UKRI Science and Technology Facilities Council

Hartree Centre

# Welcome

Science and Technology Facilities Council
Hartree Centre

# Big Data Week 7

## *Data Analysis & Probabilistic Modelling*

Dr Simon Goodchild
Data Science group leader, Hartree Centre

# Lecture overview

- Refresher of some basic statistical concepts
- Summary statistics
- Sampling and estimation
- *Break*
- Probabilistic models
  - Linear regression
  - Logistic regression
  - Time series models
  - Overfitting and regularisation
- Exploratory data analysis
- Model training

Science and Technology Facilities Council
Hartree Centre

# Statistical concepts - refresher

Science and Technology Facilities Council
Hartree Centre

# Probability space

3 elements in a probability space describing an experiment

(e.g. a single throw of a die)

Sample space $\Omega$
(set of all possible outcomes)

⚀

⚁

⚂

⚃

⚄

⚅

Event space $\mathcal{F}$
(events are sets of outcomes)

{⚀} {⚁} {⚂} …

…

{⚀ ⚁} {⚀ ⚂} …

…

{⚁ ⚂ ⚃ ⚄} …

…

{⚀ ⚁ ⚂ ⚃ ⚄ ⚅}

Probability function $P$
(assign each event a probability)

{⚀} $\rightarrow \frac{1}{6}$

…

{⚀ ⚁} $\rightarrow \frac{1}{3}$

…

{⚁ ⚂ ⚃ ⚄} $\rightarrow \frac{2}{3}$

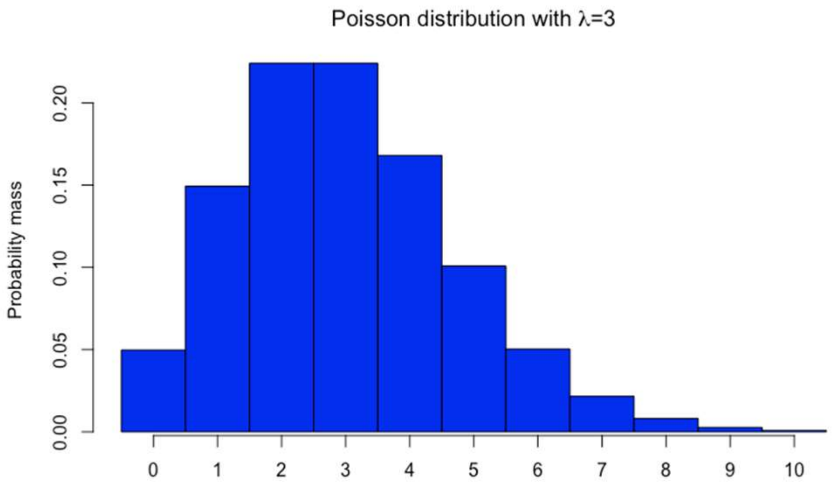
…

{⚀ ⚁ ⚂ ⚃ ⚄ ⚅} $\rightarrow 1$

GS(0

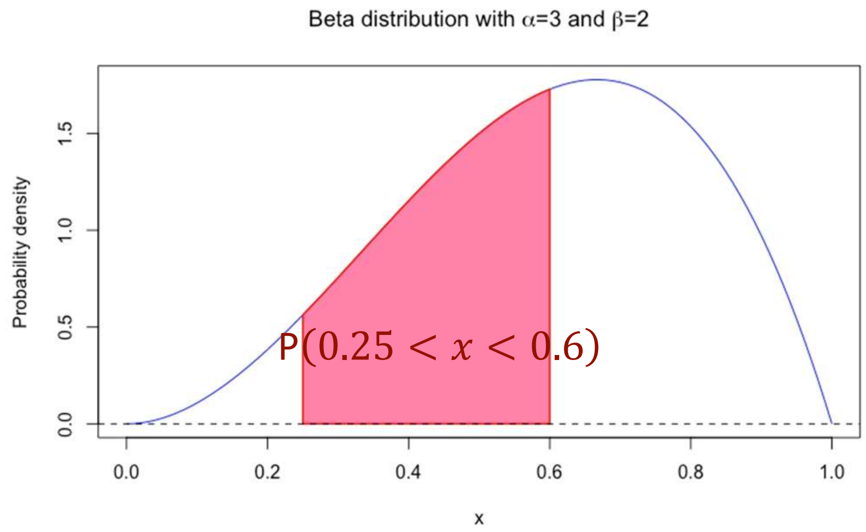
# Probability distributions

Discrete distribution




Individual probability values
All sum to 1

Continuous distribution




Probability of a range given by the area between the ends
Total area under the distribution is 1

Science and Technology Facilities Council
Hartree Centre

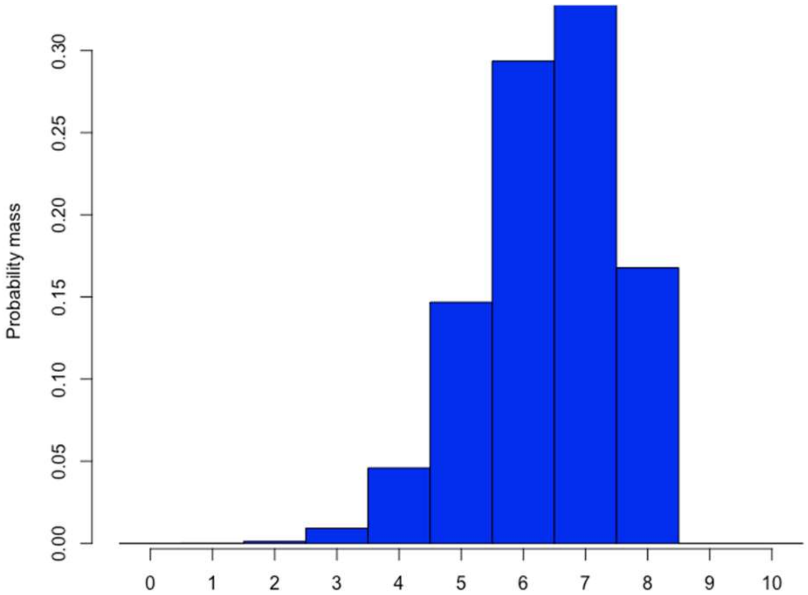
**Slide 6**

**GS(0** Perhaps introduce some of the common discrete and continuous distributions.

Goodchild, Simon (STFC,DL,HC), 2022-09-21T11:03:59.928

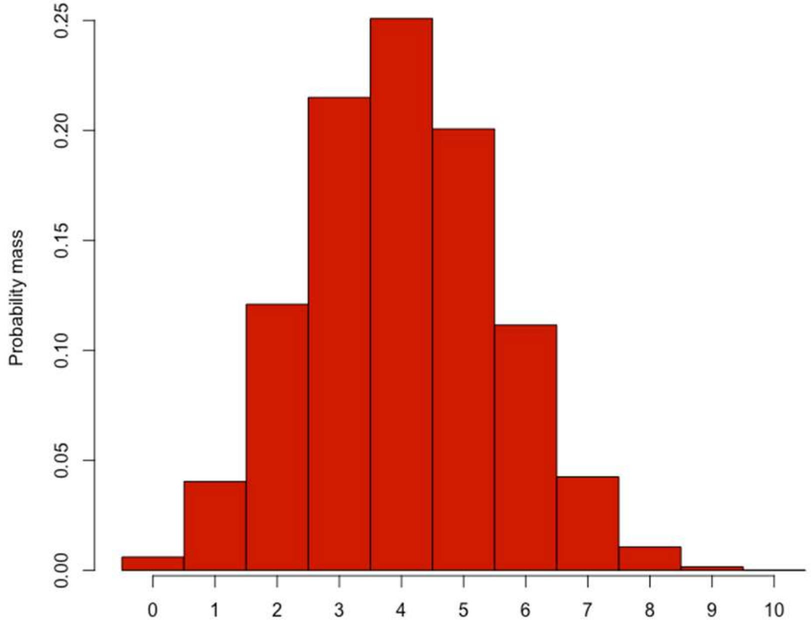
# Probability distributions

A particular probability distribution is described by a set of numerical *parameters* .




Binomial distribution with *n=8* and *p=0.8*

$$x \sim \mathrm{B}(8, 0.8)$$




Binomial distribution with *n=10* and *p=0.4*

$$x \sim \mathrm{B}(10, 0.4)$$

Science and Technology Facilities Council
Hartree Centre

Parameters often (but not always) can be interpreted:
e.g. the binomial distribution with parameters *n* and *p* gives the probability of *x* successes in *n* trials with probability *p.*

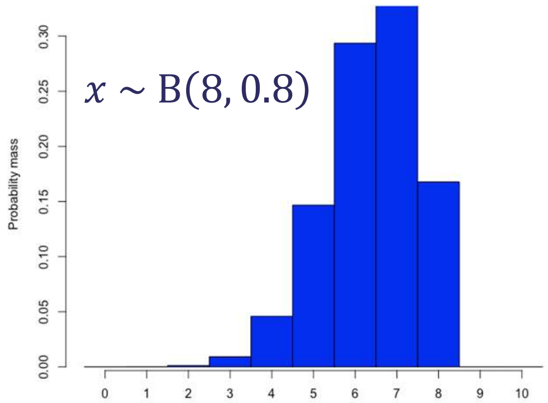
# Probability distributions

Some common discrete probability distributions:

**Bernoulli distribution**

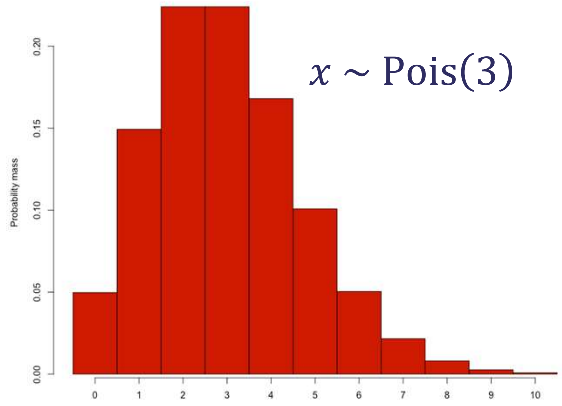
Parameter *p. G*ives the probability of a success in a trial with probability *p.*

**Binomial distribution**



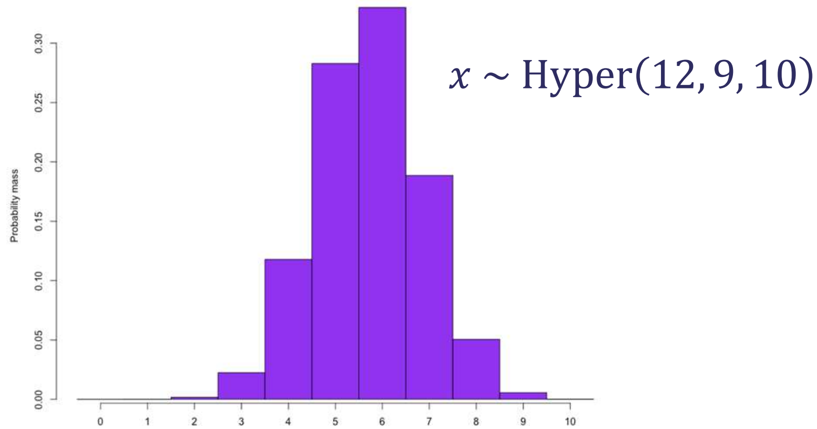

Parameters *n* and *p. G*ives the probability of *x* successes in *n* trials with probability *p.*

**Poisson distribution**




Parameter $\lambda$*. G*ives the probability of *x* events occurring in intervals with mean rate $\lambda$.

**Hypergeometric distribution**

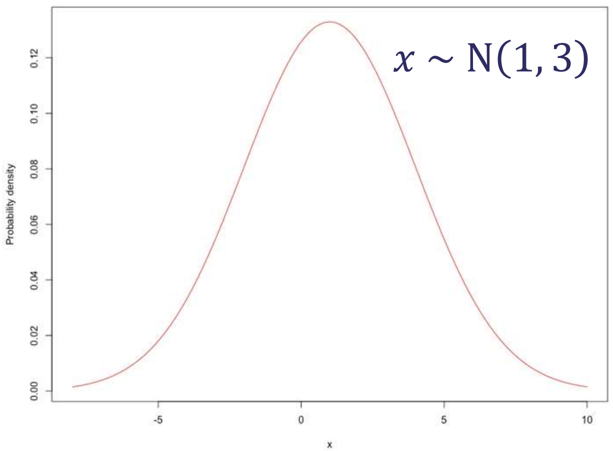



Parameters *m, n, k. G*ives the probability of getting *x* 1s when drawing *k* elements without replacement from *n* 1s and *m* 0s.

# Probability distributions

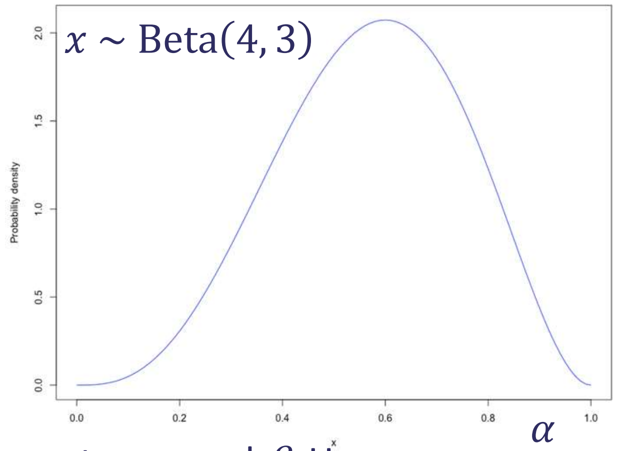
Some common continuous probability distributions:

**Normal/Gaussian distribution**



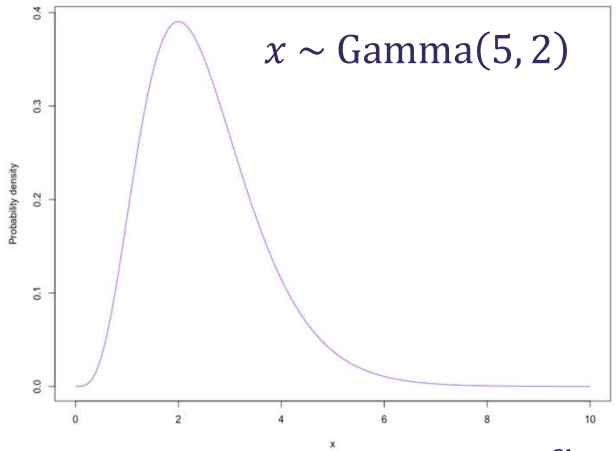

Parameters $\mu$ and $\sigma$. Has mean $\mu$ and variance $\sigma^2$

**Beta distribution**




Parameters $\alpha$ and $\beta$. Has mean $\dfrac{\alpha}{\alpha + \beta}$

**Gamma distribution**

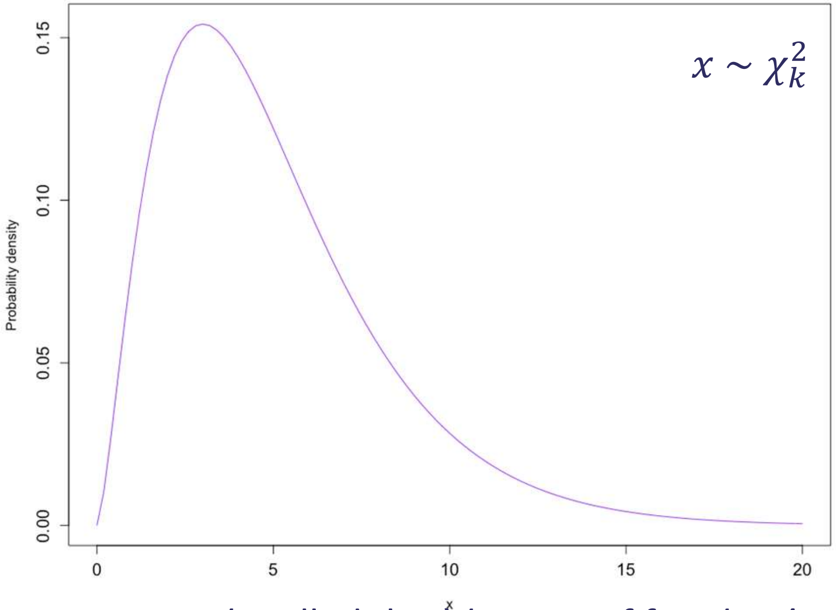



Parameters $\alpha$ and $\beta$. Has mean $\dfrac{\alpha}{\beta}$

# Probability distributions

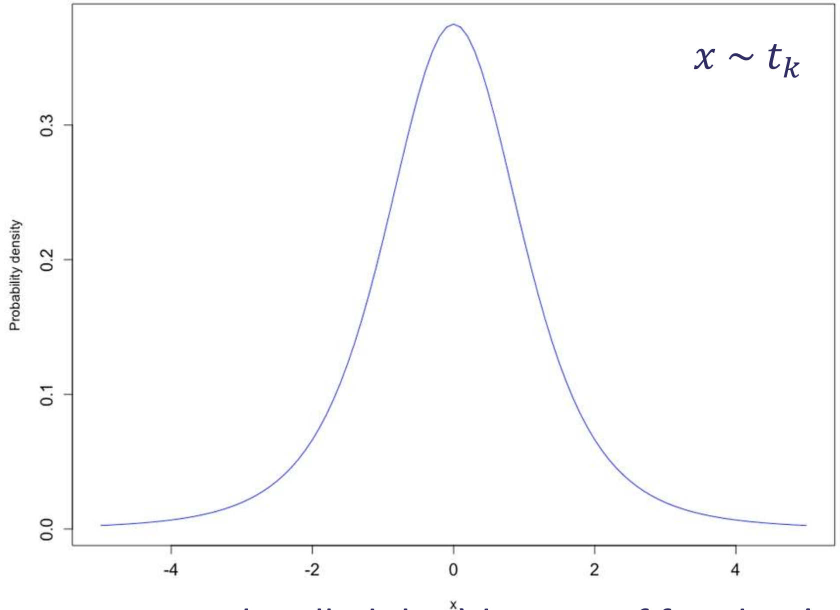
Some common continuous probability distributions:

**Chi-squared distribution**




Parameter *k*, called the *'degrees of freedom'*

**(Student's) *t*-distribution**

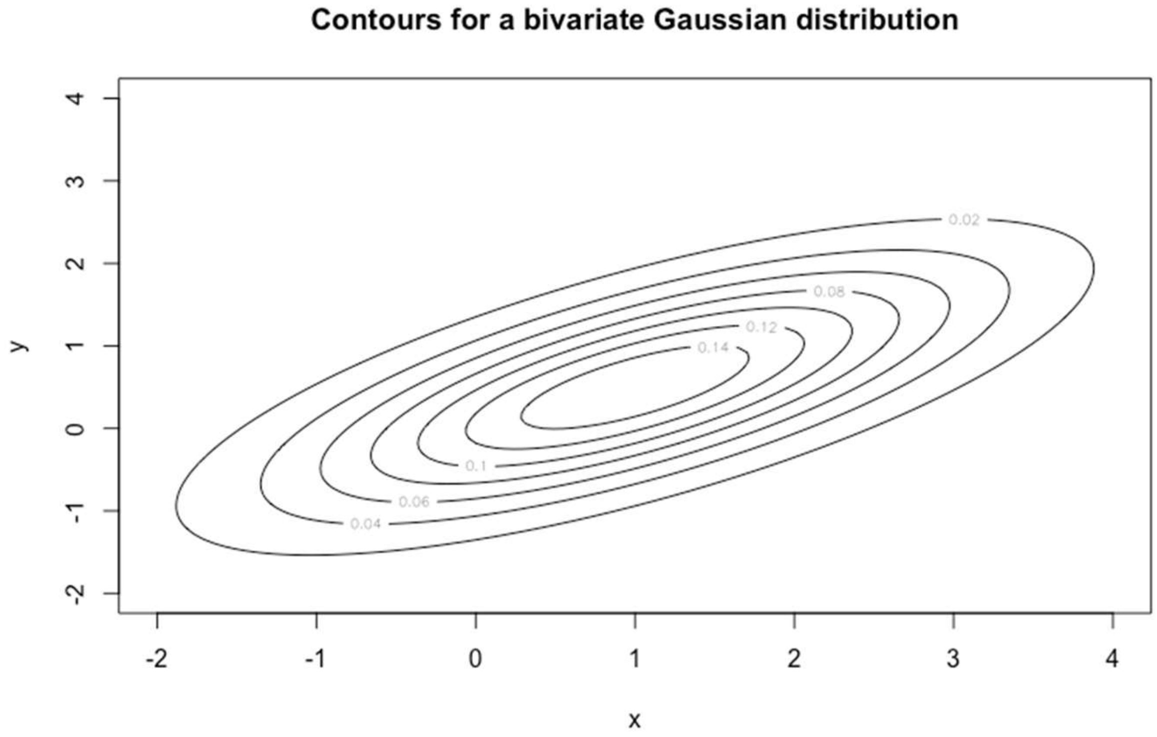



Parameter *k*, called the *'degrees of freedom'*

Science and Technology Facilities Council

Hartree Centre

# Probability distributions




This is a bivariate distribution – the total volume under the shape is 1.

The parameters are now matrices:

$$\mu = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$$

Science and Technology Facilities Council

Hartree Centre

# Expectation

The expected value of a quantity *x* is given by

$$\mathrm{E}[x] = \sum x \ \mathrm{P}(x)$$

e.g. if you roll 2 6-sided dice and add the scores together, the expectation of the total score is

$$\frac{1}{36} \times 2 + \frac{1}{18} \times 3 + \frac{1}{12} \times 4 + \frac{1}{9} \times 5 + \frac{5}{36} \times 6 + \frac{1}{6} \times 7 + \frac{5}{36} \times 8 + \frac{1}{9} \times 9 + \frac{1}{12} \times 10 + \frac{1}{18} \times 11 + \frac{1}{36} \times 12 = 7$$

This can be applied to any relevant quantity

$$\mathrm{E}[x^2] = \sum x^2 \ \mathrm{P}(x)$$

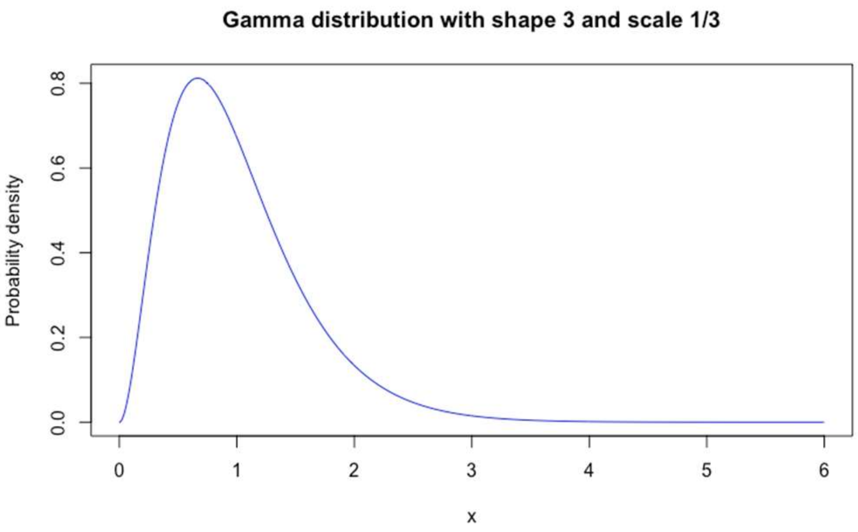
For 2 6-sided dice

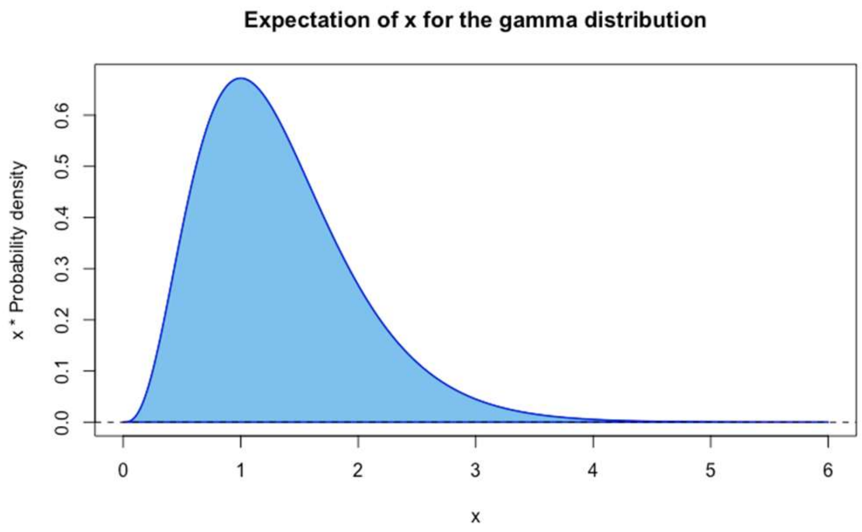
$$\frac{1}{36} \times 4 + \frac{1}{18} \times 9 + \frac{1}{12} \times 16 + \frac{1}{9} \times 25 + \frac{5}{36} \times 36 + \frac{1}{6} \times 49 + \frac{5}{36} \times 64 + \frac{1}{9} \times 81 + \frac{1}{12} \times 100 + \frac{1}{18} \times 121 + \frac{1}{36} \times 144 = \frac{329}{6}$$

Science and
Technology
Facilities Council
Hartree Centre

# Expectation

If the distribution is continuous, the expectation is given by the total area under a curve



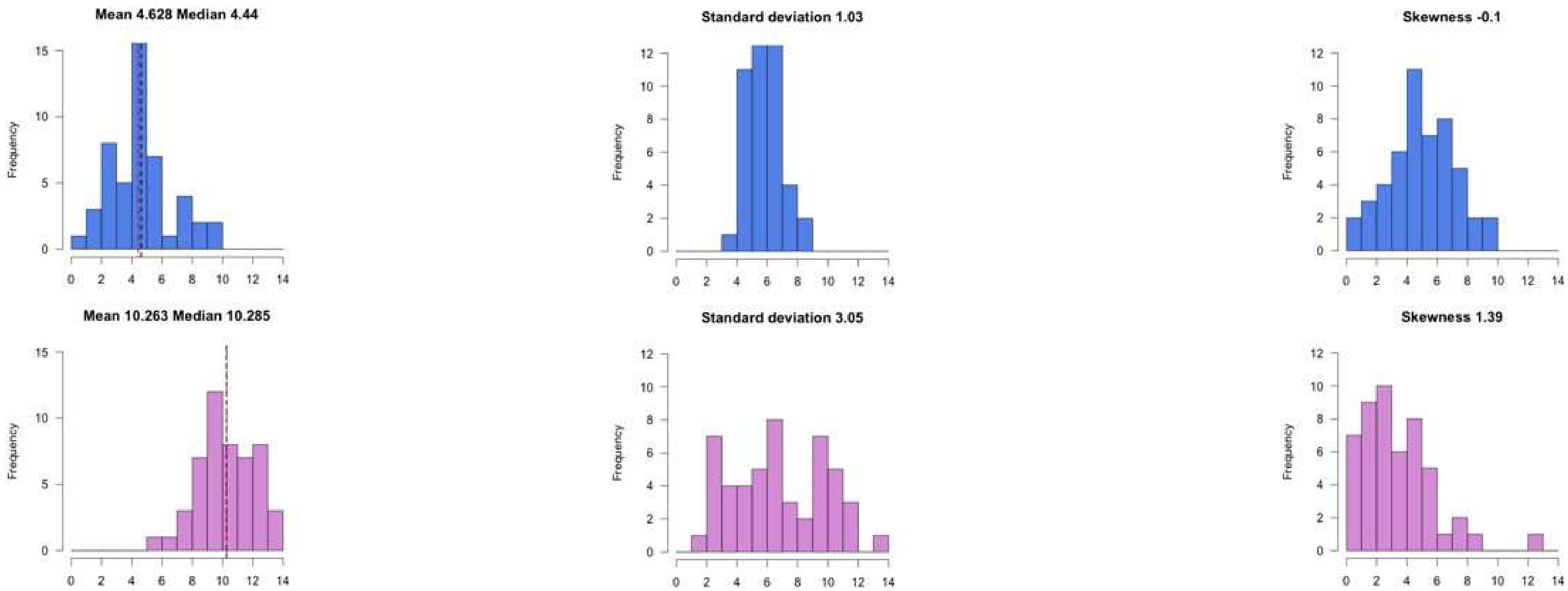




Science and Technology Facilities Council

Hartree Centre

# Summary statistics

Expectation is often used to define *summary statistics* – numbers which describe a particular property of a data set




Mean: $\mu = \mathrm{E}[x]$
describes the location

Standard deviation: $\sigma = \sqrt{\mathrm{E}[(x - \mu)^2]}$
describes the spread

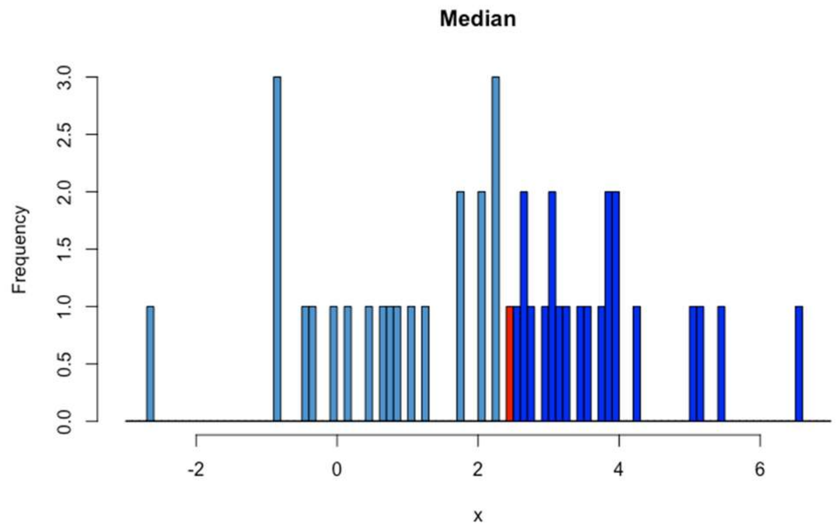
Skewness: $s = \mathrm{E}\left[\left(\frac{x-\mu}{\sigma}\right)^3\right]$
describes the asymmetry

Science and
Technology
Facilities Council

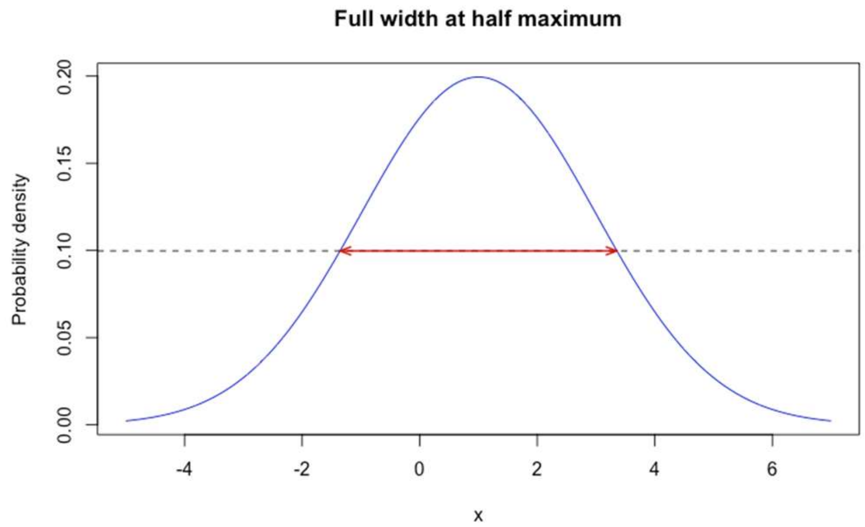
Hartree Centre

# Summary statistics

Other summary statistics can also be defined




Location: median




Spread: Full Width Half Maximum (FWHM)

Hartree Centre

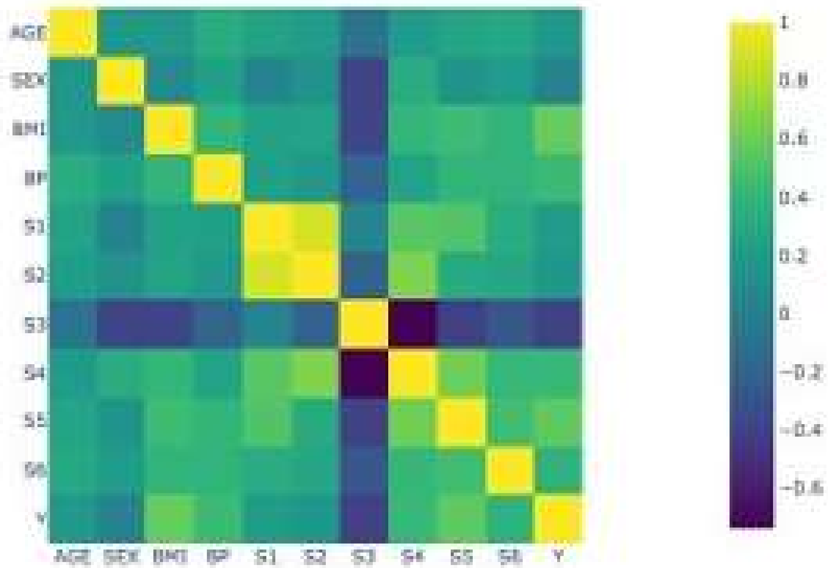
# Summary statistics

When there is more than one independent variable, other summary statistics are used

Covariance: $\Sigma(x, y) = \mathrm{E}\left[(x - \mu_x)(y - \mu_y)\right]$

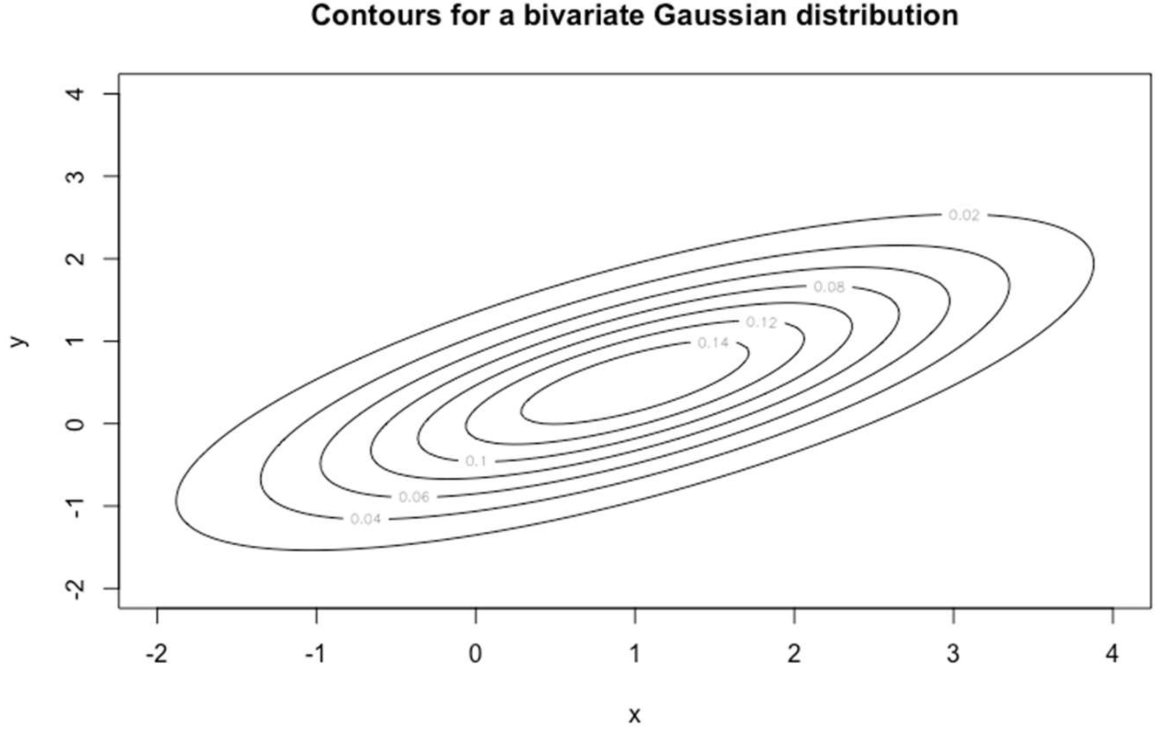
Correlation: $r(x, y) = \dfrac{\Sigma(x, y)}{\sigma_x \sigma_y}$

These can be used to measure the connection between variables.




Science and Technology Facilities Council
Hartree Centre

# Summary statistics




We can interpret the matrix parameters:

$$\mu = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \quad \text{is the mean}$$

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{is the covariance}$$

$$r = \begin{bmatrix} 1 & 0.33 \\ 0.33 & 1 \end{bmatrix} \quad \text{is the correlation}$$

Science and Technology Facilities Council

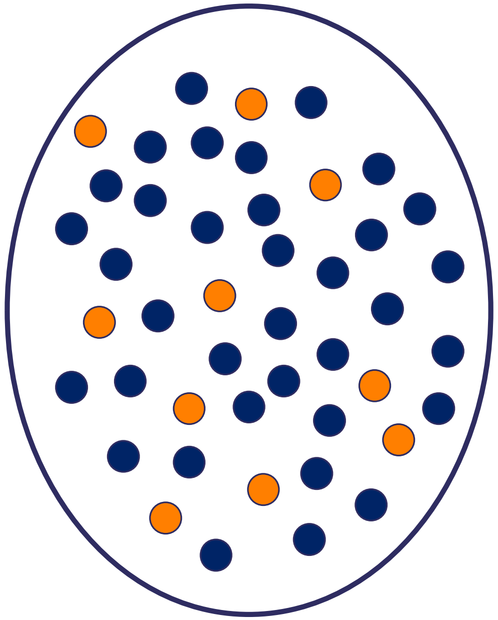
Hartree Centre

Science and Technology Facilities Council
Hartree Centre

# Sampling and estimation

Science and Technology Facilities Council
Hartree Centre

# Sampling and estimators



If a population or a data set is too large to examine globally, take a sample.

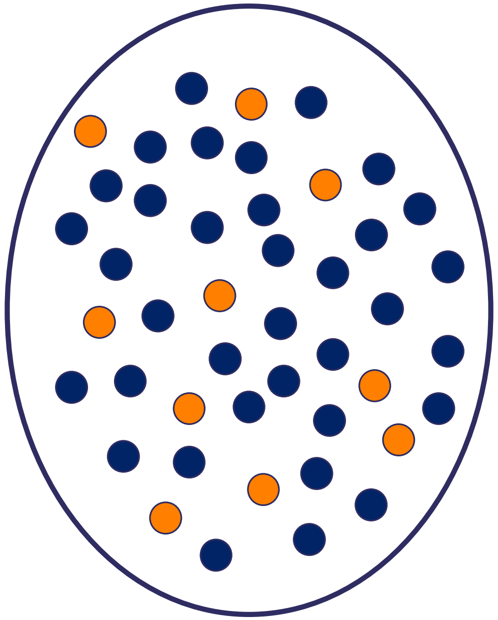
A sample is used to calculate *estimators* of population quantities that are of interest. For example the sample mean can be used as an estimator of the population mean.

Circumflex notation is often used for estimators: $\hat{\theta}$ is an estimator for parameter $\theta$.

Sampling ideas are also useful when looking at errors in statistical models, such as the *bootstrap* method.

Science and Technology Facilities Council
Hartree Centre

# Sampling



Sampling is based on the idea of a *simple random sample*.

This is a sample where every group of the sample size is equally probable.

Samples can be taken with or without replacement of the sampled point in the population.

# Properties of estimators

If $\hat{\theta}$ is an estimator of a parameter with true value $\theta$

Mean-squared error: $e^2 = \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$

Variance: $s^2 = \mathrm{E}\left[\left(\hat{\theta} - \mathrm{E}[\hat{\theta}]\right)^2\right]$ — a low variance estimator produces values close to its expectation

Bias: $b = \mathrm{E}[\hat{\theta}] - \theta$ — an *unbiased* estimator has an expectation equal to the true value

Robustness: [*various measures*] — a robust estimator is resistant to errors in the data

**Bias-variance tradeoff:**

The mean-squared error of an estimator can be decomposed
into a bias term and a variance term (and sometimes a noise term)
$e^2 = s^2 + b^2 + \sigma^2$

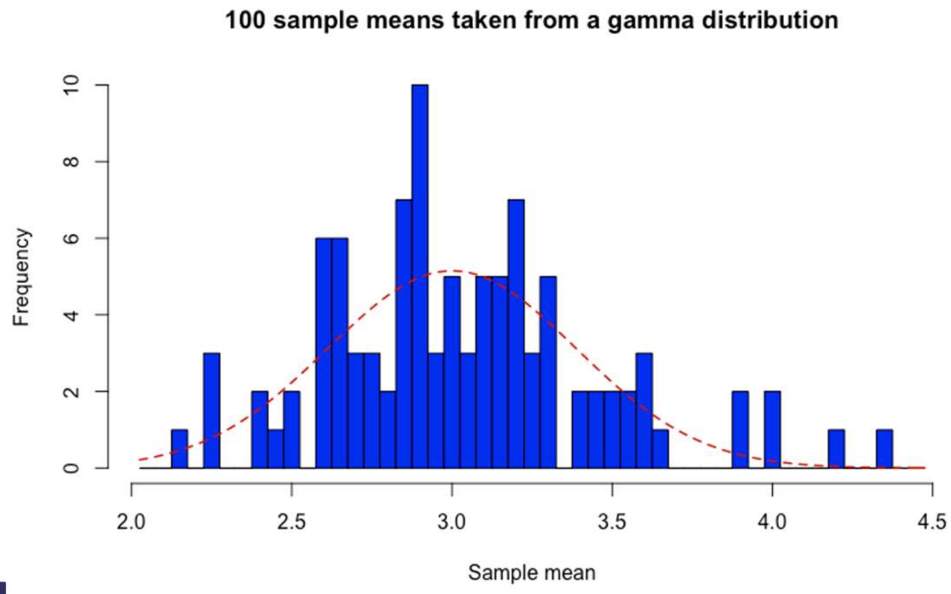
Science and Technology Facilities Council

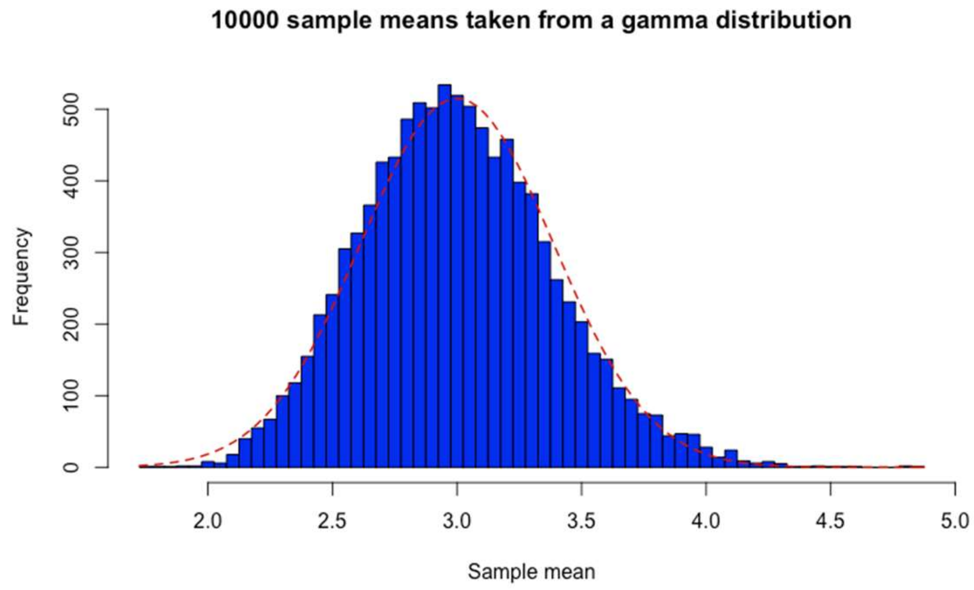
Hartree Centre

# Estimator variance

If a quantity has population mean $\mu$ and population variance $\sigma^2$
and we take a simple random sample (with replacement) of size $n$:

the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ has variance $\frac{\sigma^2}{n}$




Technology Facilities Council
Hartree Centre

# Estimator bias

If a quantity has population mean $\mu$ and population variance $\sigma^2$:

the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an unbiased estimator of the population mean

the raw sample variance $\varsigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is a biased estimator of the population variance

as $\mathrm{E}[\varsigma^2] = \frac{n-1}{n}\sigma^2$ (the difference is due to the variance of $\bar{x}$).

the usual sample variance is therefore $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ which is unbiased.

# Estimator robustness

The mean of the data set {2, 3, 4, 4, 6, 7, 7, 9, 10, 10000} is 1005.2 which is not representative of the data.

- The mean is not *robust* because it is strongly affected by very large outliers.
  This can be quantified by looking at the change in the mean when a new value is added.

The median of the data set is 6.5 which is more representative.

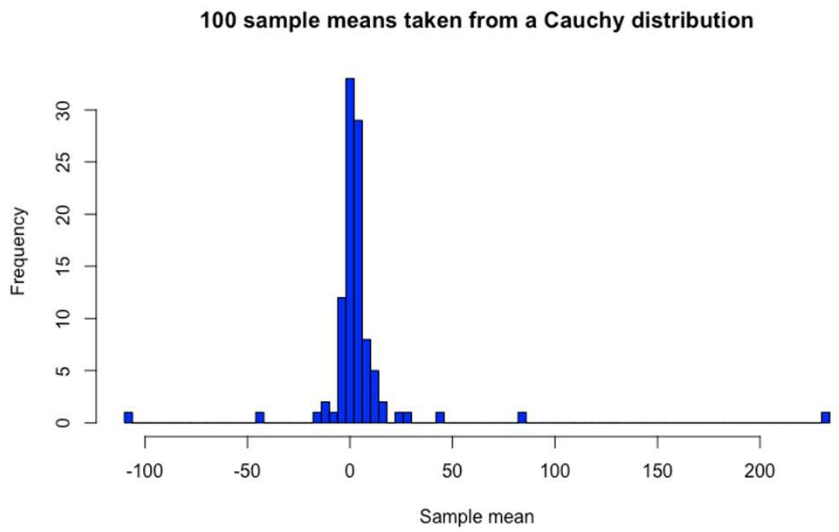
- The median is a more robust estimator of location.

The variance of a data set is also not robust to large outliers, so other measures of the spread such as the full width at half maximum (FWHM) can be used.

# Estimator robustness

Some probability distributions don't even have a defined mean and variance.

Calculate an estimator from data based on one of these distributions and it will behave oddly.

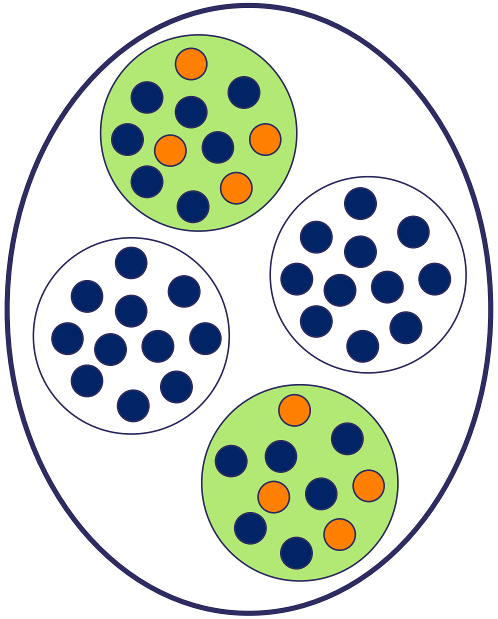



In these cases you have to use a different estimator like the median or FWHM.

Science and Technology Facilities Council

Hartree Centre

# Sampling strategies: cluster sampling



Population is divided into clusters which may be logistically separated

e.g. different farming villages in a developing country

different days to carry out sampling

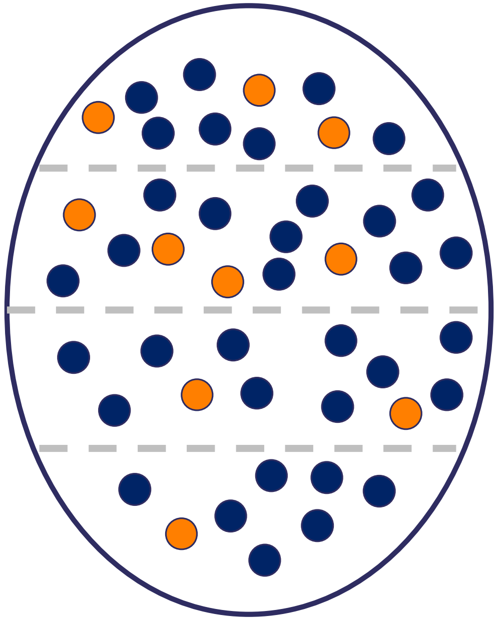
First choose a random sample of clusters
Then either sample the whole cluster (one-stage)
or a random sample from that cluster (two-stage)

Estimator properties depend on the variance between as well as within clusters

Science and Technology Facilities Council

Hartree Centre

# Sampling strategies: stratified sampling



Population is described by a feature which affects the property of interest

e.g. age affecting voting behaviour

By sampling in different proportions from different strata, the variance can be reduced.

Uses weighted averages to correct for the sampling proportions

$$\hat{p} = \sum_{i=1}^{k} \frac{n_i}{n} p_i$$

$p_i, n_i$ are the sample proportions and sizes from each stratum

$n = \sum_{i=1}^{k} n_i$ is the total population size

# Sampling strategies: a diverse population

In 1936, Democrat Franklin Roosevelt and Republican Alf Landon are standing for election as U.S. President.

Two groups took opinion polls.

- George Gallup had 50,000 responses.
- The Literary Digest magazine had 2.4 million!

**Whose poll was more accurate?**

**Their predictions:**

Gallup poll: Roosevelt win, 55.7% to 44.3%.

Literary Digest poll: Landon win, 57% to 43%.

# Sampling strategies: a diverse population

**The result:**

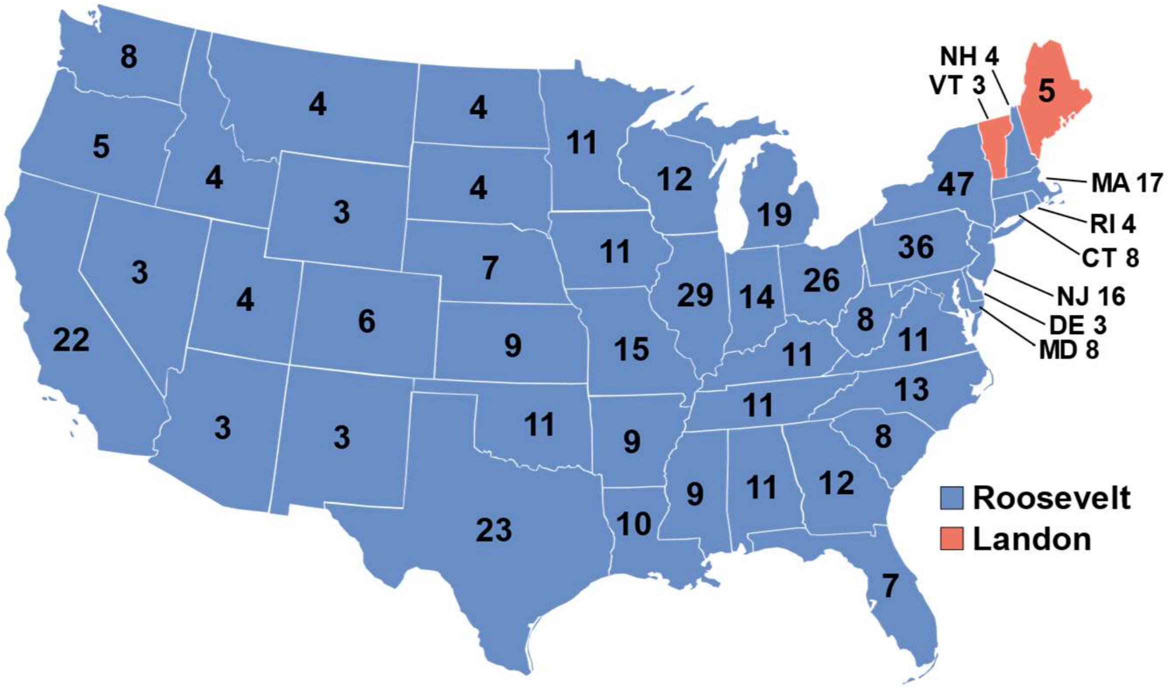



*Image credit: Wikipedia*

**Roosevelt landslide! 60.8% to 36.5% and 46 out of 48 states.**

‘A random selection of three people would have been better than a group of 300 chosen by Mr. Kinsey.’

(John Tukey, quoted in New York Times obituary, 28 July 2000)

UKRI Science and Technology Facilities Council

Hartree Centre

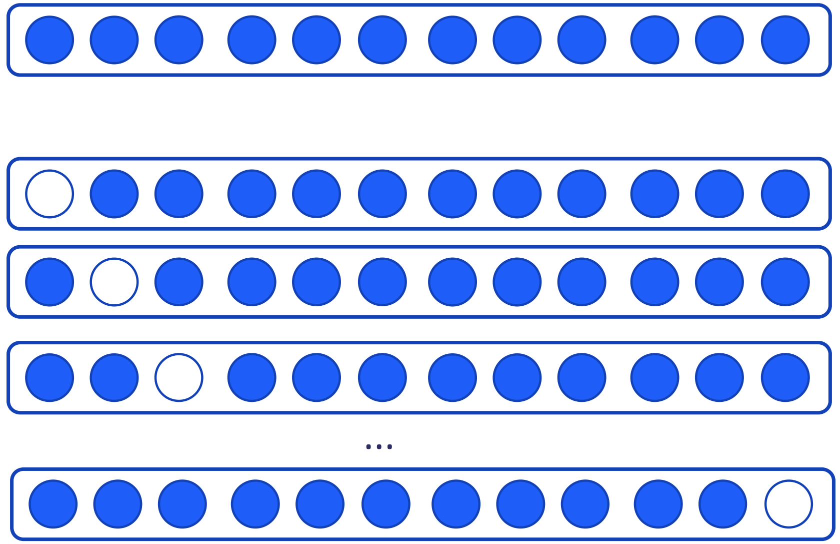
# Resampling in statistical models

*Resampling* – re-calculating statistics using a new sample drawn from the data – can be used to estimate error and bias.

e.g the *jackknife estimator*

Start with a population of *n* objects and an estimator $\hat{\theta}$

Re-calculate the estimator using all but one object – call this $\hat{\theta}_{(i)}$



$\hat{\theta}_{(1)}$

$\hat{\theta}_{(2)}$

$\hat{\theta}_{(3)}$

...

$\hat{\theta}_{(12)}$

Take the mean to get a new estimate $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}$

Science and Technology Facilities Council

Hartree Centre

# Resampling in statistical models

This new estimate can be used to estimate the bias of the estimator $\hat{\theta}$:

$$\hat{b} = (n-1)\left(\hat{\theta}_{(.)} - \hat{\theta}\right)$$

and produce a new, less biased estimator

$$\widehat{\theta'} = n\hat{\theta} - (n-1)\left(\hat{\theta}_{(.)}\right)$$

More sophisticated versions of re-sampling are part of many machine learning models
e.g. bootstrapping in a random forest.

Science and
Technology
Facilities Council
Hartree Centre

Science and Technology Facilities Council
Hartree Centre

# Break

Science and Technology Facilities Council
Hartree Centre

Science and Technology Facilities Council

Hartree Centre

# Probabilistic modelling

Science and Technology Facilities Council

Hartree Centre

# Probabilistic models

'Essentially, all models are wrong, but some are useful.'

George E.P. Box, *Empirical Model-Building and Response Surfaces*

A *probabilistic model* (or statistical model) is a set of probability distributions on a sample space.

(McCullagh, 2002)

Models can be *parametric* or *non-parametric:*

a parametric model is described by a set of parameters $\Theta$ and associated probability distributions
these parameter distributions then map to the distributions on the sample space

Science and
Technology
Facilities Council
Hartree Centre

# Probabilistic models

Probabilistic modelling often proceeds by optimising an *objective function.*

This measures the error that we would like to minimise.

# Linear regression

Take a set of $n$ observations with $p$ independent variables $\{y_i, x_{ij}\}$

A probabilistic model is a set of conditional probability distributions that give $\mathrm{P}(y|x_j)$

The linear regression model is parametric with parameters $\{\beta_j\}$

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \beta_0 + \varepsilon_i$$

or $\mathrm{y} = \mathrm{X\beta} + \mathrm{\varepsilon}$ in matrix notation

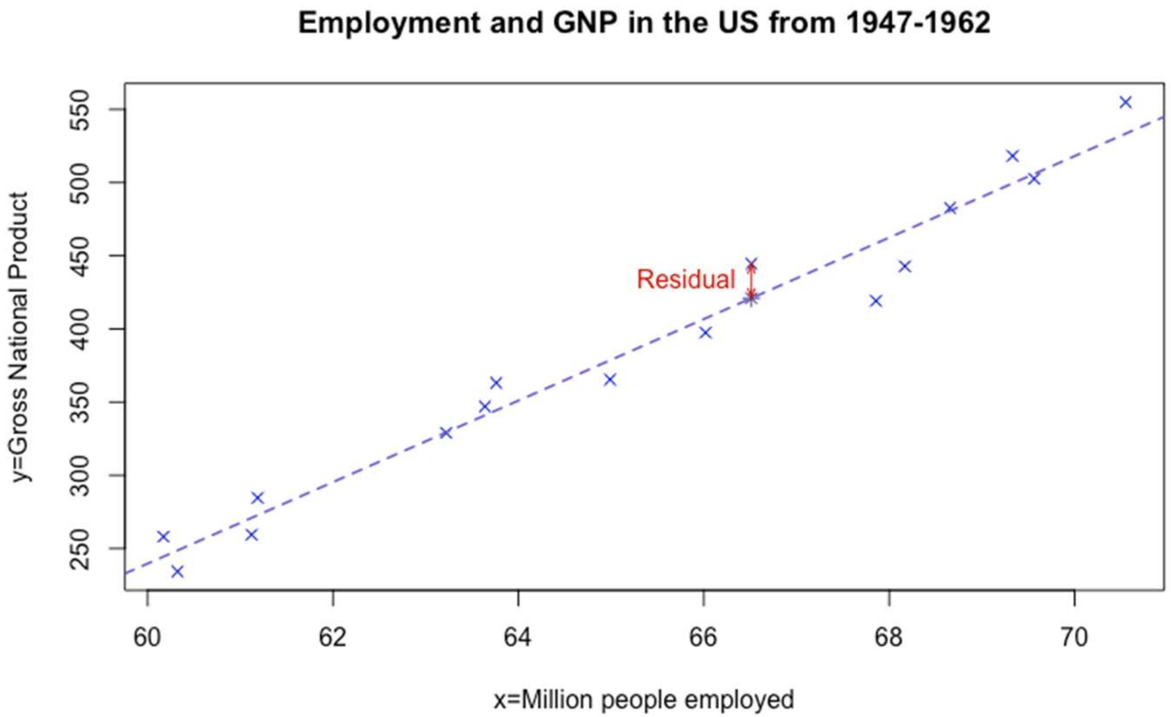
$\varepsilon_i$ is an error term, and to make this a probabilistic model we make assumptions about the errors.

Typically this is that they are independent, uncorrelated, have mean 0 and equal variances $\sigma^2$

**Design matrix**

$$\begin{array}{cc} & \leftarrow p \text{ variables} \rightarrow \\ \uparrow n \text{ observations} \downarrow & \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{32} & \cdots & x_{3p} \\ 1 & x_{41} & x_{42} & \cdots & x_{4p} \\ 1 & x_{51} & x_{52} & \cdots & x_{5p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{2p} & \cdots & x_{np} \end{bmatrix} \end{array}$$

# Linear regression




Estimated $y$-value

$$\hat{y}_i = \sum_{j=1}^{p} x_{ij}\beta_j + \beta_0$$

*Residual*: the error in each estimate

$$r_i = y_i - \hat{y}_i$$
$$= y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0$$

Sum of squared errors:

$$e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0 \right)^2$$

Science and Technology Facilities Council

Hartree Centre

GS(0

# Linear regression

The (ordinary) least-squares estimator is calculated by minimising the sum of the squared residuals

$$e = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0 \right)^2 \text{ to give } \hat{\boldsymbol{\beta}} = \left(\mathrm{X}^{\boldsymbol{T}}\mathrm{X}\right)^{-\mathbf{1}}\mathrm{X}^{\boldsymbol{T}}\mathrm{y}$$

The dependent variable estimate is $\hat{\mathrm{y}} = \mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}\left(\mathrm{X}^{\boldsymbol{T}}\mathrm{X}\right)^{-\mathbf{1}}\mathrm{X}^{\boldsymbol{T}}\mathrm{y}$

If the errors are independent, uncorrelated, have mean 0 and equal variances $\sigma^2$ then this estimator is

- unbiased
- has the lowest variance of any linear unbiased estimator (Gauss-Markov theorem)

However, linear regression models can be vulnerable to outliers, especially if there is not much data.

Science and Technology Facilities Council

Hartree Centre

**GS(0** Highlight the different terms in this equation?

Goodchild, Simon (STFC,DL,HC), 2022-09-21T11:06:12.564

# Linear regression

The least-squares estimator parameters are $\hat{\beta} = \left(\mathrm{X}^T\mathrm{X}\right)^{-1}\mathrm{X}^T\mathrm{y}$

If there is only one independent variable, the design matrix is

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \\ 1 & x_{41} \\ 1 & x_{51} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$$

Therefore:

$$\mathrm{X}^T\mathrm{X} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

$$\left(\mathrm{X}^T\mathrm{X}\right)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

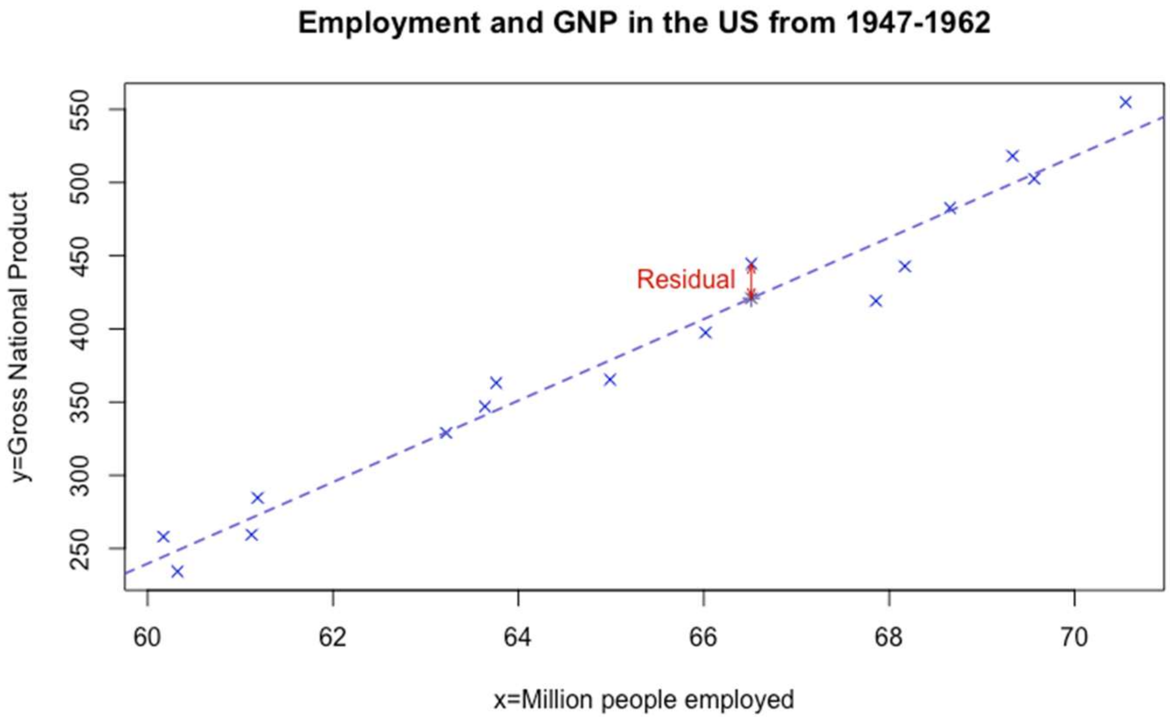
$$\mathrm{X}^T\mathrm{y} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$\hat{\beta} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i \\ n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix}$$

Science and Technology Facilities Council

Hartree Centre

# Linear regression




*Residual*: the error in the estimate

$$r_i = y_i - \hat{y}_i$$

The error variance can be estimated using the residuals

$$s^2 = \frac{\mathbf{r}^T \mathbf{r}}{n - p}$$

Science and Technology Facilities Council

Hartree Centre

GS(0

# Linear regression

A more restrictive assumption is that the errors are are independent, identically distributed $\varepsilon \sim N(0, \sigma^2 \mathrm{I})$
In this case the least-squares estimate is also the *maximum likelihood estimator*

$$L(\beta;\mathrm{x}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left( \frac{-1}{2\sigma^2} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0 \right)^2 \right)$$

The likelihood is the expression for the probability of **x**, but seen as a function of the parameters.

Under this assumption:

the parameter estimates have a Gaussian distribution $\hat{\beta} \sim N(\beta, (\mathrm{X}^T\mathrm{X})^{-1}\sigma^2)$

the error estimate has a chi-squared distribution $s^2 \sim \chi^2_{n-p}$

the parameter with the estimated error has a *t*-distribution

This can be used to estimate a confidence interval for the parameters.

Science and
Technology
Facilities Council
Hartree Centre

**Slide 41**

---

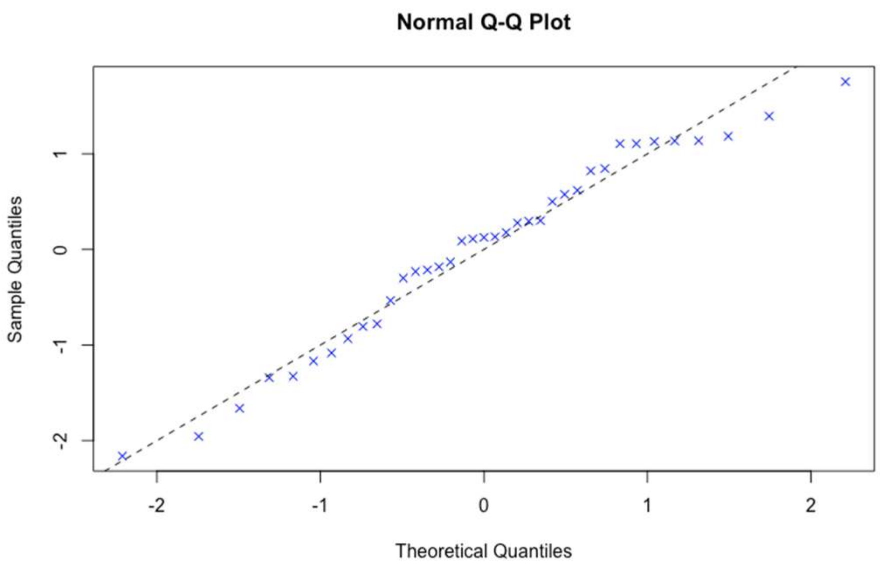
**GS(0** Show an example of this

Goodchild, Simon (STFC,DL,HC), 2022-09-21T15:38:36.820

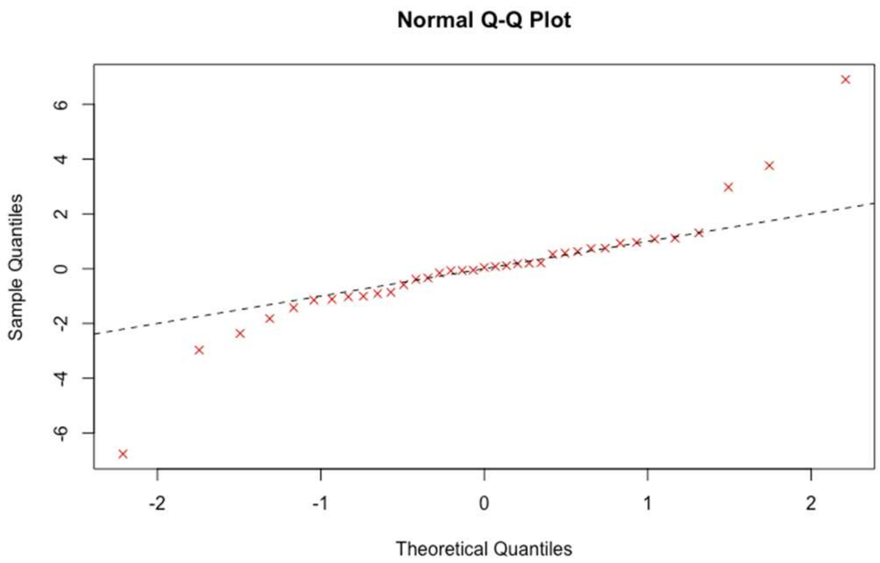
GS(0

# Linear regression

The assumption of normally distributed errors can be tested using a Q-Q plot: of the quantiles of the distribution against the quantiles of a normal distribution




Normally distributed errors




Non-normal (*t*-distributed) errors

Science and Technology Facilities Council

Hartree Centre

**GS(0** Show what happens when you try linear regression with non-normal errors?

Goodchild, Simon (STFC,DL,HC), 2022-10-12T16:12:30.494

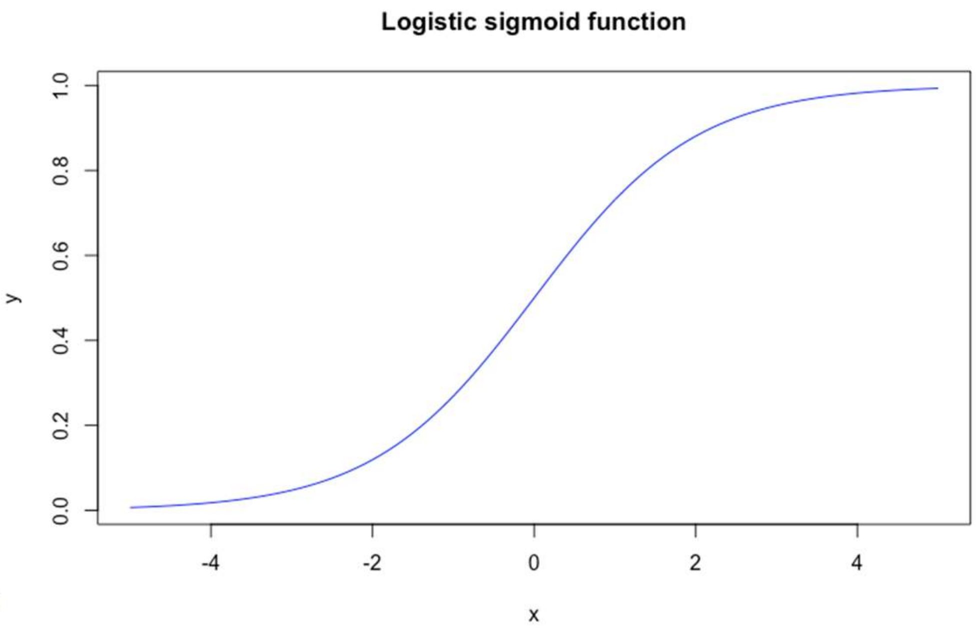
# Logistic regression

Model for a binary classification.

Take a set of $n$ observations with $p$ independent variables $\{y_i, x_{ij}\}$ where now $y$ is 0 or 1

Create a linear model as before: $z_i = \sum_{j=1}^{p} x_{ij}\beta_j$

Convert this model into an 0 or 1 prediction by using a *link function* $\mathrm{P}(y_i = 1) = \frac{1}{1 + e^{-z_i}}$




This can also be viewed as a linear model for the log-odds

$$\log\left(\frac{P}{1-P}\right) = \sum_{j=1}^{p} x_{ij}\beta_j$$

Science and Technology Facilities Council

Hartree Centre

# Logistic regression

Unlike linear regression, logistic regression doesn't have an algebraic solution – the model has to be fitted numerically

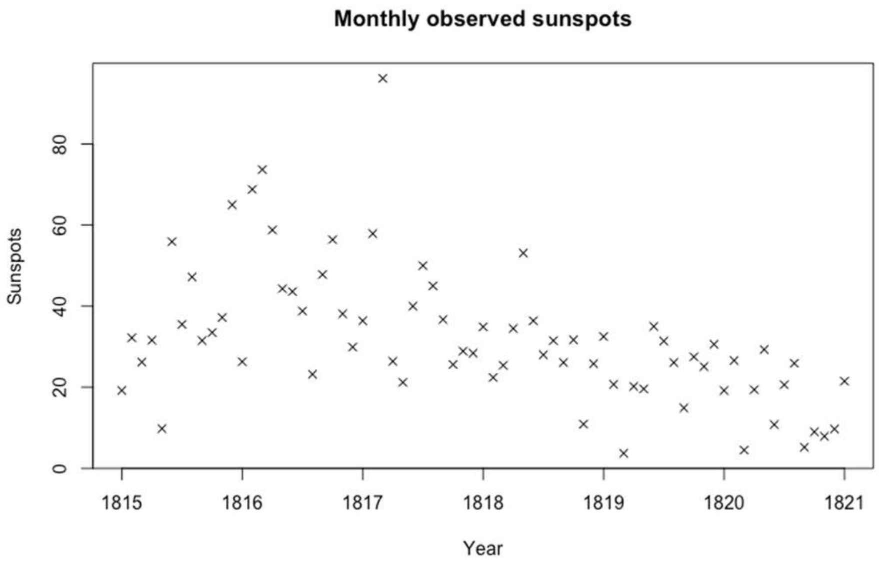
Usually use the method of maximum likelihood again.

Logistic regression is one of a class of *generalised linear models* where a parameter of interest is modelled as a linear function.

# Time series models

A time series is a series of observations taken at regular time intervals.

$X_t$ for $t = 1,2,3, \dots$




A (weakly) stationary time series has a constant mean and autocovariance.

Mean $\mu = \mathrm{E}[X_t]$

Autocovariance $A_\tau = \mathrm{E}[X_t X_{t-\tau}] - \mu^2$

Science and Technology Facilities Council
Hartree Centre

GS(0

# Time series models

Autoregressive (AR) model of order *p* $$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

Moving average (MA) model of order *q* $$X_t = c + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t$$

Autoregressive moving average (ARMA) model of order (*p*,*q*) $$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

If the time series is not stationary, can difference it:

Autoregressive integrated moving average (ARIMA) model of order (*p*,*d*,*q*) applies ARMA(*p*,*q*) to the *d* differences of the time series

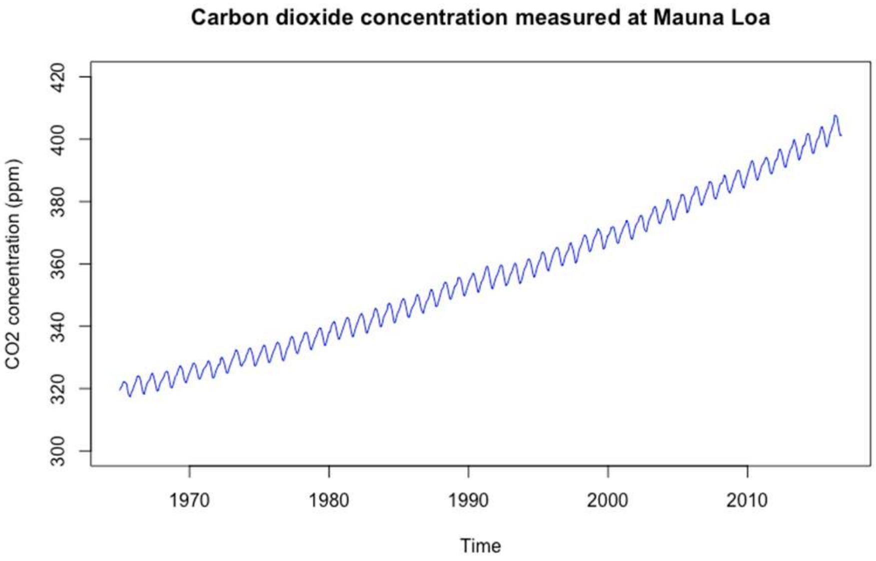
**Slide 46**

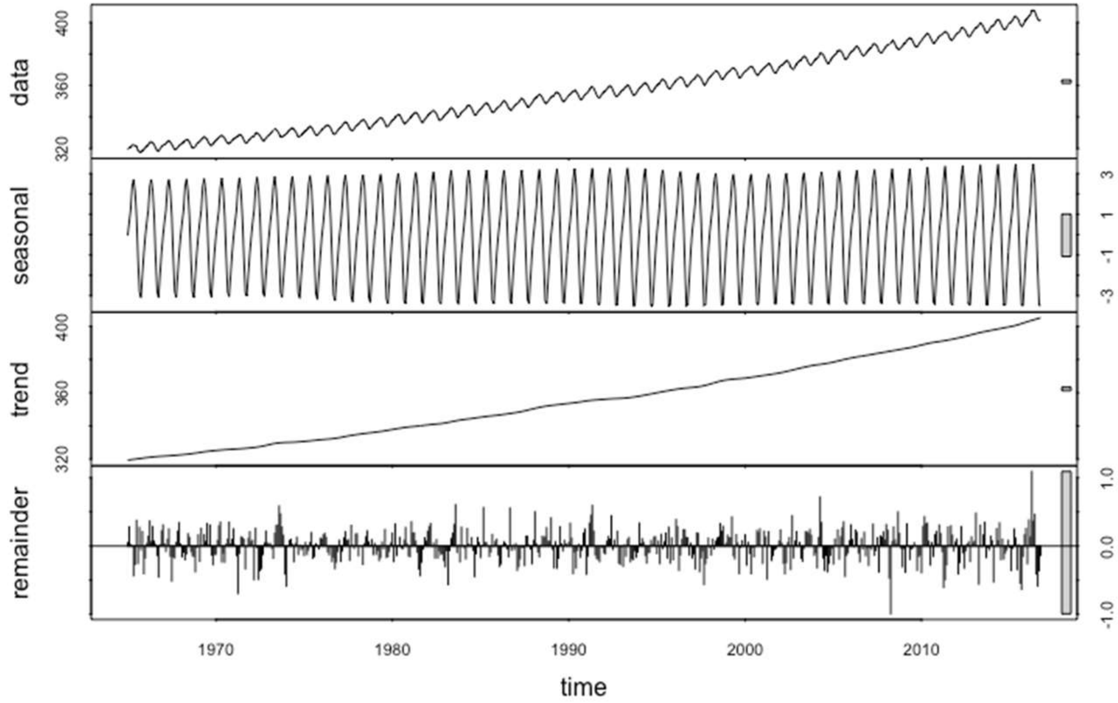
---

**GS(0** Show some examples of ARMA models.

Goodchild, Simon (STFC,DL,HC), 2022-09-21T11:03:18.835

# Time series: trend and seasonality



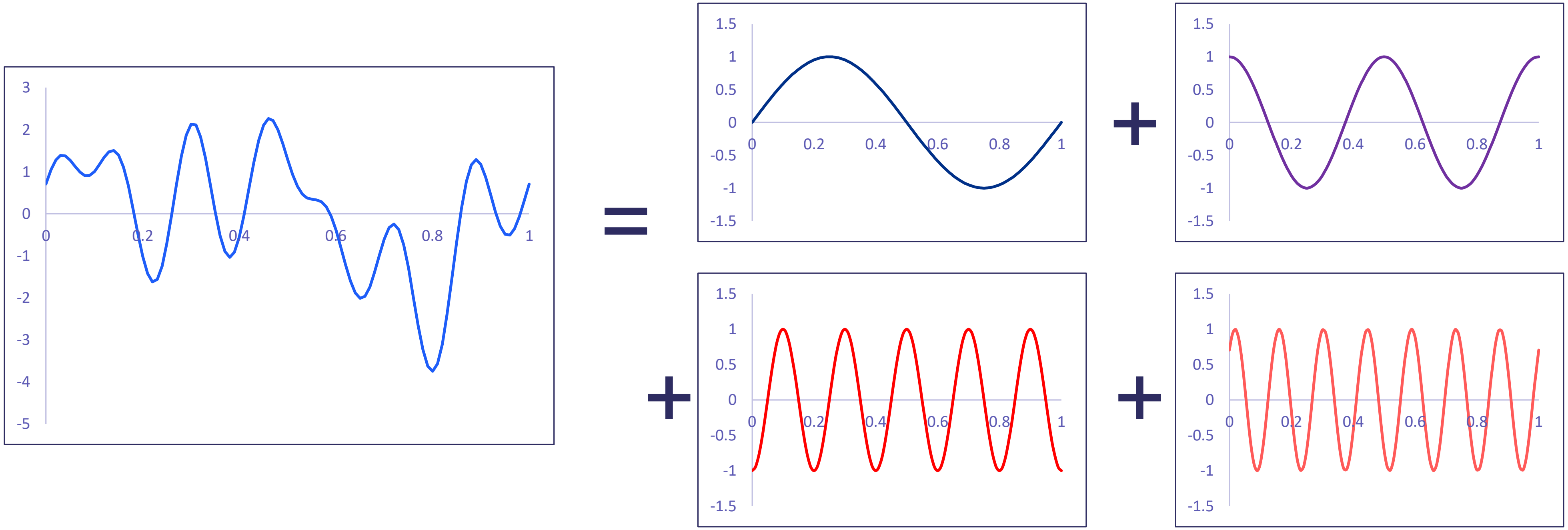




The STL method (Seasonal Decomposition of Time Series by Loess (Locally Estimated Scatterplot Smoothing) decomposes a time series into trend and seasonality.

Science and Technology Facilities Council
Hartree Centre

# Time series: seasonality

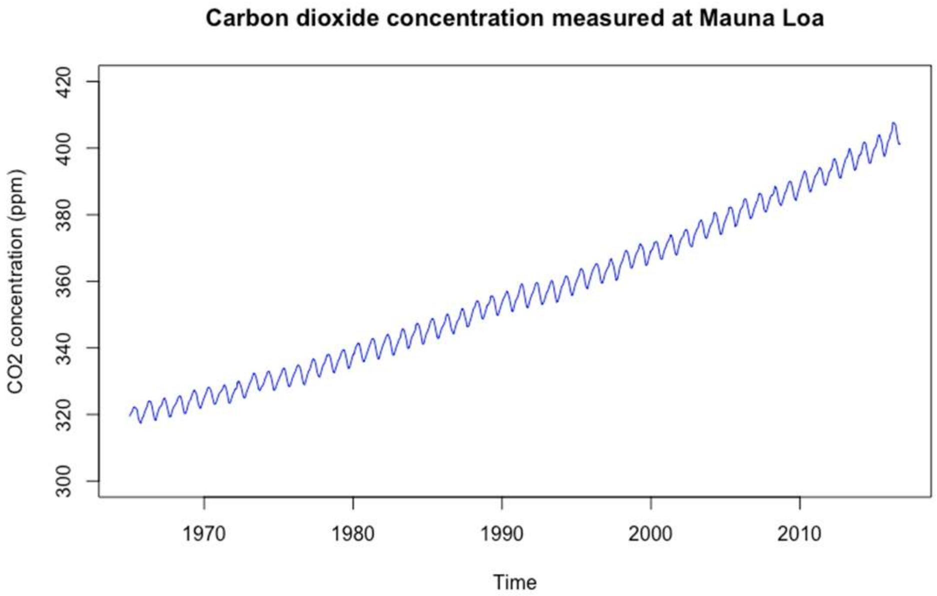
Any function* on a finite interval [0, L] can be expressed as the sum of a series of sine and cosine waves

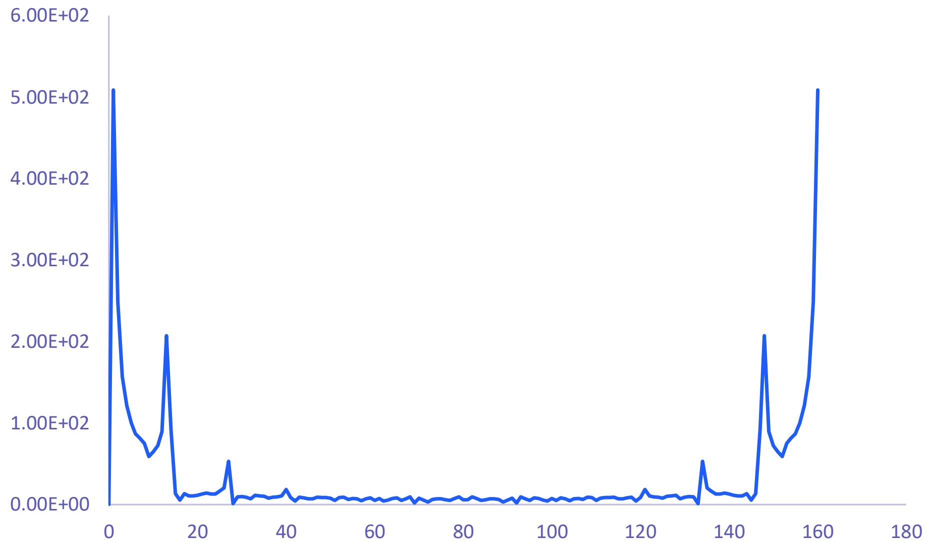



*subject to some conditions on behaviour

Science and Technology Facilities Council
Hartree Centre

# Time series: seasonality




Fourier transform of the $CO_2$ data

Science and Technology Facilities Council

Hartree Centre

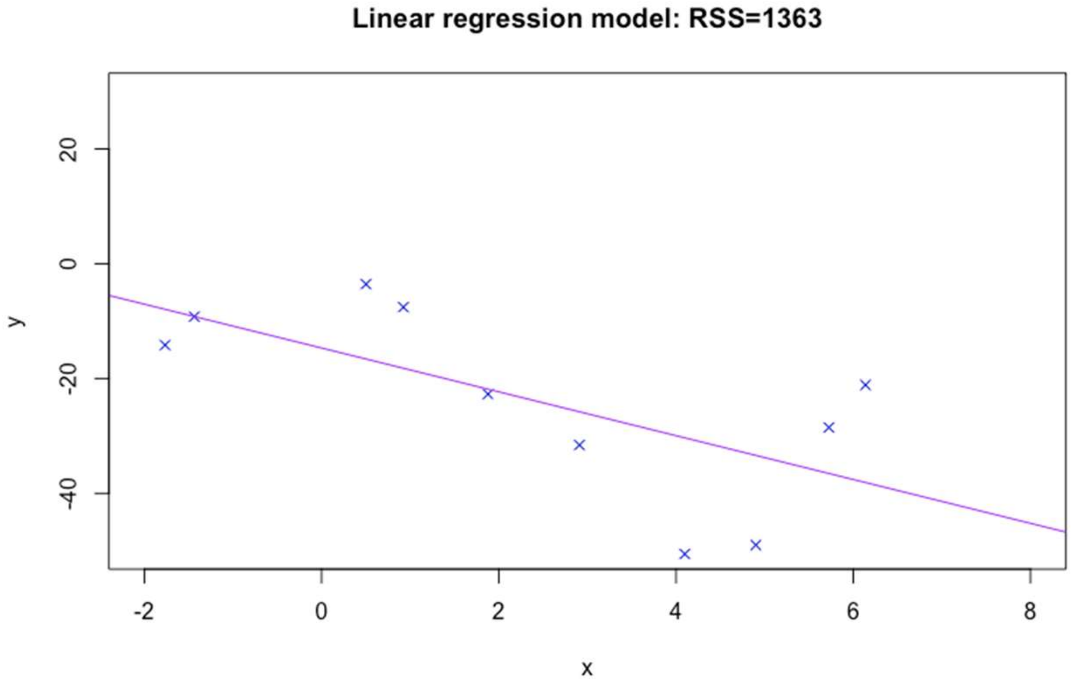
# Polynomial regression

Extension of linear regression where the model is now a polynomial

$$y_i = \beta_0 + \sum_{k=1}^{m} \beta_k \, x_i^k + \varepsilon_i$$

If the different $x_i^k$ are assumed uncorrelated this can be fit using least-squares.

# Underfitting

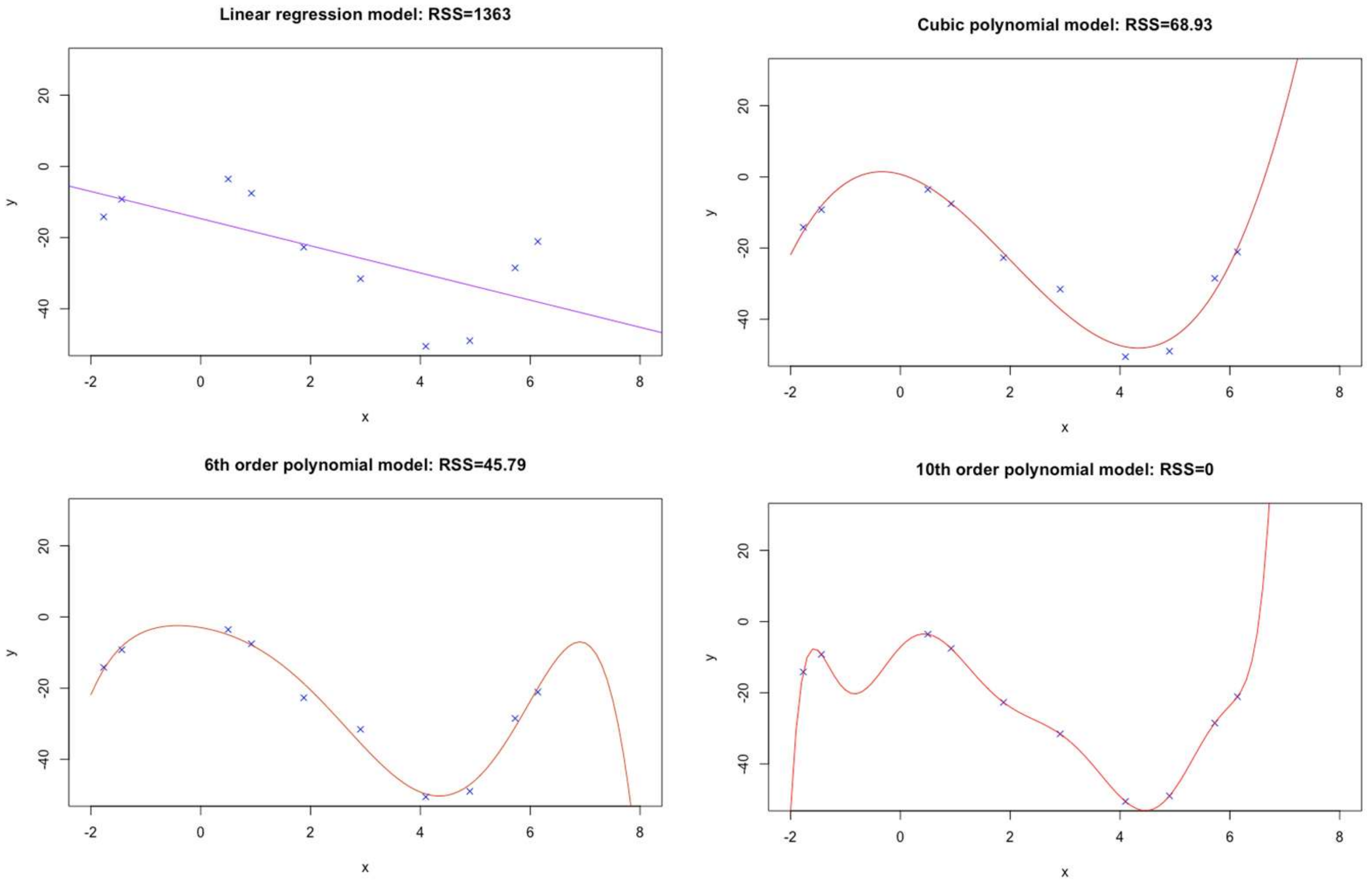



An underfit model doesn't have enough complexity to capture all the behaviour in the data.

Science and Technology Facilities Council
Hartree Centre

# Overfitting

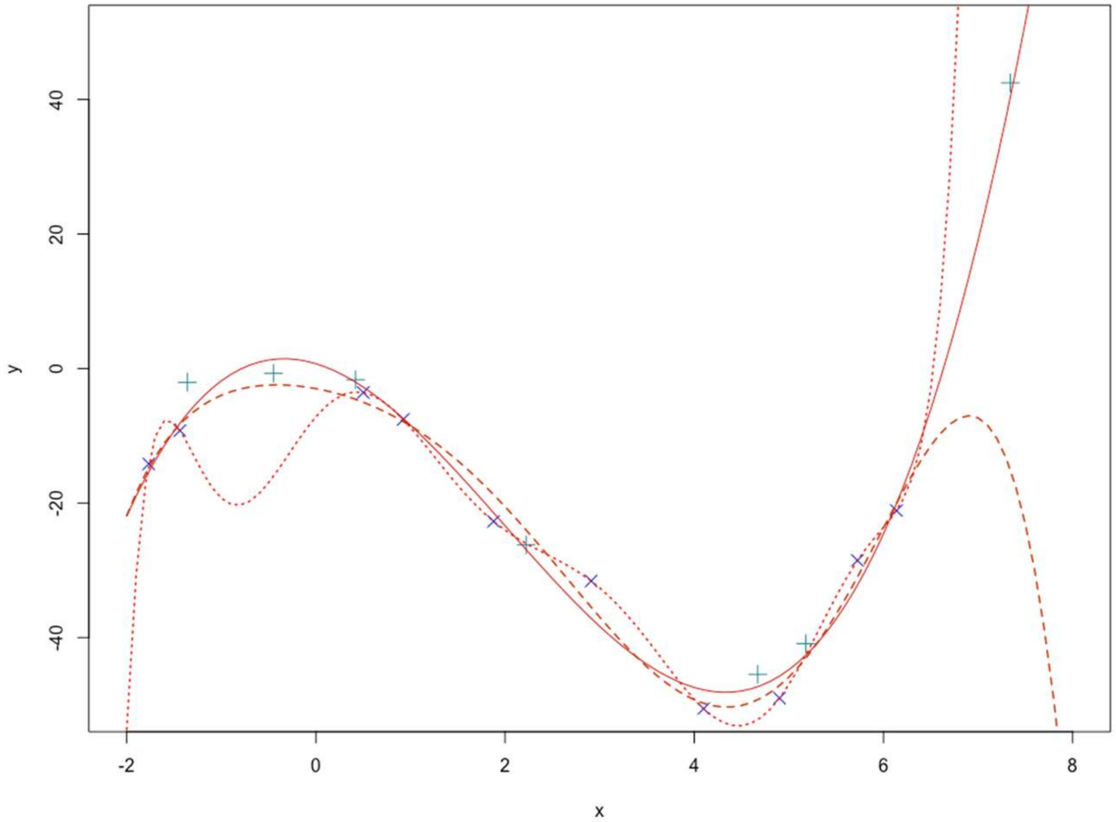
With enough terms in a polynomial model, any set of points can be fitted exactly (as long as all the independent variables are different).




Science and Technology Facilities Council

Hartree Centre

# Overfitting




When new samples are taken, the overfit models decrease in accuracy.

Science and Technology Facilities Council
Hartree Centre

# Regularisation

Regularisation reduces the complexity of a model by adding a penalty for having more coefficients

Instead of minimising the mean-squared error

$$e = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0 \right)^2$$

ridge regression minimises

$$e = \sum_{i=1}^{n} \left( y_i - \sum_{j=1}^{p} x_{ij}\beta_j - \beta_0 \right)^2 + \lambda \left( \sum_{j=1}^{p} \beta_j^2 \right)$$

This estimator is no longer unbiased, but can be less prone to overfitting

Science and Technology Facilities Council
Hartree Centre

Science and Technology Facilities Council

Hartree Centre

# Exploratory data analysis

Science and Technology Facilities Council

Hartree Centre

# Exploratory data analysis

*Exploratory data analysis* or EDA is the process of preparing data for probabilistic modelling.

- Removing erroneous values
- Checking the validity of data
- Finding patterns and features which may inform model-building

The process from EDA to modelling is not straightforward
Model building may identify problems in the data which then need to be investigated.

GS(0

# Errors

| | Yr | Mn | Date | Date.1 | CO2 |
|---|---|---|---|---|---|
| 1 | 1958 | 1 | 21200 | 1958.0411 | -99.99 |
| 2 | 1958 | 2 | 21231 | 1958.126 | -99.99 |
| 3 | 1958 | 3 | 21259 | 1958.2027 | 315.69 |
| 4 | 1958 | 4 | 21290 | 1958.2877 | 317.45 |
| 5 | 1958 | 5 | 21320 | 1958.3699 | 317.5 |
| 6 | 1958 | 6 | 21351 | 1958.4548 | -99.99 |
| 7 | 1958 | 7 | 21381 | 1958.537 | 315.86 |
| 8 | 1958 | 8 | 21412 | 1958.6219 | 314.93 |
| 9 | 1958 | 9 | 21443 | 1958.7068 | 313.21 |
| 10 | 1958 | 10 | 21473 | 1958.789 | -99.99 |

This is the raw CO2 data used earlier.

The highlighted values are errors and need to be discarded

Errors may be denoted by 0, a plainly absurd (e.g. negative value) or a non-number placeholder like 'NA'

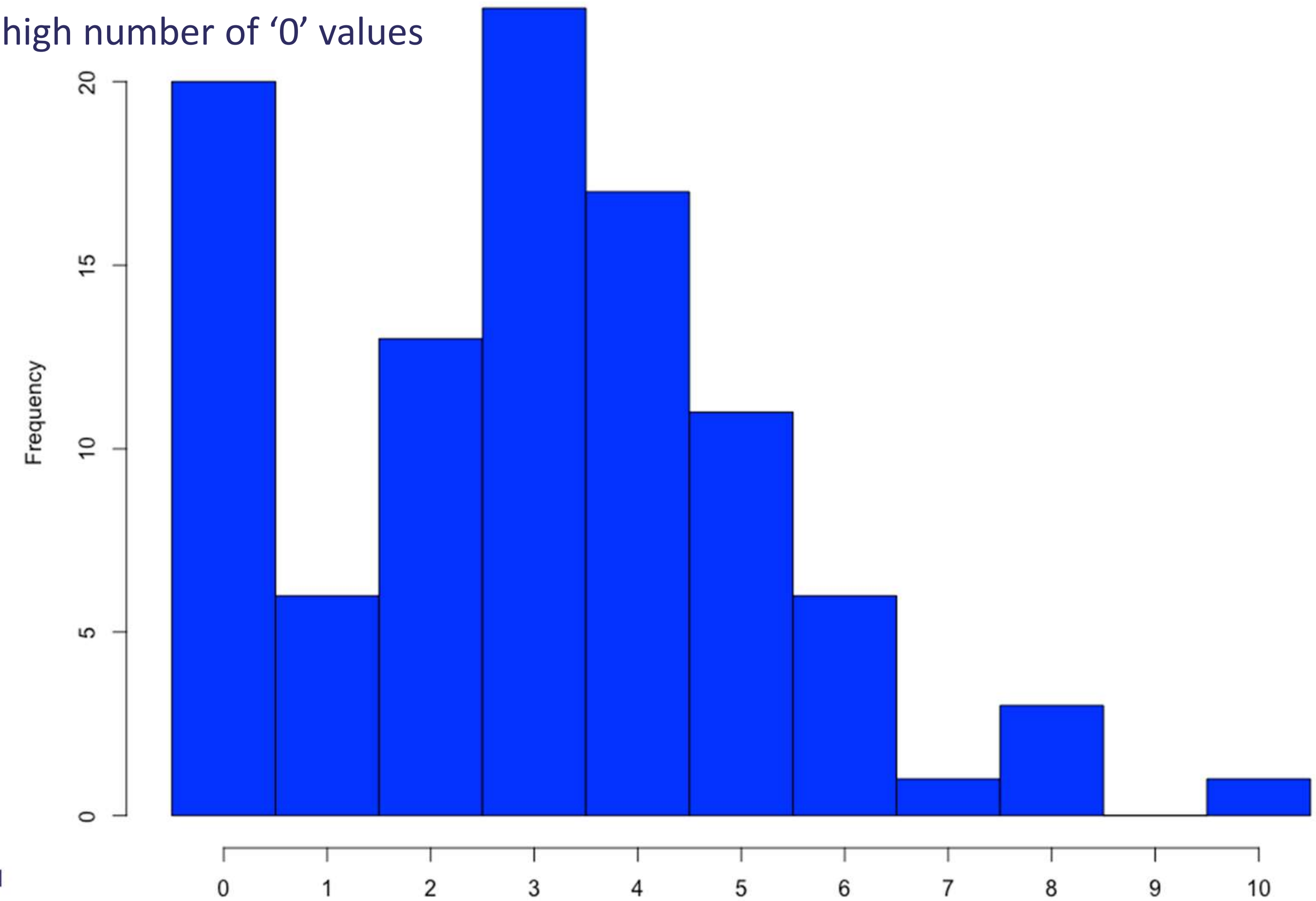
Science and Technology Facilities Council
Hartree Centre

**Slide 57**

**GS(0** It might be an idea to demonstrate the technique of finding 0 errors using a histogram.
Goodchild, Simon (STFC,DL,HC), 2022-10-12T12:30:43.224

# Errors

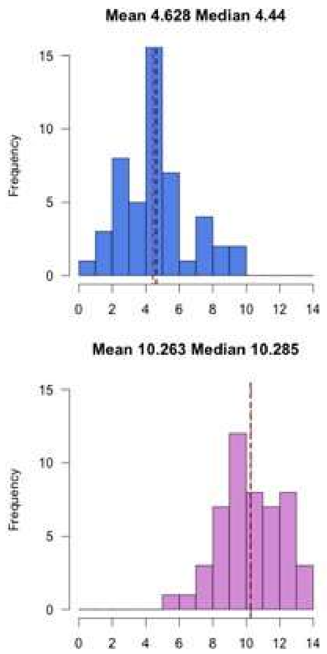
Note the unusually high number of ‘0’ values



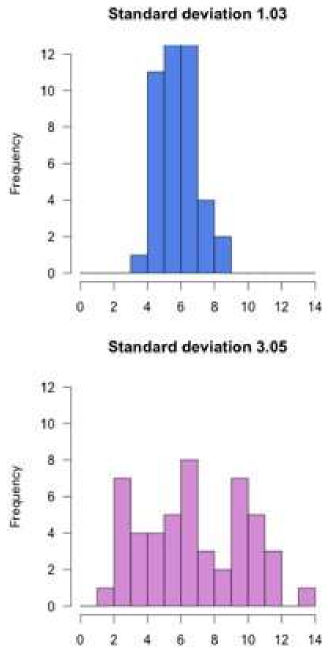

Science and Technology Facilities Council
Hartree Centre

# Summary statistics

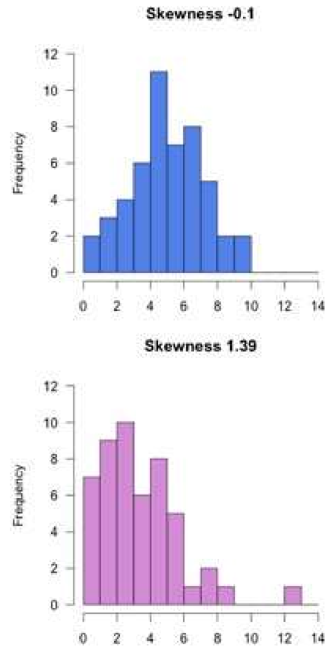
Location:
mean or median




Spread:
variance or FWHM




Asymmetry:
skewness



GS(0

# Summary statistics at scale

There are two ways of calculating sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Directly

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$$

Using the total sum of squares

Normally the second way is used as you only need one subtraction, but with a very large dataset the two terms in the second way may get very large, so the subtraction may have numerical problems.

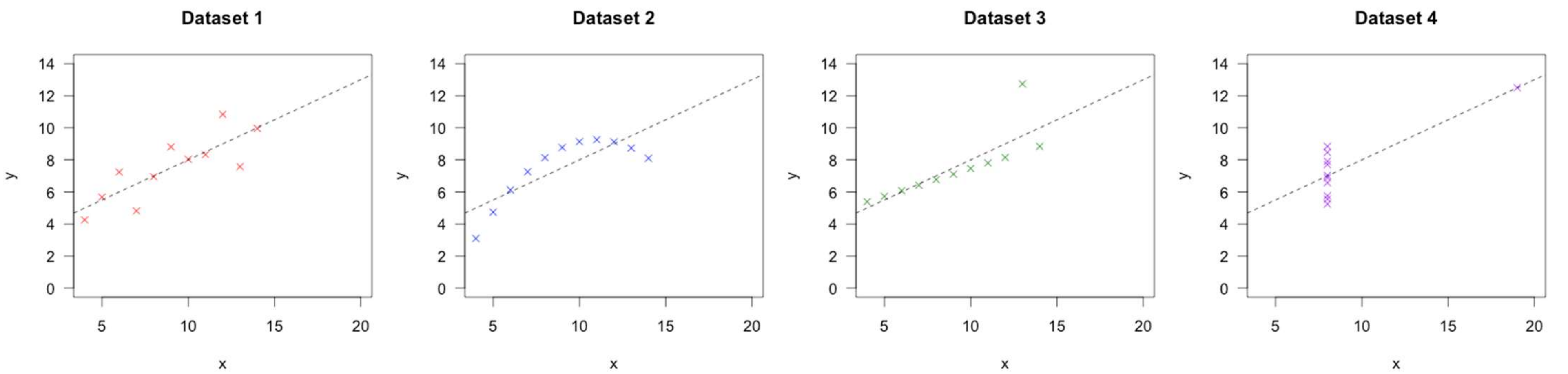
**Slide 60**

**GS(0** Illustrate the problem directly with some floating point numbers?

Goodchild, Simon (STFC,DL,HC), 2022-09-21T11:08:19.691

# Anscombe’s Quartet

| | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
|---|---|---|---|---|
| Mean x | 9.00 | 9.00 | 9.00 | 9.00 |
| Standard deviation x | 3.12 | 3.12 | 3.12 | 3.12 |
| Mean y | 7.50 | 7.50 | 7.50 | 7.50 |
| Standard deviation y | 2.03 | 2.03 | 2.03 | 2.03 |
| Gradient | 0.50 | 0.50 | 0.50 | 0.50 |
| Intercept | 3.00 | 3.00 | 3.00 | 3.00 |

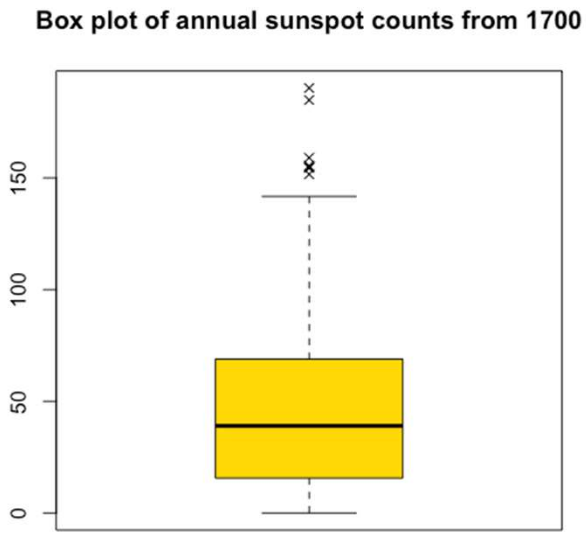



Science and Technology Facilities Council
Hartree Centre

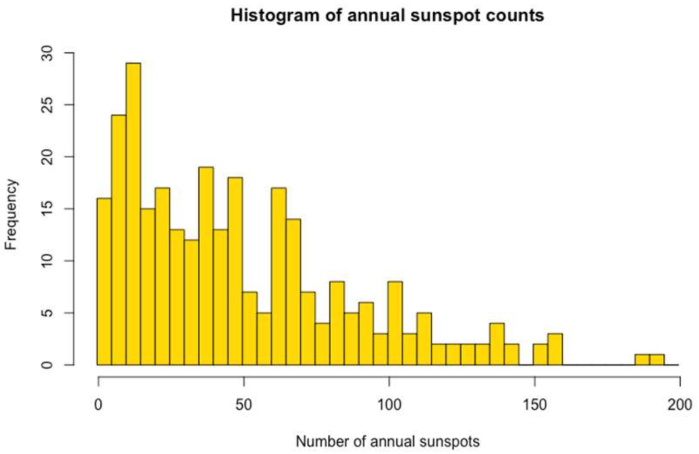
⇒ **Always visualise your data!**

# Data visualisation techniques

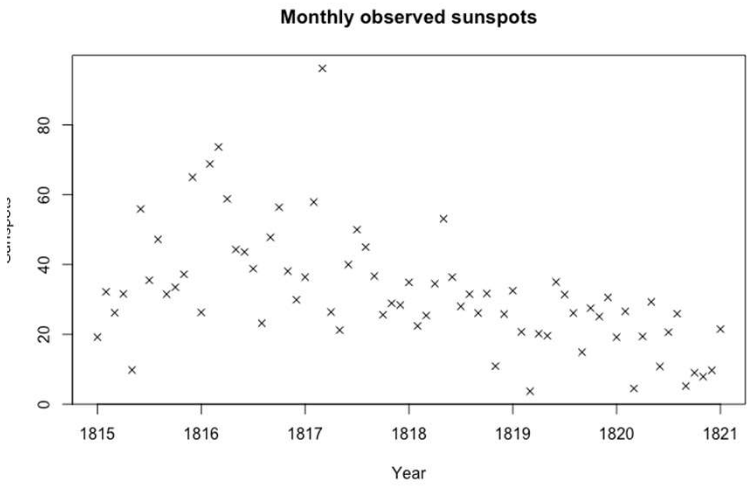
Box/box and whiskers plot




Histogram




Scatter plot

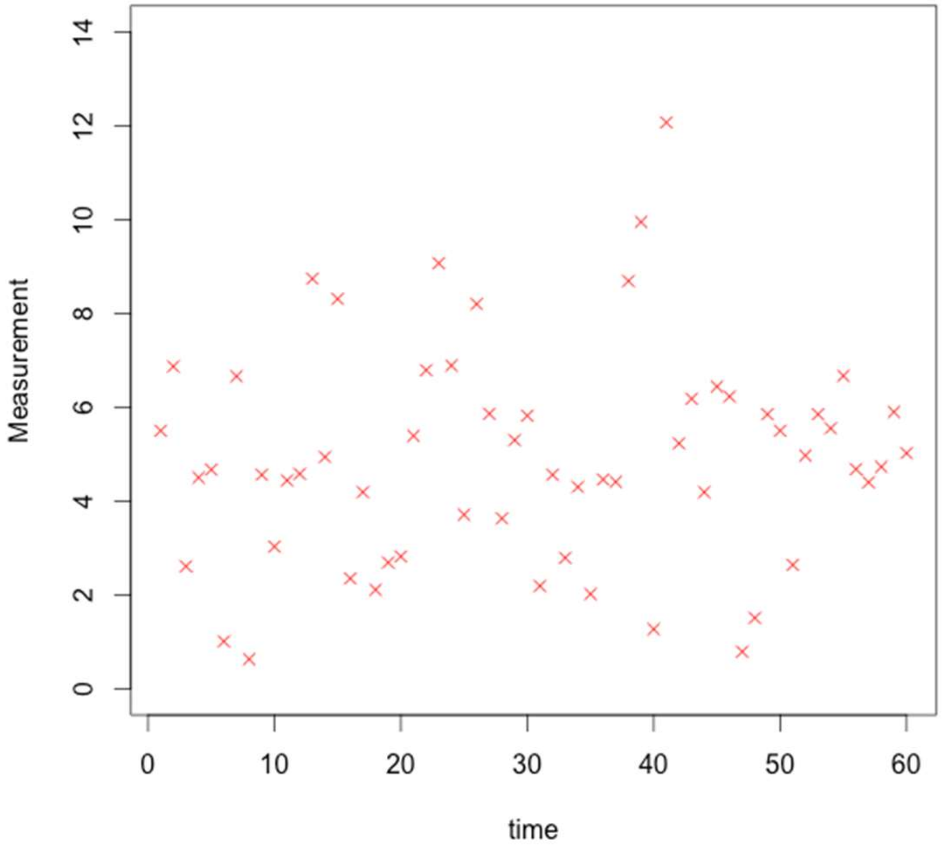



Science and Technology Facilities Council

Hartree Centre

# Anomaly detection

**Problem:**
Given this series of measurements, are any of these unusual?




Science and Technology Facilities Council
Hartree Centre

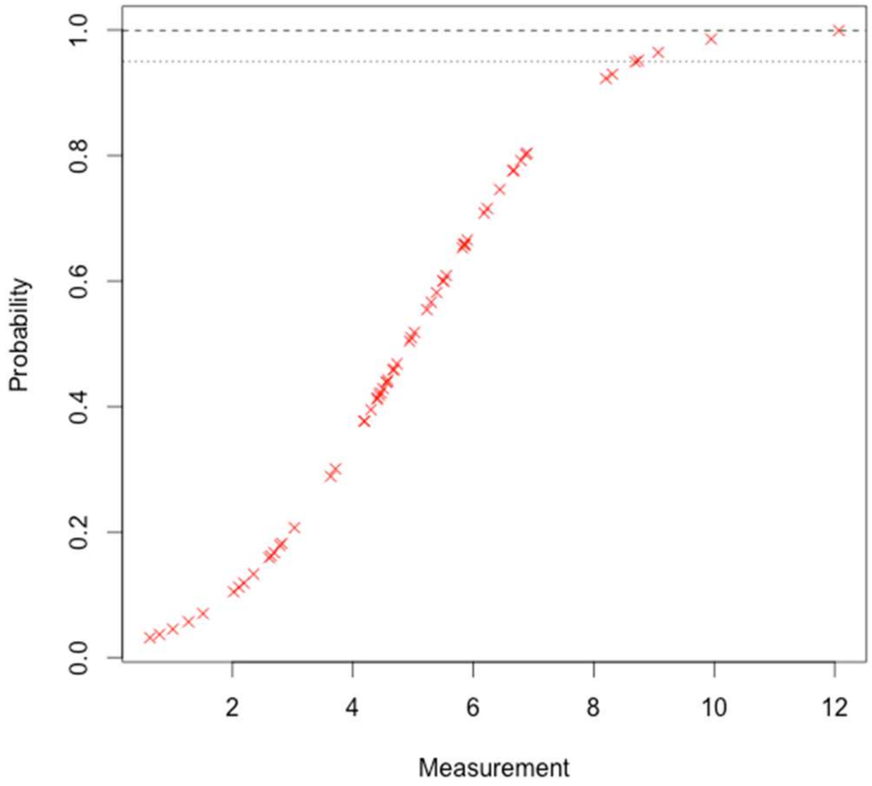
# Anomaly detection

**Modelling approach:** build a statistical model for the data.

Model this data with a normal distribution

Estimate the p-values for the data points based on the model

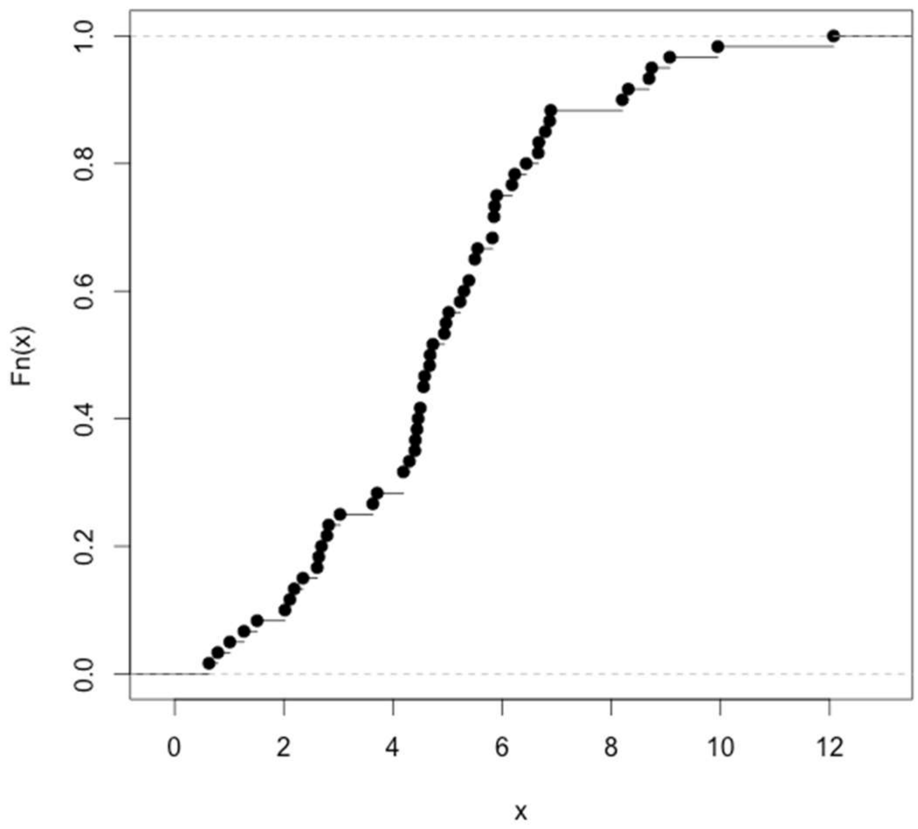
**Careful:** The more measurements we take, the more likely we are to see unusual values.




Science and Technology Facilities Council
Hartree Centre

# Anomaly detection

**Non-parametric approach:** estimate the probability of each point based on the others.




This is called the *empirical cumulative distribution function* or ECDF

Science and Technology Facilities Council

Hartree Centre

# EDA tips

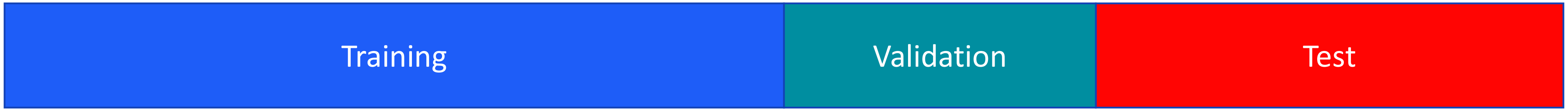
- **Always visualise your data**
- Watch out for errors and anomalies which may mess up your calculations
- Don’t just look at the start and end of big data files: take a random sample

Science and
Technology
Facilities Council
Hartree Centre

# Model training


Split data into training set (to build model) and test set (to assess its accuracy)


Split data into training set (to build model), validation set (to test model parameters) and test set (to assess its accuracy)



Science and Technology Facilities Council

Hartree Centre

# Questions?

Science and Technology Facilities Council

Hartree Centre

# Thank You

Dr Simon Goodchild
simon.goodchild@stfc.ac.uk
sggoodc@liv.ac.uk