.

We have used two approximations here:

(a) We used a normal approximation (CLT).
(b) We approximated $\sigma(p)$ by $\sigma(\bar{x})$.

Both are good approximations.

Opinion polls often mention "$\pm 3\%$ error". Note that

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

since $p(1-p)$ has its maximum at $p = \frac{1}{2}$. Then, we have

since $\sigma^2(p) \leq \frac{1}{4}$.

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

Suppose $X_1, X_2, \ldots, X_n$ is a random sample with $\mathbf{E}(X_i) = \theta$ and $\text{var}(X_i) = \{\sigma(\theta)\}^2$, for some known function σ.

Then, for large n, the CLT we have

$$\mathbf{P}(-z_{\alpha/2} \leq \frac{\bar{X} - \theta}{\sigma(\theta)/\sqrt{n}} \leq z_{\alpha/2}) \approx 1 - \alpha\,,$$

or

$$\mathbf{P}(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma(\theta)}{\sqrt{n}} \leq \theta \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma(\theta)}{\sqrt{n}}) \approx 1 - \alpha\,.$$

As $\sigma(\theta)$ depends on θ, replace it by the estimate $\sigma(\bar{x})$, giving a confidence interval with endpoints

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma(\bar{x})}{\sqrt{n}}\,.$$

This uses approximations (a) and (b) above.

6. Linear Regression.

Suppose we measure two variables in the same population:
x, the 'explanatory variable'
y, the 'response variable'

Example 1 Suppose $x =$ the age of a child and $y =$ the height of a child.

Example 2 Suppose $x =$ the latitude of a (Northern Hemisphere) city and $y =$ the average temperature in the city.

We may ask the following questions:

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 a linear regression:

$$y = \alpha + \beta x + \text{‘error’}\,.$$

Note that a linear relationship does not necessarily imply that x causes y.

We suppose that

$$Y_i = \alpha + \beta x_i + \epsilon_i \,, \qquad (4)$$

for $i = 1, 2, \ldots, n$, where

$x_1, x_2, \ldots x_n$	are known constants,
$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$	are i.i.d. N$(0, \sigma^2)$: 'random errors',
α, β	are unknown parameters.

Note The Y_i are random variables, e.g. denoting the average temperature in city i. (y_i is the observed value of Y_i.)

The x_i do *not* correspond to random variables, e.g. x_i is the latitude of city i.

Two common objectives are:

1. To estimate α and β (i.e. find the 'best' straight line).

2. To determine whether the mean of Y really depends on x? (i.e. is $\beta \neq 0$?)

We focus on estimating α and β, and suppose that σ^2 is known.

From (4),

$$Y \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)\,.$$

If $f(y_i; \alpha, \beta)$ is a normal p.d.f. with mean $\alpha + \beta x_i$ and variance σ^2, then the likelihood of observing $y_1, y_2, \ldots, y_n$ is

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

and so

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

Maximizing $l(\alpha, \beta)$ over α and β is equivalent to minimizing the sum of squares

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

Thus, the MLEs of α and β are also called the *least squares estimators*.

We want to minimize $\sum_{i=1}^{n}$(vertical distance)^2.

Now,

$$
\begin{aligned}
Y_i &= \alpha + \beta x_i + \epsilon_i \\
&= \alpha + \beta \bar{x} + \beta(x_i - \bar{x}) + \epsilon_i \\
&= a + b w_i + \epsilon_i \, ,
\end{aligned}
$$

where $a = \alpha + \beta \bar{x}$, $b = \beta$ and $w_i = x_i - \bar{x}$.

We work in terms of the new parameters a and b, and note that $\sum_{i=1}^{n} w_i = 0$.

The MLEs/least squares estimators of a and b minimize

$$
S(a,b) = \sum_{i=1}^{n} (y_i - a - b w_i)^2 \, .
$$

Since S is a function of two variables, a and b, we use partial differentiation to minimize:

$$
\begin{aligned}
\frac{\partial S}{\partial a} &= -2 \sum_{i=1}^{n} (y_i - a - b w_i) \, , \\
\frac{\partial S}{\partial b} &= -2 \sum_{i=1}^{n} w_i (y_i - a - b w_i) \, .
\end{aligned}
$$

So, if $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$ then

$$\begin{aligned} \sum_{i=1}^{n} y_i &= na + b \sum_{i=1}^{n} w_i \,, \\ \sum_{i=1}^{n} w_i y_i &= a \sum_{i=1}^{n} w_i + b \sum_{i=1}^{n} w_i^2 \,. \end{aligned}$$

Hence, the MLEs are

$$\begin{aligned} \widehat{a} &= \bar{Y} \,, \\ \widehat{b} &= \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i^2} \,. \end{aligned}$$

If we had minimized $S(\alpha, \beta)$ over α and β, we would have obtained

$$\begin{aligned} \widehat{\alpha} &= \bar{Y} - \widehat{\beta}\bar{x} \,, \\ \widehat{\beta} &= \widehat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \,. \end{aligned}$$

The *fitted regression line* is $y = \widehat{\alpha} + \widehat{\beta}x$.

The point $(\bar{x}, \bar{y})$ always lies on this line.

Further Aspects of Linear Regression.

Regression Through the Origin.

We could choose to fit the best line of the form $y = \beta x$. The relevant model is:

$$Y_i = \beta x_i + \epsilon_i \,,$$

where $i = 1, 2, \ldots, n$, with $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ i.i.d. N$(0, \sigma^2)$, x_i known constants and β an unknown parameter.

We would estimate β by minimizing

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

Polynomial Regression.

We could include an x^2 term in the model:

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i\,,$$

and estimate α, β, γ by minimizing

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

The simplest way to see if a linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$ is appropriate is to plot the points (x_i, y_i), $i = 1, 2, \ldots, n$.

Although computer packages may be used to fit a regression (i.e. find the MLEs of α and β), you should always plot the points to see whether it is sensible to describe the variation in Y as a linear function of x.

Consider the model

$$Y_i = a + bw_i + \epsilon_i\,,$$

where $w_i = x_i - \bar{x}$.

We have

$$\begin{aligned} \widehat{a} &= \frac{1}{n}\sum_{i=1}^{n} Y_i\,, \\ \widehat{b} &= \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i^2}\,. \end{aligned}$$

Are these MLEs unbiased?

Note $\mathbf{E}(Y_i) = a + bw_i$, so

$$\begin{aligned}
\mathbf{E}(\hat{a}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(a + bw_i) \\
&= \frac{1}{n}(na + b\sum_{i=1}^{n} w_i) \\
&= a\,,
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}(\hat{b}) &= \frac{1}{\sum_{i=1}^{n} w_i^2}\cdot\mathbf{E}(\sum_{i=1}^{n} w_i Y_i) \\
&= \frac{1}{\sum_{i=1}^{n} w_i^2}\cdot\sum_{i=1}^{n} w_i(a + bw_i) \\
&= \frac{1}{\sum_{i=1}^{n} w_i^2}\cdot(a\sum_{i=1}^{n} w_i + b\sum_{i=1}^{n} w_i^2) \\
&= b\,.
\end{aligned}$$

We can also calculate the variances of $\widehat{a}$ and $\widehat{b}$. First,

$$\text{var}(\widehat{a}) = \text{var}(\bar{Y}) = \frac{\sigma^2}{n},$$

and

$$\begin{aligned}
\text{var}(\widehat{b}) &= \frac{1}{\left(\sum_{i=1}^{n} w_i^2\right)^2} \cdot \text{var}(\sum_{i=1}^{n} w_i Y_i) \\
&= \frac{1}{\left(\sum_{i=1}^{n} w_i^2\right)^2} \cdot \sum_{i=1}^{n} w_i^2 \cdot \text{var}(Y_i) \\
&= \frac{1}{\left(\sum_{i=1}^{n} w_i^2\right)^2} \cdot \sum_{i=1}^{n} w_i^2 \cdot \sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} w_i^2}.
\end{aligned}$$

In the models:

1. $Y_i = \alpha + \beta x_i + \epsilon_i$;

2. $Y_i = a + b(x_i - \bar{x}) + \epsilon_i$,

$b = \beta$ is usually the parameter of interest. (We are rarely interested in a or α.)

Confidence Interval for β

We note that since

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i^2}$$

is a linear combination of $Y_1, Y_2, \ldots, Y_n$, $\widehat{\beta}$ is normally distributed.

So, from the above calculations, $\widehat{\beta} \sim \mathrm{N}(\beta, \sigma_{\beta}^2)$ where $\sigma_{\beta}^2 = \sigma^2 / \sum_{i=1}^{n} w_i^2$.

Hence,

$$\frac{\widehat{\beta} - \beta}{\sigma_\beta} \sim \mathsf{N}(0, 1)\,.$$

So,

$$\mathbf{P}(-z_{\alpha/2} \leq \frac{\widehat{\beta} - \beta}{\sigma_\beta} \leq z_{\alpha/2}) = 1 - \alpha$$

(N.B. $\alpha = 0.05$ is *NOT* a regression parameter) and therefore

$$\mathbf{P}(\widehat{\beta} - z_{\alpha/2} \cdot \sigma_\beta \leq \beta \leq \widehat{\beta} + z_{\alpha/2} \cdot \sigma_\beta) = 1 - \alpha\,.$$

If we assume σ is known then σ_β is also known, and therefore the endpoints of a $1-\alpha$ confidence interval for β are

$$\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i^2} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{\sum_{i=1}^{n} w_i^2}} . \qquad (5)$$

In practice, however, σ^2 is rarely known.

It turns out that an unbiased estimate of σ^2 is

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

If we use the square root of this in place of σ in (5), we get an approximate $100(1-\alpha)\%$ confidence interval for β.

As usual, n must be large for a good approximation. In fact, it would be more accurate to use a t-distribution, than the $\mathrm{N}(0,1)$ distribution.