2. Parameter Estimation and Goodness of Fit – Part One

In the frequentist approach, parameter estimation requires the definition of a lot of mathematical machinery

- **Random sample of size M, drawn from underlying pdf**

2. Parameter Estimation and Goodness of Fit – Part One

In the frequentist approach, parameter estimation requires the definition of a lot of mathematical machinery

- **Random sample of size M , drawn from underlying pdf**
- **Sampling distribution, derived from underlying pdf**

  (depends on underlying pdf, and on M)

2. Parameter Estimation and Goodness of Fit – Part One

In the frequentist approach, parameter estimation requires the definition of a lot of mathematical machinery

- **Random sample of size M , drawn from underlying pdf**

- **Sampling distribution, derived from underlying pdf**

  (depends on underlying pdf, and on M)

- **Define an *estimator* – function of sample used to estimate parameters of the pdf**

2. Parameter Estimation and Goodness of Fit – Part One

In the frequentist approach, parameter estimation requires the definition of a lot of mathematical machinery

- **Random sample of size M , drawn from underlying pdf**
- **Sampling distribution, derived from underlying pdf**

  (depends on underlying pdf, and on M)
- **Define an *estimator* – function of sample used to estimate parameters of the pdf**
- **Hypothesis test – to decide if estimator is 'acceptable', for the given sample size**

2. Parameter Estimation and Goodness of Fit – Part One

In the frequentist approach, parameter estimation requires the definition of a lot of mathematical machinery

- **Random sample** of size M, drawn from underlying pdf

How do we decide what makes an 'acceptable' estimator?

estimate parameters of the pdf

- **Hypothesis test** – to decide if estimator is 'acceptable', for the given sample size

Example: measuring the wavelength of a spectral line

True wavelength = z_0 (fixed but unknown parameter)

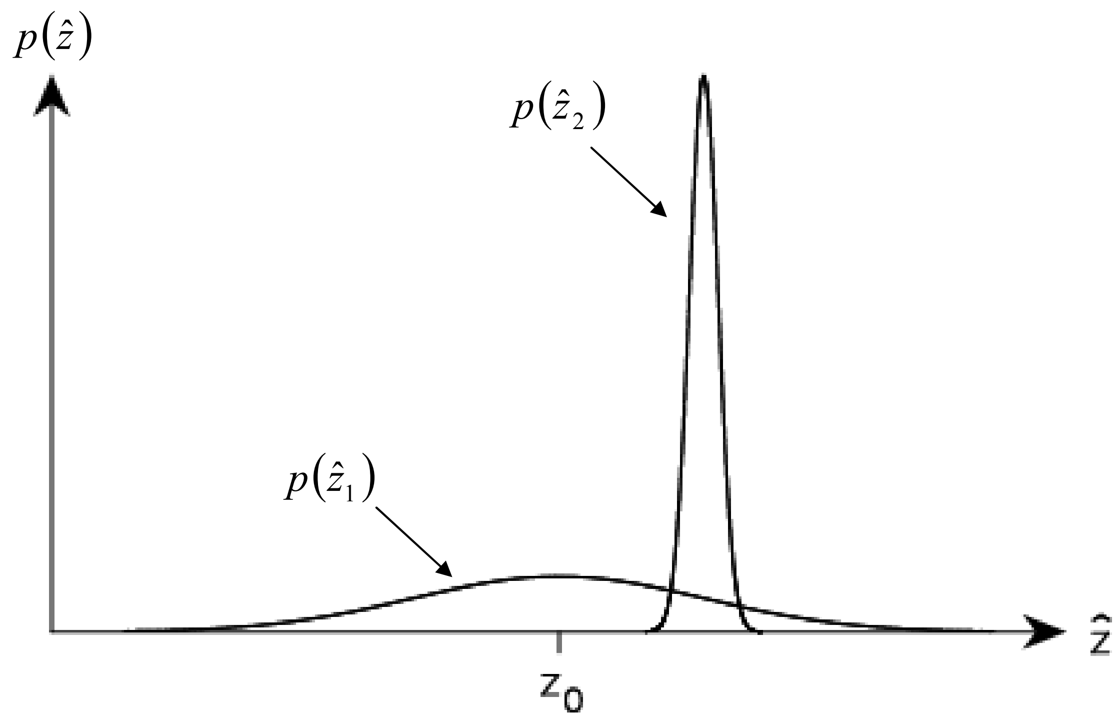
Compute sampling distribution for $\hat{z}_1$ and $\hat{z}_2$, modelling errors

1. Low dispersion spectrometer

Unbiased:

Repeat observation a large number of times $\Rightarrow$ average estimate is equal to z_0

$$E(\hat{z}_1) = \int \hat{z}_1 \, p(\hat{z}_1; z_0) \, d\hat{z}_1 = z_0$$



Example: measuring the wavelength of a spectral line

True wavelength = z_0 (fixed but unknown parameter)

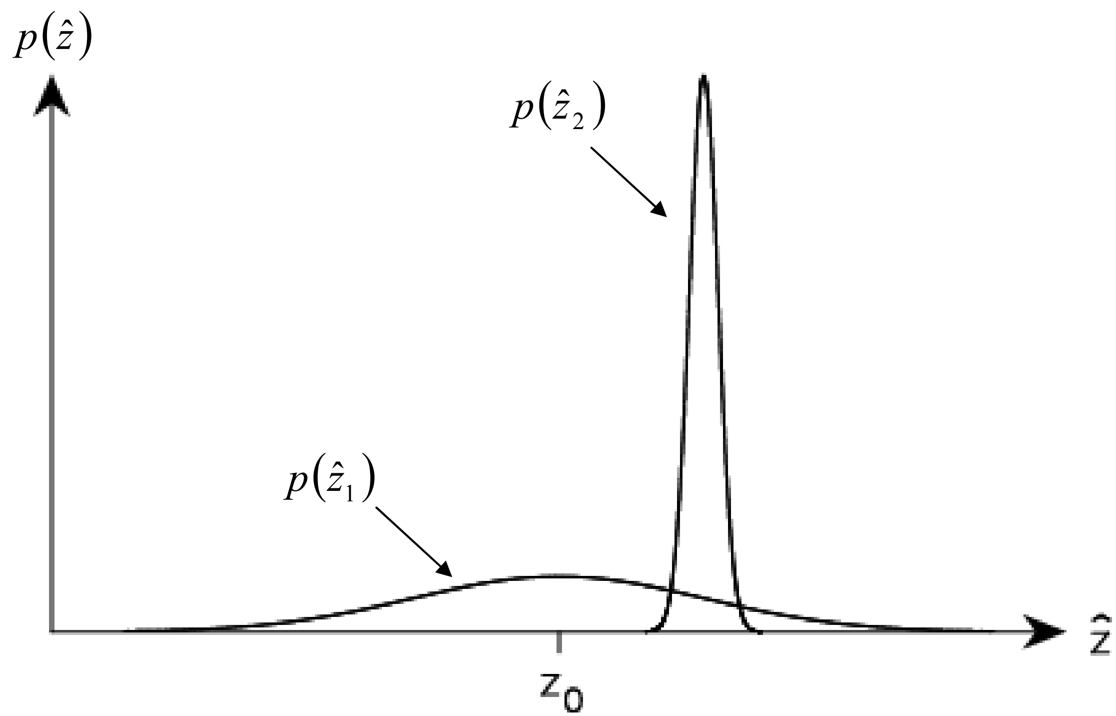
Compute sampling distribution for $\hat{z}_1$ and $\hat{z}_2$, modelling errors

1. Low dispersion spectrometer

   Unbiased:

   Repeat observation a large number of times
   $\Rightarrow$ average estimate is equal to z_0

$$E(\hat{z}_1) = \int \hat{z}_1 \, p(\hat{z}_1; z_0) \, d\hat{z}_1 = z_0$$



BUT $\mathrm{var}[\hat{z}_1]$ **is large**

Example: measuring the wavelength of a spectral line

True wavelength = z_0 (fixed but unknown parameter)

Compute sampling distribution for $\hat{z}_1$ and $\hat{z}_2$, modelling errors

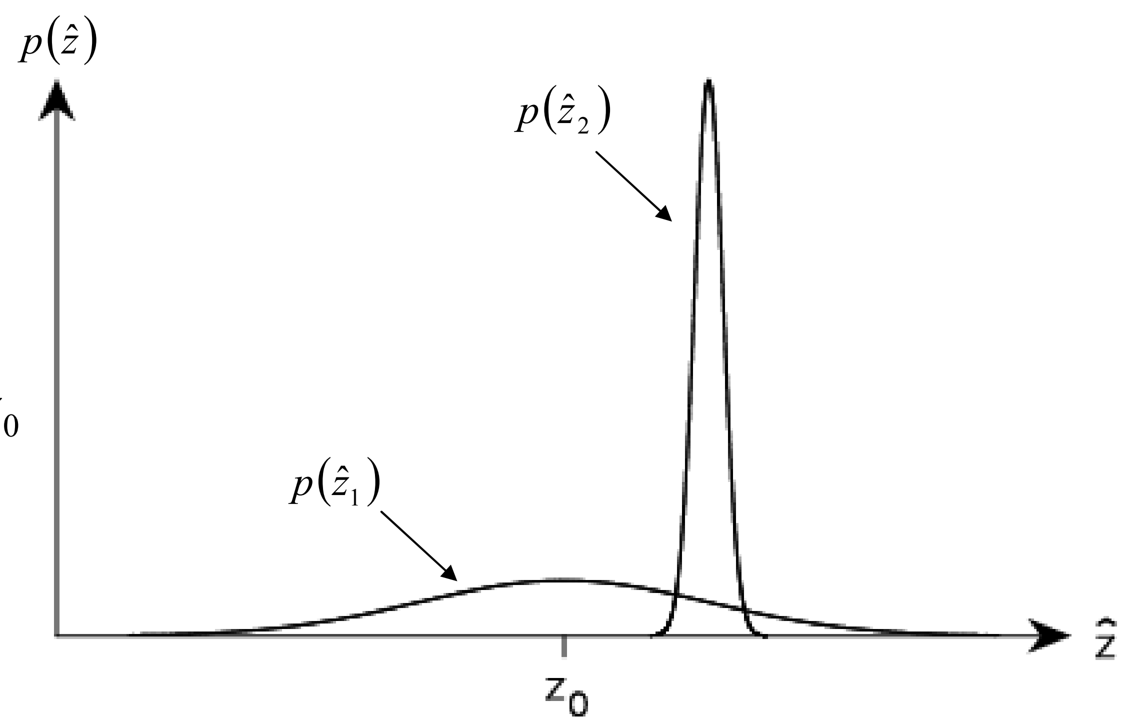
2. High dispersion spectrometer

but faulty physicist!
(e.g. wrong calibration)

Biased:

$$E(\hat{z}_2) = \int \hat{z}_2\, p(\hat{z}_2; z_0)\, d\hat{z}_2 \neq z_0$$

BUT $\mathrm{var}[\hat{z}_2]$ is small



Example: measuring the wavelength of a spectral line

True wavelength = z_0 (fixed but unknown parameter)

Compute sampling distribution for $\hat{z}_1$ and $\hat{z}_2$, modelling errors

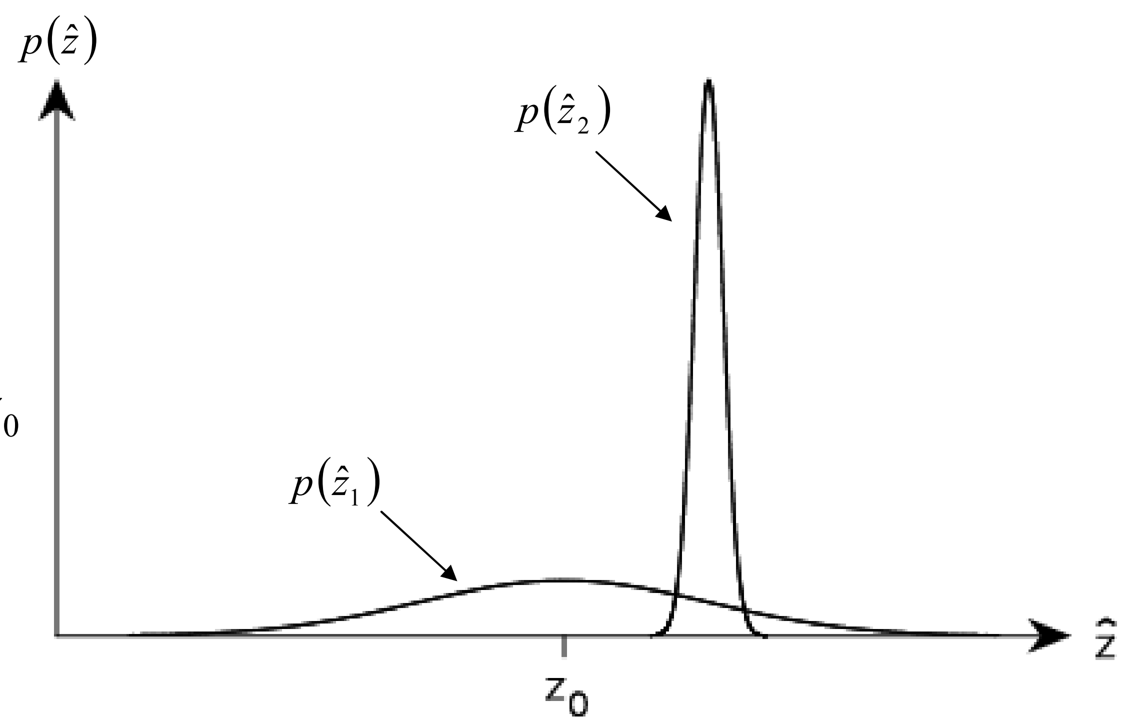
2. High dispersion spectrometer

   but faulty physicist!
   (e.g. wrong calibration)

   Biased:

$$E(\hat{z}_2) = \int \hat{z}_2 \, p(\hat{z}_2; z_0) \, d\hat{z}_2 \neq z_0$$

BUT $\text{var}[\hat{z}_2]$ is small



Better choice of estimator (if we can correct bias)

The Sample Mean

$\{x_1, \ldots, x_M\}$ = random sample from pdf $p(x)$ with mean μ and variance σ^2

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$$ = **sample mean**

Can show that $E(\hat{\mu}) = \mu$ **unbiased estimator**

But bias is defined formally in terms of an infinite set of randomly chosen samples, each of size M.

The Sample Mean

$\{x_1, \ldots, x_M\}$ = random sample from pdf $p(x)$ with mean μ and variance σ^2

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$$ = **sample mean**

Can show that $E(\hat{\mu}) = \mu$ **unbiased estimator**

But bias is defined formally in terms of an infinite set of randomly chosen samples, each of size M.

What can we say with a finite number of samples, each of finite size?

The Sample Mean

$\{x_1, \ldots, x_M\}$ = random sample from pdf $p(x)$ with mean μ and variance σ^2

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} x_i$$ = **sample mean**

Can show that

$$E(\hat{\mu}) = \mu$$

unbiased estimator

and

$$\mathrm{var}[\hat{\mu}] = \frac{\sigma^2}{M}$$

as sample size increases, sample mean increasingly concentrated near to true mean

Linear correlation

Given sampled data $\{(x_i, y_i); i = 1,\ldots,n\}$ we can **estimate** the linear correlation between the variables as follows:

Pearson's product moment correlation coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\quad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Sample mean

Sample standard deviation

Linear correlation

Given sampled data $\{(x_i, y_i); i = 1, \ldots, n\}$ we can **estimate** the linear correlation between the variables as follows:

Pearson's product moment correlation coefficient

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ $\qquad s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

Sample mean

Sample standard deviation

If $p(x, y)$ is bivariate normal then r is an estimator of ρ

Linear correlation

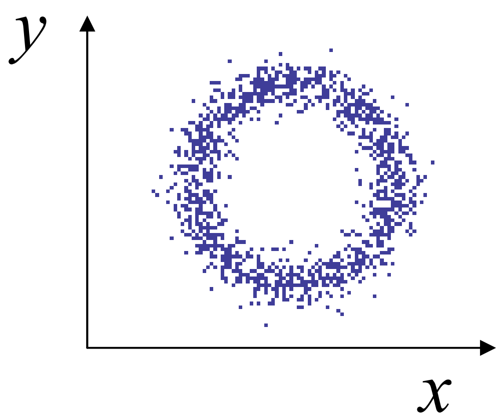
We can also rewrite the formula for r in the slightly simpler forms:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

or

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Question 4: Estimate r for the sample $\{(x,y)\}$ data shown in the graph below



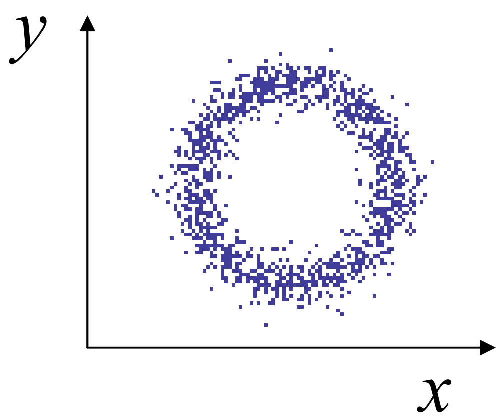
A $r = 0$

B $r = 0.5$

C $r = 1$

D $r = -1$

Question 4: Estimate r for the sample $\{(x, y)\}$ data shown in the graph below



A $r = 0$

B $r = 0.5$

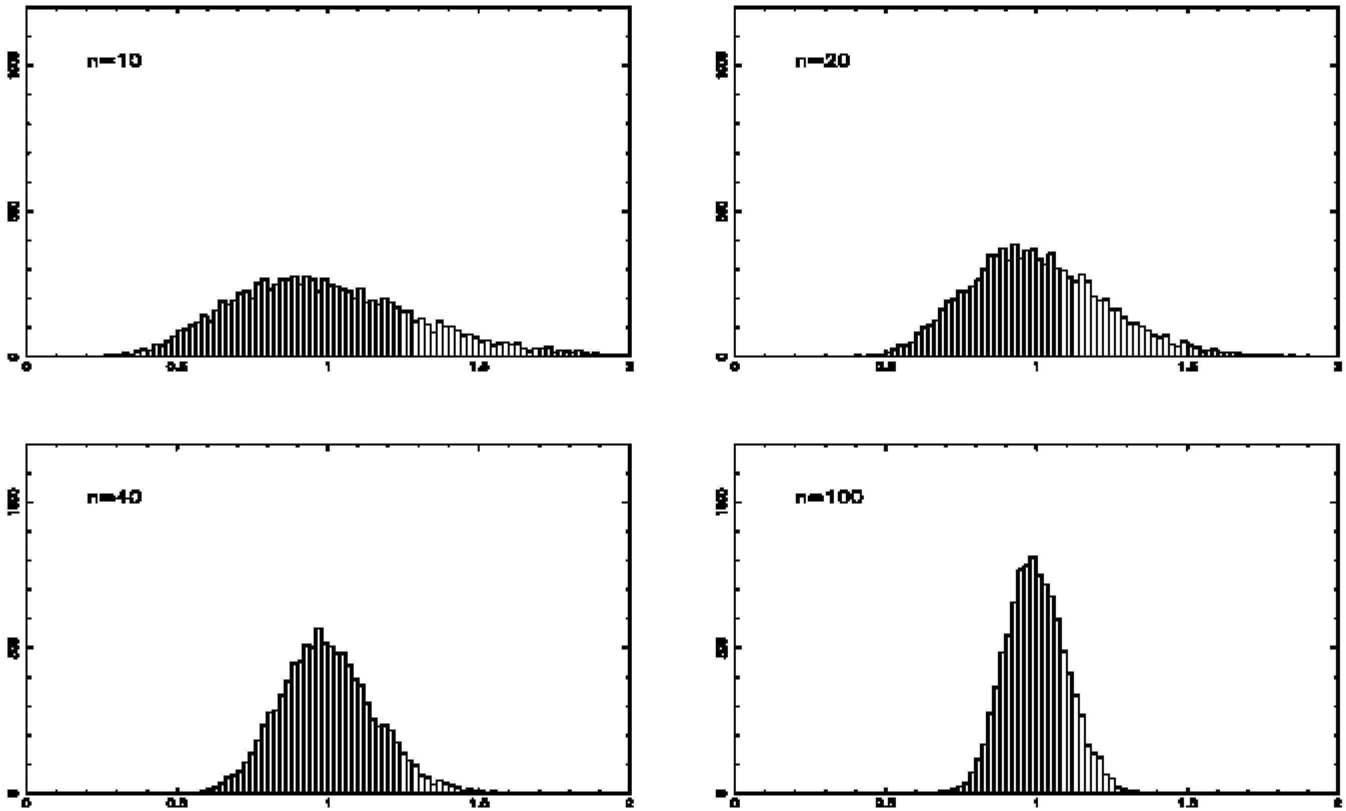
C $r = 1$

D $r = -1$

The Central Limit Theorem

For *any* pdf with finite variance σ^2, as $M \to \infty$

$\hat{\mu}$ follows a normal pdf with mean μ and variance σ^2 / M



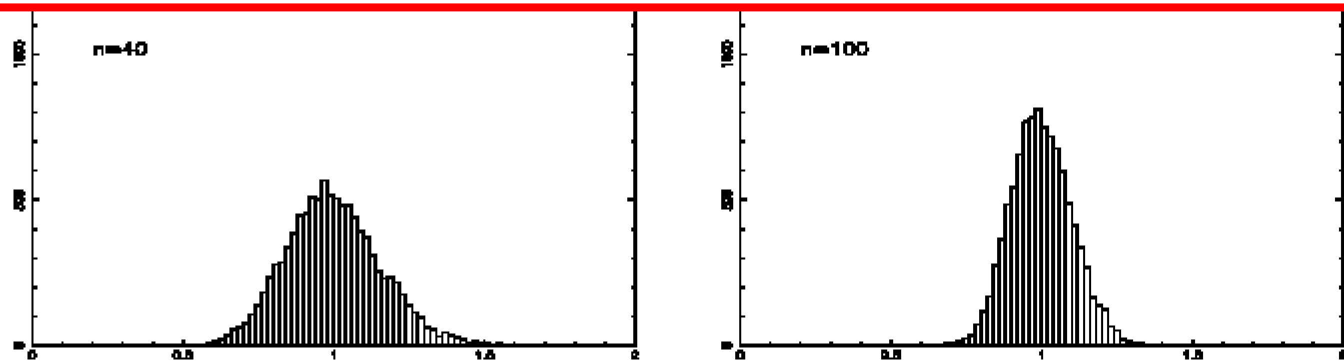
The Central Limit Theorem

For *any* pdf with finite variance σ^2, as $M \rightarrow \infty$

$\hat{\mu}$ follows a normal pdf with mean μ and variance σ^2 / M

Explains importance of normal pdf in statistics.

But still based on asymptotic behaviour of an infinite ensemble of samples that we didn't actually observe!



The Central Limit Theorem

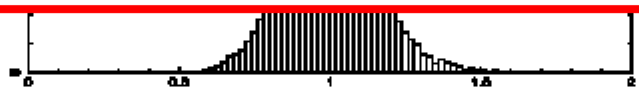
For *any* pdf with finite variance σ^2, as $M \to \infty$

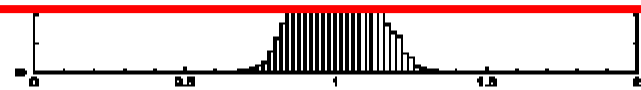
$\hat{\mu}$ follows a normal pdf with mean μ and variance σ^2 / M

Explains importance of normal pdf in statistics.

But still based on asymptotic behaviour of an infinite ensemble of samples that we didn't actually observe!

No 'hard and fast' rule for defining 'good' estimators. FPT invokes a number of principles – e.g. least squares, maximum likelihood





Method of Least Squares

- ‘workhorse’ method for fitting lines and curves to data in the physical sciences
- method often encountered (as a ‘black box’?) in elementary courses
- useful demonstration of underlying statistical principles
- simple illustration of fitting straight line to (x,y) data

Ordinary Linear Least Squares

Suppose that the scatter in a plot of $\{x_i, y_i\}$ is assumed to arise from errors in only one of the two variables. This case is called **Ordinary Least Squares**. We then call x the **independent variable**, and y the **dependent variable**. Thus we suppose that we can write, for each data point:-

$$y_i = a + bx_i + \epsilon_i$$

where ϵ_i is known as the **residual** of the i^{th} data point – i.e. the difference between the observed value of y_i, and the value predicted by the best-fit straight line, characterised by parameters a and b.

Ordinary Linear Least Squares

We assume that the $\{\epsilon_i\}$ are an independently and identically distributed random sample from some underlying pdf with mean zero and variance σ^2 – i.e. the residuals are equally likely to be positive or negative and all have equal variance.

The **least squares estimators** of a and b minimise

$$S = \chi^2(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

and $\hat{a}_{\text{LS}}$ and $\hat{b}_{\text{LS}}$ satisfy

$$\frac{\partial S}{\partial a} = 0 \quad \text{when} \quad a = \hat{a}_{\text{LS}} \qquad \frac{\partial S}{\partial b} = 0 \quad \text{when} \quad b = \hat{b}_{\text{LS}}$$

Ordinary Linear Least Squares

We assume that the $\{\epsilon_i\}$ are an independently and identically distributed random sample from some underlying pdf with mean zero and variance σ^2 – i.e. the residuals are equally likely to be positive or negative and all have equal variance.

$$S = \sum_{i=1}^{n} \varepsilon_i^2$$

The **least squares estimators** of a and b minimise

$$S \quad = \quad \chi^2(a,b) \quad = \quad \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

and $\hat{a}_{\rm LS}$ and $\hat{b}_{\rm LS}$ satisfy

$$\frac{\partial S}{\partial a} = 0 \quad \text{when} \quad a = \hat{a}_{\rm LS} \qquad\qquad \frac{\partial S}{\partial b} = 0 \quad \text{when} \quad b = \hat{b}_{\rm LS}$$

Solving these equations, $\hat{a}_{\mathrm{LS}}$ and $\hat{b}_{\mathrm{LS}}$ are given by

$$\hat{a}_{\mathrm{LS}} = \frac{\sum y_i \sum x_i^2 - \sum y_i x_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\hat{b}_{\mathrm{LS}} = \frac{n \sum y_i x_i - \sum y_i \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

We can show that

$$E\left(\hat{a}_{LS}\right) = a_{LS}$$

$$E\left(\hat{b}_{LS}\right) = b_{LS}$$

i.e. LS estimators are ***unbiased***.

Also

$$\mathrm{var}(\hat{a}_{\mathrm{LS}}) = \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\mathrm{var}(\hat{b}_{\mathrm{LS}}) = \frac{\sigma^2 n}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

and

$$\mathrm{cov}(\hat{a}_{\mathrm{LS}}, \hat{b}_{\mathrm{LS}}) = \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

We can show that

$$E(\hat{a}_{LS}) = a_{LS}$$

$$E(\hat{b}_{LS}) = b_{LS}$$

i.e. LS estimators are ***unbiased***.

Also

$$\mathrm{var}(\hat{a}_{\mathrm{LS}}) = \frac{\sigma^2 \sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\mathrm{var}(\hat{b}_{\mathrm{LS}}) = \frac{\sigma^2 n}{n \sum x_i^2 - (\sum x_i)^2}$$

and

$$\mathrm{cov}(\hat{a}_{\mathrm{LS}}, \hat{b}_{\mathrm{LS}}) = \frac{-\sigma^2 \sum x_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Choosing the $\{x_i\}$ so that $\sum x_i = 0$ we can make $\hat{a}_{LS}$ and $\hat{b}_{LS}$ independent.

Weighted Linear Least Squares

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Define

$$S = \chi^2(a,b) = \sum_{i=1}^{n} \left[\frac{y_i - (a + bx_i)}{\sigma_i} \right]^2$$

Again we find Least Squares estimators of a and b satisfying

$$\frac{\partial S}{\partial a} = 0 \qquad \frac{\partial S}{\partial b} = 0$$

Solving, we find

$$\hat{a}_{\mathrm{WLS}} = \frac{\sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \sum \frac{y_i x_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2}$$

$$\hat{b}_{\mathrm{WLS}} = \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{y_i x_i}{\sigma_i^2} - \sum \frac{y_i}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2}$$

Also

$$\mathrm{var}(\hat{a}_{\mathrm{WLS}}) = \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

$$\mathrm{var}(\hat{b}_{\mathrm{WLS}}) = \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

$$\mathrm{cov}\left(\hat{a}_{\mathrm{WLS}}, \hat{b}_{\mathrm{WLS}}\right) = \frac{-\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

Also

$$\mathrm{var}(\hat{a}_{\mathrm{WLS}}) = \frac{\sum \frac{x_i^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

$$\mathrm{var}(\hat{b}_{\mathrm{WLS}}) = \frac{\sum \frac{1}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

$$\mathrm{cov}\left(\hat{a}_{\mathrm{WLS}}, \hat{b}_{\mathrm{WLS}}\right) = \frac{-\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2}$$

In the case where σ_i^2 is constant, for all i, these formulae reduce to those for the unweighted case.

Extensions and Generalisations

- Errors on *both* variables?

Need to modify merit function accordingly.

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2}$$

Renders equations *non-linear*; no simple analytic solution!

See e.g. Numerical Recipes 15.3

Extensions and Generalisations

- General linear models?

e.g.

$$y(x) = a_1 + a_2 x + a_3 x^2 + \cdots + a_M x^{M-1}$$

We have

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2$$

Can formulate as a matrix equation and solve for parameters

See e.g. Numerical Recipes 15.4

Define $\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix}$ $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

Vector of model parameters

Vector of observations

$$\boldsymbol{X} = \begin{bmatrix} X_1(x_1) \cdots X_1(x_M) \\ \vdots \quad\quad \vdots \\ X_N(x_1) \cdots X_N(x_M) \end{bmatrix}$$

Matrix of model basis functions

Model:

Vector of model parameters

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{\varepsilon}$$

Vector of observations

Design matrix of model basis functions

Vector of errors

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

where we assume ε_i is drawn from some pdf with mean zero and variance σ^2

Weighting by errors

Define $\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_M \end{bmatrix}$

Vector of model parameters

$$\boldsymbol{b} = \begin{bmatrix} y_1/\sigma_1 \\ \vdots \\ y_N/\sigma_N \end{bmatrix}$$

Vector of weighted observations

$$\boldsymbol{A} = \begin{bmatrix} \frac{X_1(x_1)}{\sigma_1} & \cdots & \frac{X_1(x_M)}{\sigma_1} \\ \vdots & & \vdots \\ \frac{X_N(x_1)}{\sigma_N} & \cdots & \frac{X_N(x_M)}{\sigma_N} \end{bmatrix}$$

Design matrix

Weighted Model:

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{a} + \boldsymbol{e}$$

Vector of model parameters

Vector of weighted observations

Weighted design matrix of model basis functions

Vector of weighted errors

$$\boldsymbol{e} = \begin{bmatrix} \varepsilon_1 / \sigma_1 \\ \vdots \\ \varepsilon_N / \sigma_N \end{bmatrix}$$

where we assume ε_i is drawn from some pdf with mean zero and variance σ_i^2

We solve for the parameter vector $\hat{\boldsymbol{a}}_{LS}$ that minimises

$$\boldsymbol{S} = \boldsymbol{e}^{T} \cdot \boldsymbol{e} = \sum_{i=1}^{n} e_i^{\,2}$$

This has solution

$$\hat{\boldsymbol{a}}_{LS} = \left(\boldsymbol{A}^{T}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{T} \cdot \boldsymbol{b}$$

$M \times M$ *matrix*

and

$$\operatorname{cov}\left(\hat{\boldsymbol{a}}_{LS}\right) = \left(\boldsymbol{A}^{T}\boldsymbol{A}\right)^{-1}$$

Inverting $\left(A^T A\right)$ can be hazardous, particularly if A is a ***sparse*** matrix and/or close to singular.

Some inversion methods will break down, since they may give a formal solution, but are highly unstable to round-off error in the data.

Remedy: solution via Singular Value Decomposition.

From linear algebra theory:

Any $N \times M$ matrix can be decomposed as the product of an $N \times M$ column-orthogonal matrix $\mathbf{U}$, an $M \times M$ diagonal matrix $\mathbf{W}$ with positive or zero elements (the *singular* values) and the transpose of an $M \times M$ orthogonal matrix $\mathbf{V}$

From linear algebra theory:

Any $N \times M$ matrix can be decomposed as the product of an $N \times M$ column-orthogonal matrix $\mathbf{U}$, an $M \times M$ diagonal matrix $\mathbf{W}$ with positive or zero elements (the *singular* values) and the transpose of an $M \times M$ orthogonal matrix $\mathbf{V}$

M parameters

N observations

$$\begin{pmatrix} \\ \\ \mathbf{A} \\ \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ \mathbf{U} \\ \\ \\ \end{pmatrix} \cdot \begin{pmatrix} w_1 & & & \\ & w_2 & & \\ & & \cdots & \\ & & \cdots & \\ & & & w_M \end{pmatrix} \cdot \begin{pmatrix} \\ \mathbf{V}^T \\ \\ \end{pmatrix}$$

$$\sum_{i=1}^{N} U_{ik} U_{in} = \delta_{kn} \qquad \begin{array}{l} 1 \le k \le M \\ 1 \le n \le M \end{array}$$

$$\sum_{j=1}^{M} V_{jk} V_{jn} = \delta_{kn} \qquad \begin{array}{l} 1 \le k \le M \\ 1 \le n \le M \end{array}$$

Let the vectors $\mathbf{U}_{(i)}\ i = 1, \ldots, M$ denote the columns of $\mathbf{U}$ (each one is a vector of length N)

Let the vectors $\mathbf{V}_{(i)}; i = 1, \ldots, M$ denote the columns of $\mathbf{V}$ (each one is a vector of length M)

It can be shown that the solution to the general linear model satisfies

$$\hat{\boldsymbol{a}}_{LS} = \sum_{i=1}^{M} \left(\frac{\mathbf{U}_{(i)} \cdot \mathbf{b}}{w_i} \right) \mathbf{V}_{(i)}$$

$$\hat{a}_{LS} = \sum_{i=1}^{M} \left(\frac{\mathbf{U}_{(i)} \cdot \mathbf{b}}{w_i} \right) \mathbf{V}_{(i)}$$

Very small values of w_i will amplify any round-off errors in $\mathbf{b}$

Solution:

For these very small singular values, set $\frac{1}{w_i} = 0$.

This suppresses their noisy contribution to the least-squares solution for the parameters $\hat{a}_{LS}$.

SVD acts as a noise filter – see Section 5

Extensions and Generalisations

- Non-linear models? $y_i^{\text{model}} \equiv y^{\text{model}}(x_i; \theta_1, \ldots, \theta_k)$

Model parameters

Suppose $y_i^{\text{obs}} = y_i^{\text{model}} + \epsilon_i$

ε_i drawn from pdf with mean zero, variance σ_i^2

Then

$$S = \chi^2 = \sum_{i=1}^{n} \left[\frac{y_i^{\text{obs}} - y_i^{\text{model}}}{\sigma_i} \right]^2$$

Extensions and Generalisations

- Non-linear models? $$y_i^{\text{model}} \equiv y^{\text{model}}(x_i; \theta_1, \ldots, \theta_k)$$

Model parameters

Suppose $$y_i^{\text{obs}} = y_i^{\text{model}} + \epsilon_i$$

ε_i drawn from pdf with mean zero, variance σ_i^2

Then

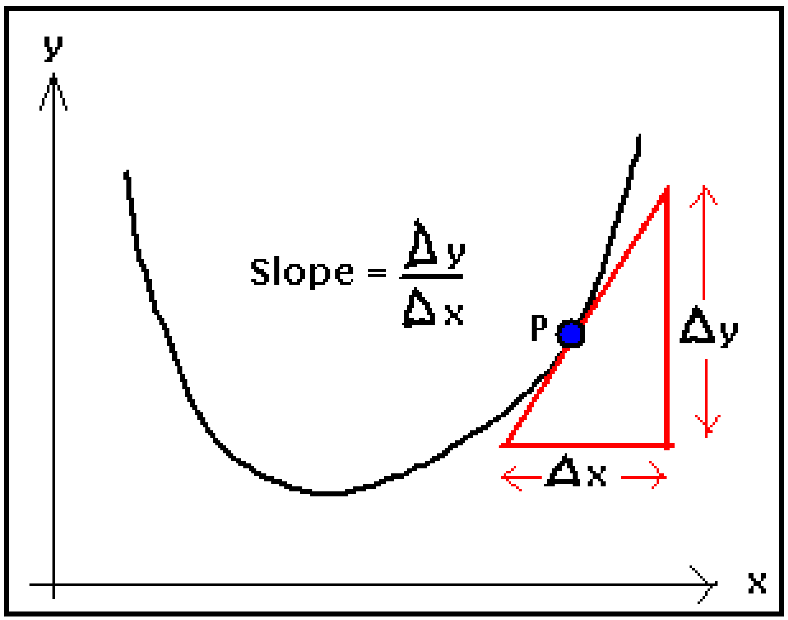
$$S = \chi^2 = \sum_{i=1}^{n} \left[\frac{y_i^{\text{obs}} - y_i^{\text{model}}}{\sigma_i} \right]^2$$

But no simple analytic method to minimise sum of squares
(e.g. no analytic solutions to $\partial S / \partial \theta_i = 0$)

Extensions and Generalisations

- Non-linear models?

Methods of solution often involve assuming *Taylor expansion* of χ^2 around minimum, and solving by gradient descent



See e.g.

Numerical Recipes 15.5

and Section 6

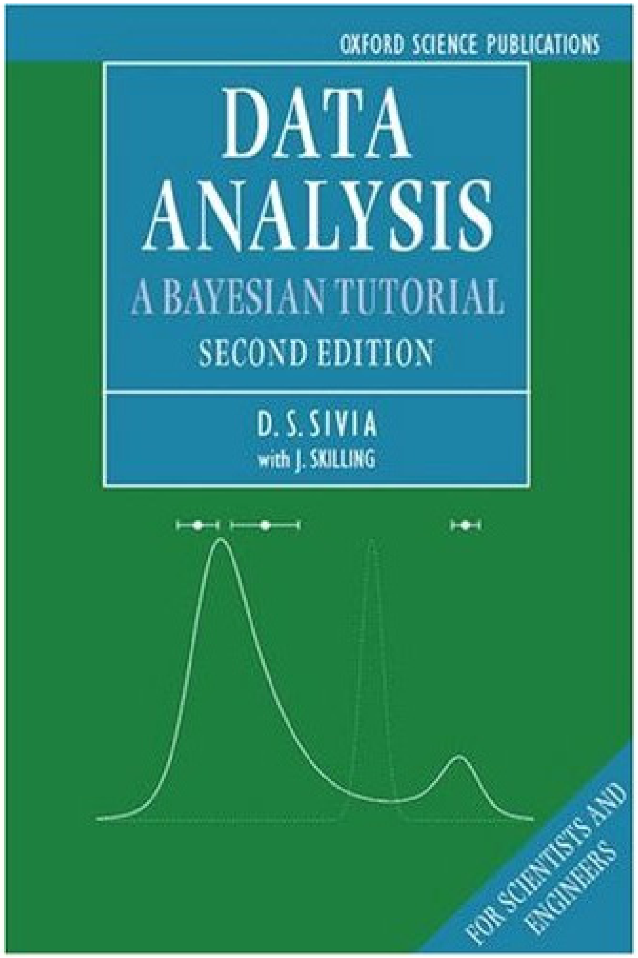
Extensions and Generalisations

- Correlated errors?

  We need to define a **covariance matrix** $C_{ij} = \mathrm{cov}(x_i, x_j)$

$$\chi^2 = \sum_i \sum_j \left(y_i - y_i^{\text{model}}\right)\left[C_{ij}\right]^{-1}\left(y_j - y_j^{\text{model}}\right)$$

See e.g. Gregory, Chapter 10



Sivia Chapter 3 gives a very clear discussion of least squares fitting within a Bayesian framework.

In particular, contrasts, for Gaussian residuals:

- known σ
- unknown σ → Student's t

See also Section 3

The principle of maximum likelihood

Frequentist approach:

A parameter is a fixed (but unknown) constant

From actual data we can compute Likelihood,

$L =$ probability of obtaining the observed data, given the value of the parameter θ

The principle of maximum likelihood

Frequentist approach:

A parameter is a fixed (but unknown) constant

From actual data we can compute Likelihood,

$L =$ probability of obtaining the observed data, given the value of the parameter θ

Now define **likelihood function**: (infinite) family of curves generated by regarding L as a function of θ, for data fixed.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$

The principle of maximum likelihood

Frequentist approach:

A parameter is a fixed (but unknown) constant

From actual data we can compute Likelihood,

L = probability of obtaining the observed data, given the value of the parameter θ

Now define **likelihood function**: (infinite) family of curves generated by regarding L as a function of θ, for data fixed.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$

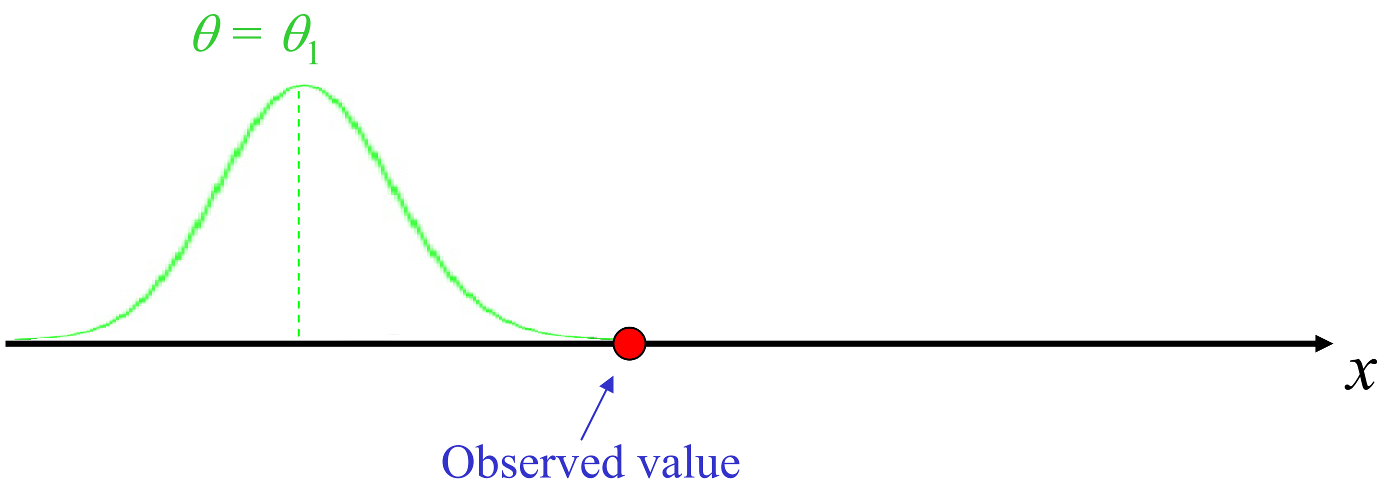
We set the parameter equal to the value that makes the actual data sample we *did* observe – out of all the possible random samples we *could have* observed – the most likely.

Aside: Likelihood function has same definition in Bayesian probability theory, but subtle difference in meaning and interpretation – no need to invoke idea of (infinite) ensemble of different samples.

Principle of Maximum Likelihood

A good estimator of θ maximises L -

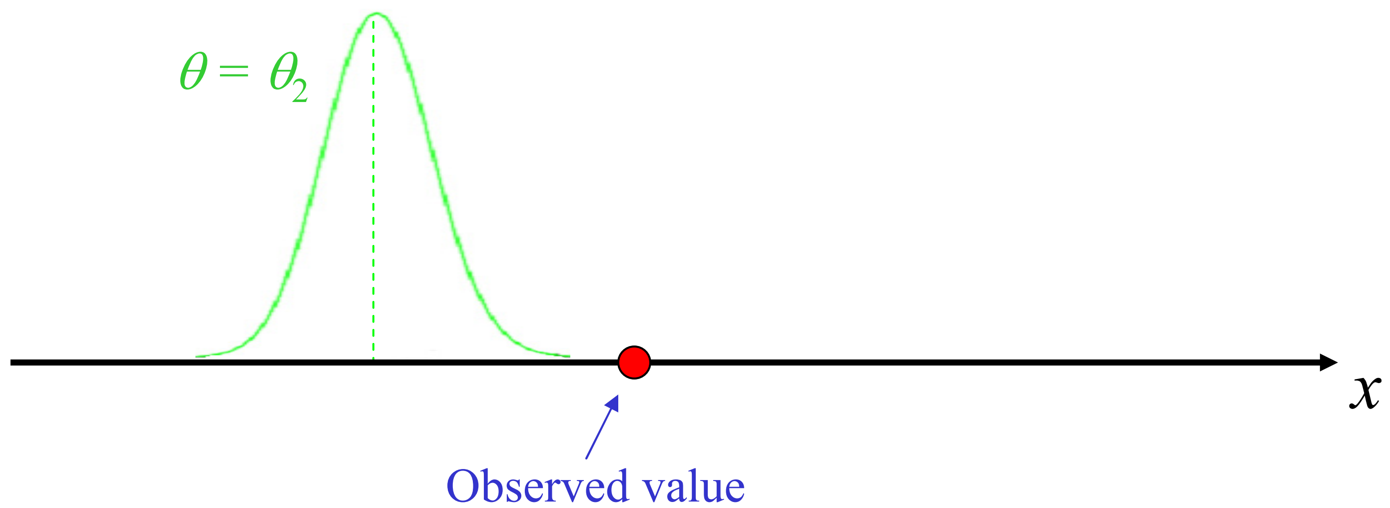
i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$



Principle of Maximum Likelihood

A good estimator of θ maximises L -

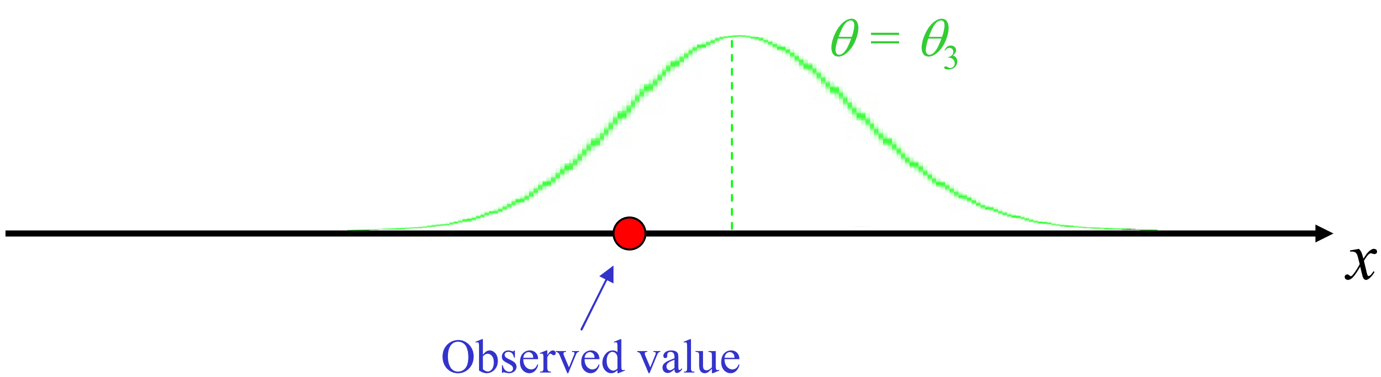
i.e. $\dfrac{\partial L}{\partial \theta} = 0$ and $\dfrac{\partial^2 L}{\partial \theta^2} < 0$



Principle of Maximum Likelihood

A good estimator of θ maximises L -

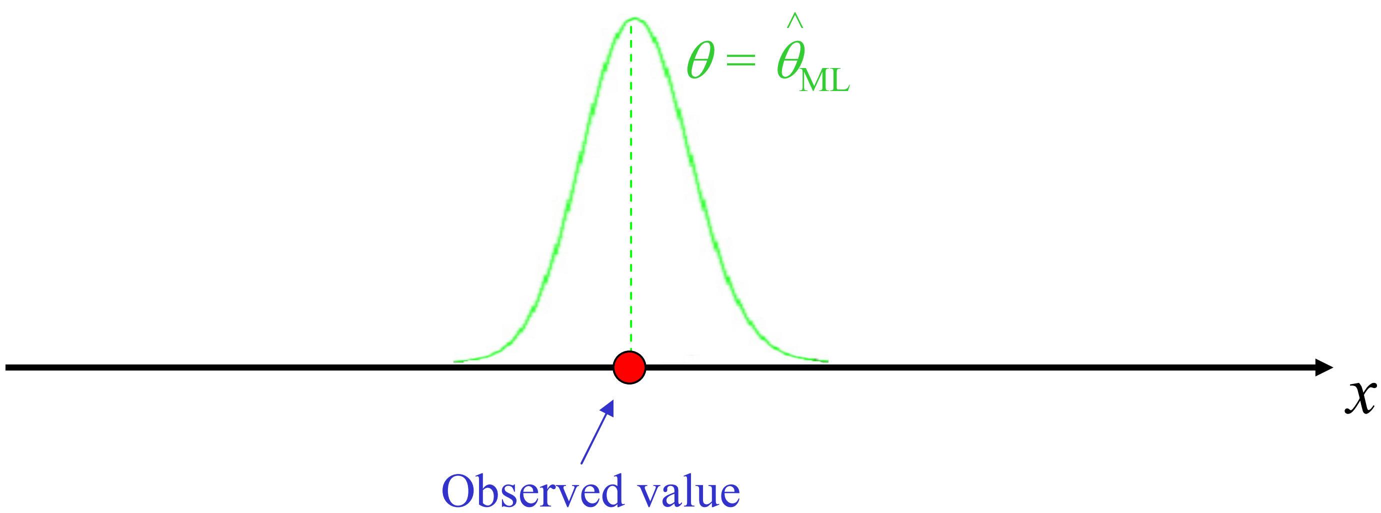
i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$



Principle of Maximum Likelihood

A good estimator of θ maximises L -

i.e. $\frac{\partial L}{\partial \theta} = 0$ and $\frac{\partial^2 L}{\partial \theta^2} < 0$



Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again **weighted least squares** for the simple model $y_i = a + bx_i + \epsilon_i$

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again weighted least squares for the simple model $y_i \quad = \quad a + bx_i + \epsilon_i$

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Likelihood $$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{\varepsilon_i^2}{\sigma_i^2} \right]$$

Question 5: How can we justify writing the likelihood as a product?

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2}\right]$$

A Because the residuals are all equal to each other

B Because the residuals are all Gaussian

C Because the residuals are all positive

D Because the residuals are all independent

Question 5: How can we justify writing the likelihood as a product?

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left[-\frac{1}{2}\frac{\varepsilon_i^2}{\sigma_i^2}\right]$$

A Because the residuals are all equal to each other

B Because the residuals are all Gaussian

C Because the residuals are all positive

D Because the residuals are all independent

Least squares as maximum likelihood estimators

To see the maximum likelihood method in action, let's consider again weighted least squares for the simple model $y_i = a + bx_i + \epsilon_i$

Suppose the i^{th} residual, $\{\epsilon_i\}$, is assumed to be drawn from some underlying pdf with mean zero and variance σ_i^2, where the variance is allowed to be different for each residual.

Let's assume the pdf is a Gaussian

Likelihood $$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{\varepsilon_i^2}{\sigma_i^2} \right]$$

(note: L is a product of 1-D Gaussians because we are assuming the ε_i are independent)

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \qquad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right]$$

and the ML estimators of a and b satisfy $\partial L / \partial a = 0$ and $\partial L / \partial b = 0$

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \quad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right]$$

and the ML estimators of a and b satisfy $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$

But maximising L is equivalent to maximising $\ell = \ln L$

Here
$$\ell = -\frac{n}{2}\ln(2\pi) - \ln\sum_{i=1}^{n}\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i}{\sigma_i}\right)^2$$

$$= \text{constant} - \frac{1}{2}S$$

This is exactly the same sum of squares we defined earlier

Substitute $\varepsilon_i = y_i - a - bx_i$

$$\Rightarrow \qquad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2} \frac{(y_i - a - bx_i)^2}{\sigma_i^2} \right]$$

and the ML estimators of a and b satisfy $\partial L / \partial a = 0$ and $\partial L / \partial b = 0$

But maximising L is equivalent to maximising $\ell = \ln L$

Here $$\ell = -\frac{n}{2}\ln(2\pi) - \ln \sum_{i=1}^{n} \sigma_i - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

$$= \text{constant} - \frac{1}{2} S$$

This is exactly the same sum of squares we defined earlier

So in this case maximising L is **exactly equivalent** to minimising the sum of squares.

i.e. for Gaussian, independent errors, ML and weighted LS estimators are identical.