Statistics and Scientific Method

Peter J Diggle

Lancaster University

CHICAS centre for health informatics, computing, and statistics

Course text

Diggle, P.J. and Chetwynd, A.G. (2011). *Statistics and Scientific Method: an Introduction for Students and Researchers*. Oxford: Oxford University Press.

AIMS Rwanda, March 2022

Additional references

Box, G.E.P, Hunter, W.G. and Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building* New York: Wiley

Cox, D.R. (1972). Regression models and life tables (with Discussion).*Journal of the Royal Statistical Society* B, **34**, 187–220.

Diggle, P.J. (1990). *Time Series: a biostatistical introduction*. Oxford: Oxford University Press.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data (second edition)*. Oxford: Oxford University Press.

Diggle, P.J. and Giorgi, E. (2019). *Model-based Geostatistics: Methods and Applications in Global Public Health*. Boca Raton: CRC Press

Diggle, P.J., Menezes, R. and Su, T.-L. (2010). Geostatistical analysis under preferential sampling (with Discussion).*Applied Statistics*, **59**, 191–232.

Eglen, S.J. and Wong, J.C.T. (2008). Spatial constraints underlying the retinal mosaics of two types of horizontal cells in cat and macaque. *Visual Neuroscience*, **25**, 209–213.

Graff-Lonnevig, V. and Browaldh, L. (1990). Twelve hours bronchodilating effect of inhaled Formoterol in children with asthma: a double-blind cross-over study versus Salbutamol. *Clinical and Experimental Allergy*, **20**, 429–432.

Grimmett, G. and Stirzaker, D. (2020). *Probability and Random Processes: Fourth Edition.* Oxford: Oxford University Press

Lee, P.M. (2012). *Bayesian Inference: an Introduction (4th edition)*. Chichester: Wiley

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (second edition). London : Chapman and Hall.

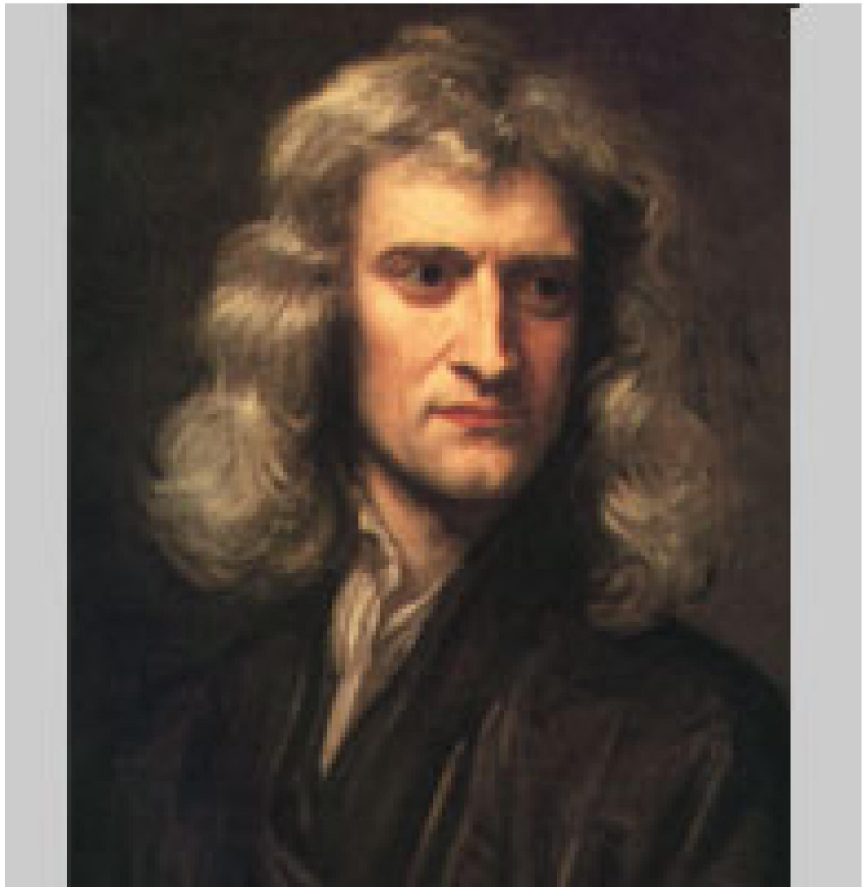
Mercer, W.B. and Hall, A.D. (1911). The experimental error of field trials. *Journal of Agricultural Science*, **4**, 107–132.

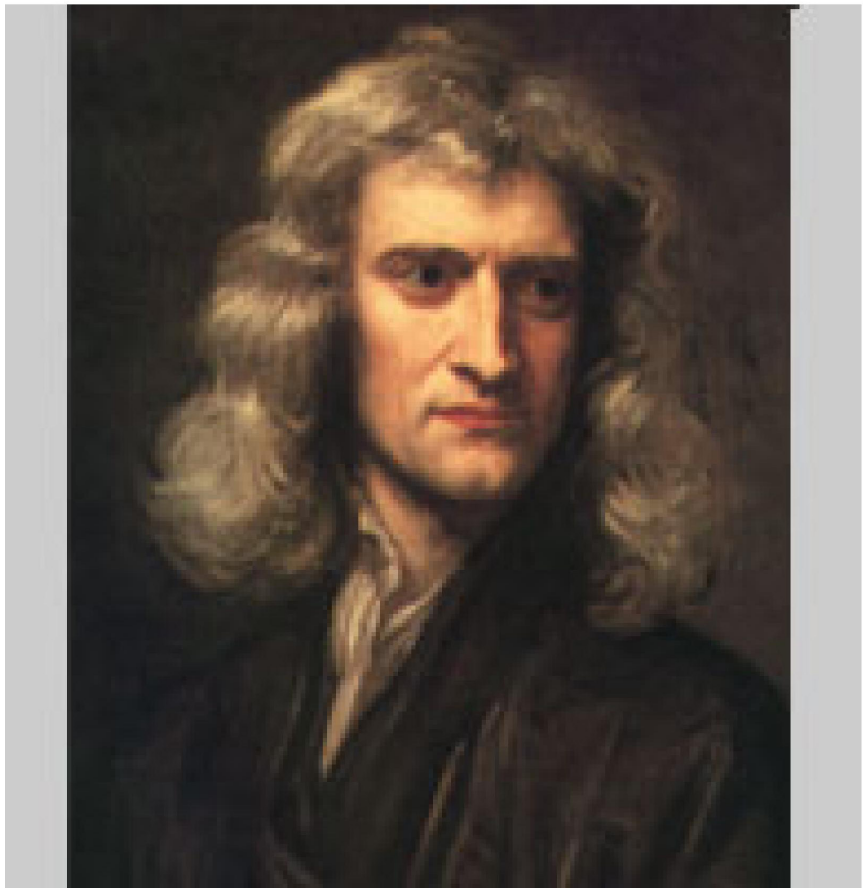
Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, **8**, 158–183.

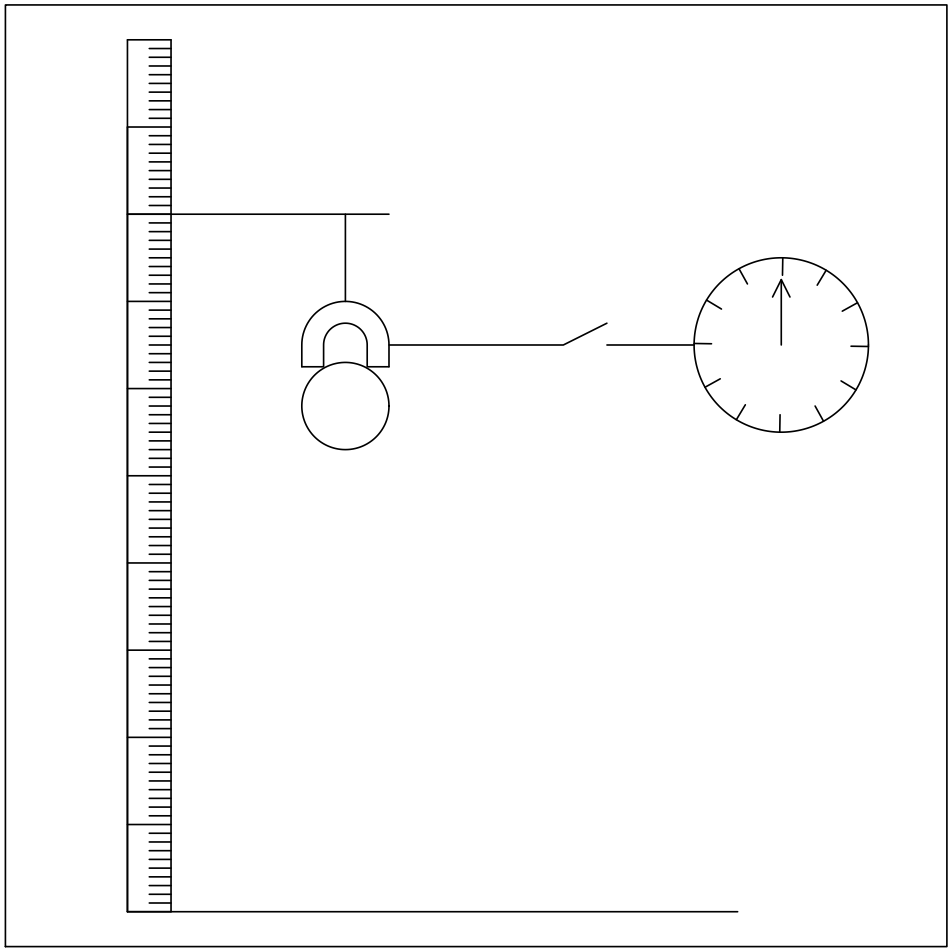
Overview





Isaac Newton (1643–1727)

An experiment to illustrate Newton's Law



Experimental results

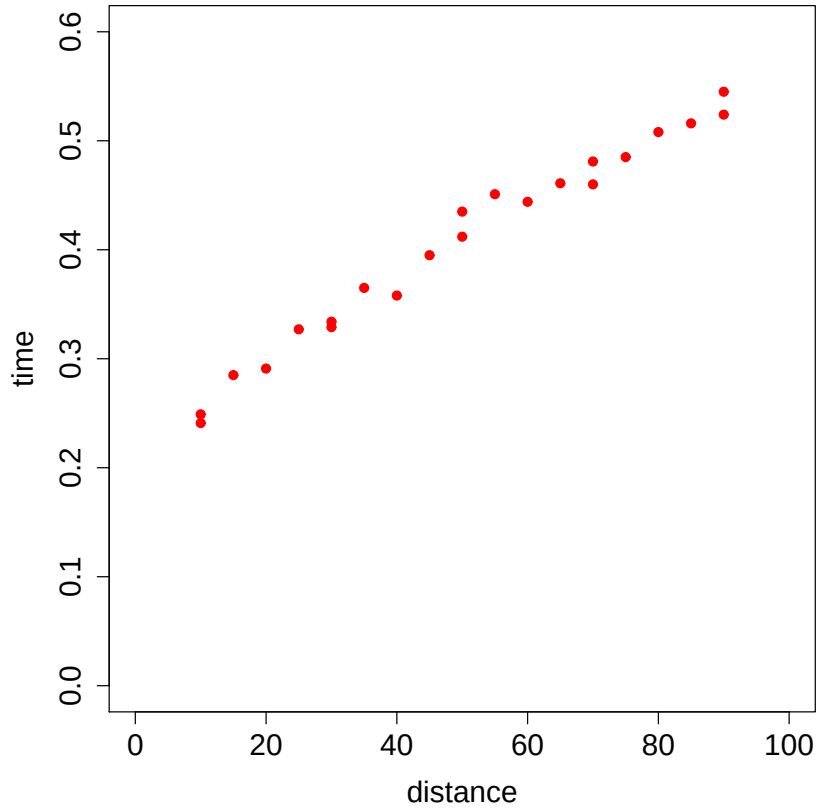
The data reproduced below are the results obtained by one student from 22 experimental runs.

t(sec)	d(cm)	t(sec)	d(cm)	t(sec)	d(cm)
0.241	10	0.358	40	0.460	70
0.249	10	0.395	45	0.485	75
0.285	15	0.435	50	0.508	80
0.291	20	0.412	50	0.516	85
0.327	25	0.451	55	0.524	90
0.329	30	0.444	60	0.545	90
0.334	30	0.461	65		
0.365	35	0.481	70		

Points for discussion

- **choice of design points, d?**
- **inputs and outputs?**
- **sources of systematic variation in results?**
- **sources of random variation in results?**

Experimental results in graphical form



- **How would you describe the relationship between distance and time?**
- **Why did I put time on the *y*-axis of the graph, rather than on the *x*-axis?**

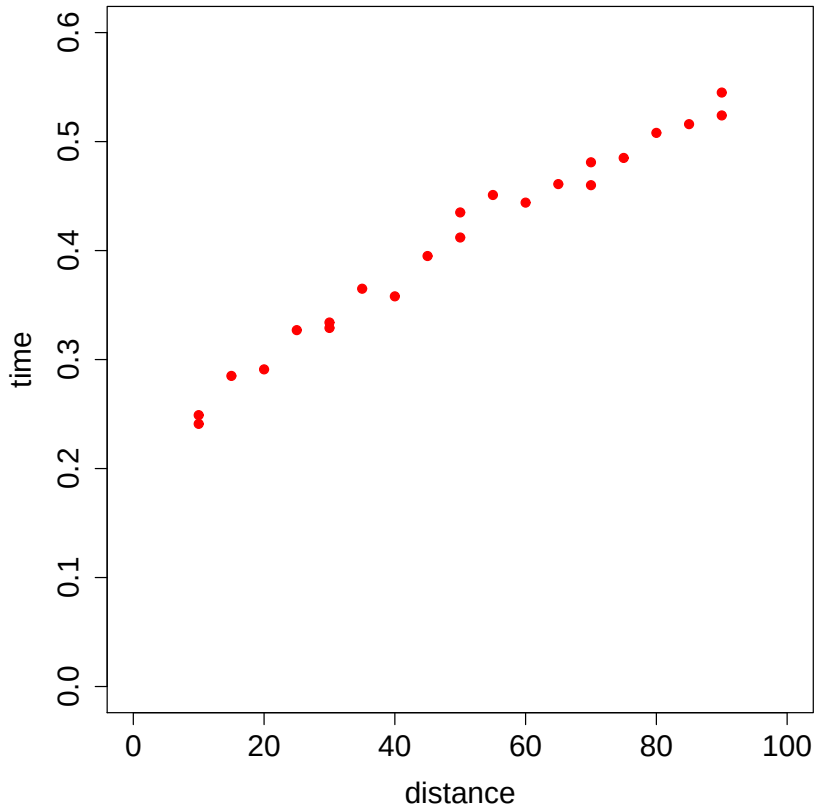
Newtonian mechanics

Newton's law states that the vertical distance d travelled in time t by a body initially at rest and falling under the influence of gravity is given by the formula

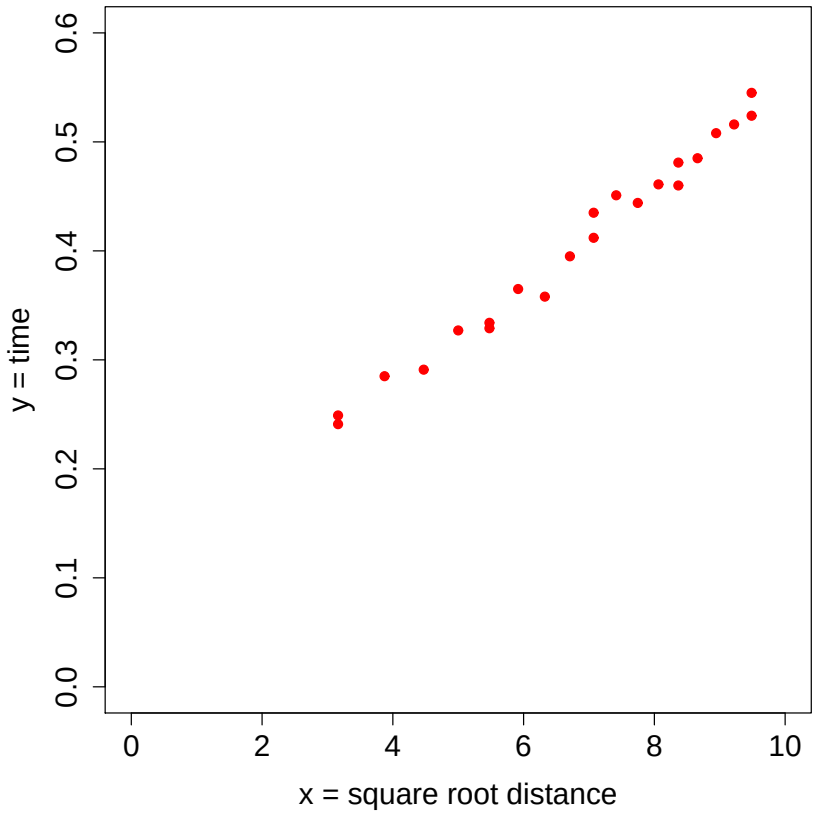
$$d = \frac{1}{2}gt^2$$

where g is a constant.

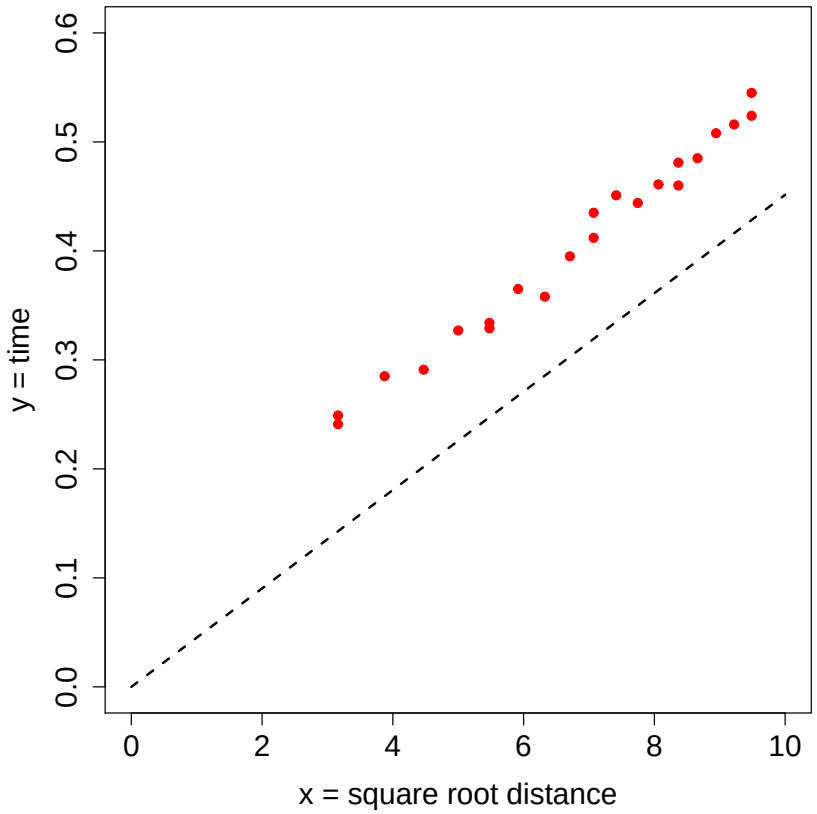
Newtonian mechanics and experimental data



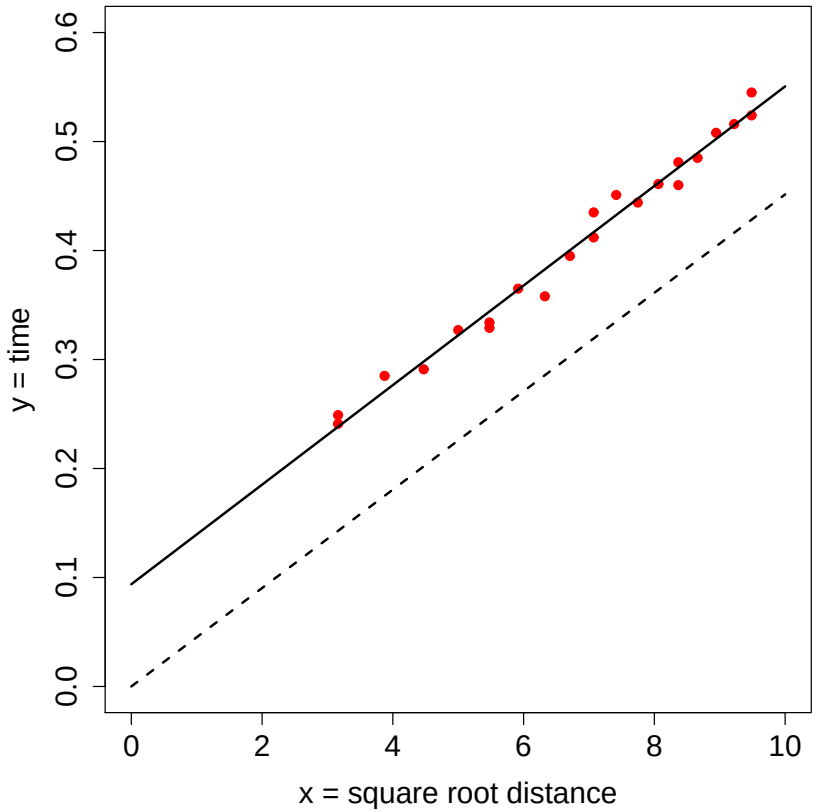
- **The experimentally observed relationship is non-linear, as predicted by Newton's law.**



- **Transformation of the data, from distance d to the new variable $x = \sqrt{d}$ makes the relationship linear, also as predicted by Newton's law.**

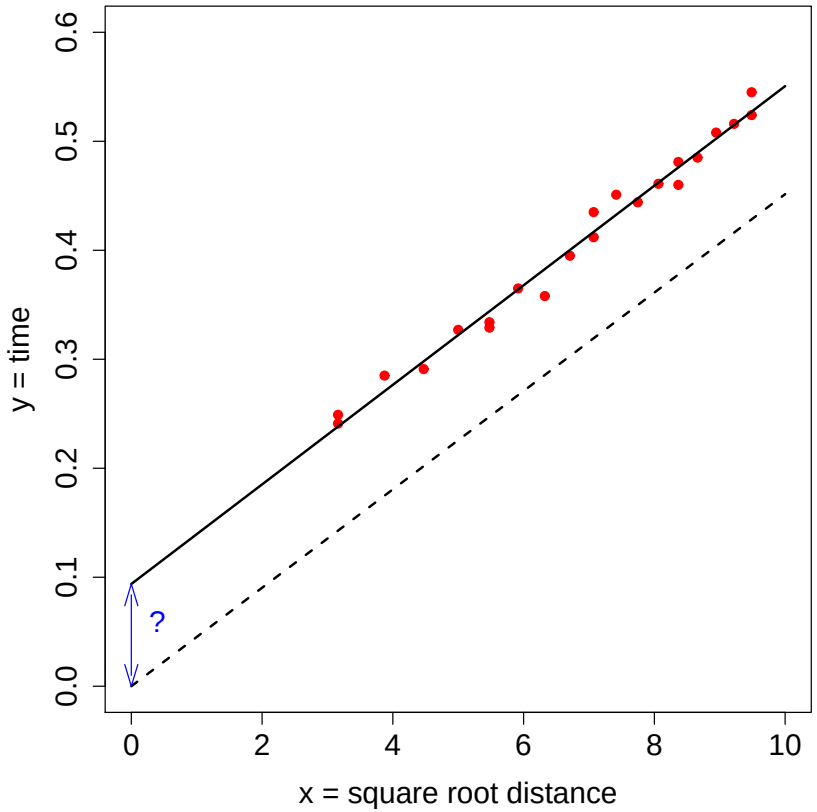


- **But Newton's law does not fit the data!**



- **We need to add an intercept to the straight line**

.



- **What does the intercept represent?**
- **What does the variation of the data around the fitted line represent?**

A statistical model for the experiment

Start with Newtonian mechanics,

$$d = \frac{1}{2} g \times t^2$$

Now transform to $x = \sqrt{d}$ and $Y = t$, and put Y on the left-hand side of the equation,

$$x = \sqrt{g/2} \times Y \qquad Y = \beta \times x \quad (\beta = \sqrt{2/g})$$

Now incorporate the effects of the experimenter's reaction-time,

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

Interpreting the model

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

- α represents mean reaction time
- $\beta = \sqrt{2/g}$ is the quantity of scientific interest
- Z is a random error, which varies independently between different runs of the experiment

Statistical method

- **Design**

  What data should be collected in order to answer a scientific question as precisely as possible?

- **Modelling**

  How can the variation in the data be described mathematically, so that the description:

  - **is not demonstrably inconsistent with the data**
  - **incorporates the underlying science, to the extent that this is well understood**
  - **is a simple as possible subject to the above two constraints**

- **Inference**

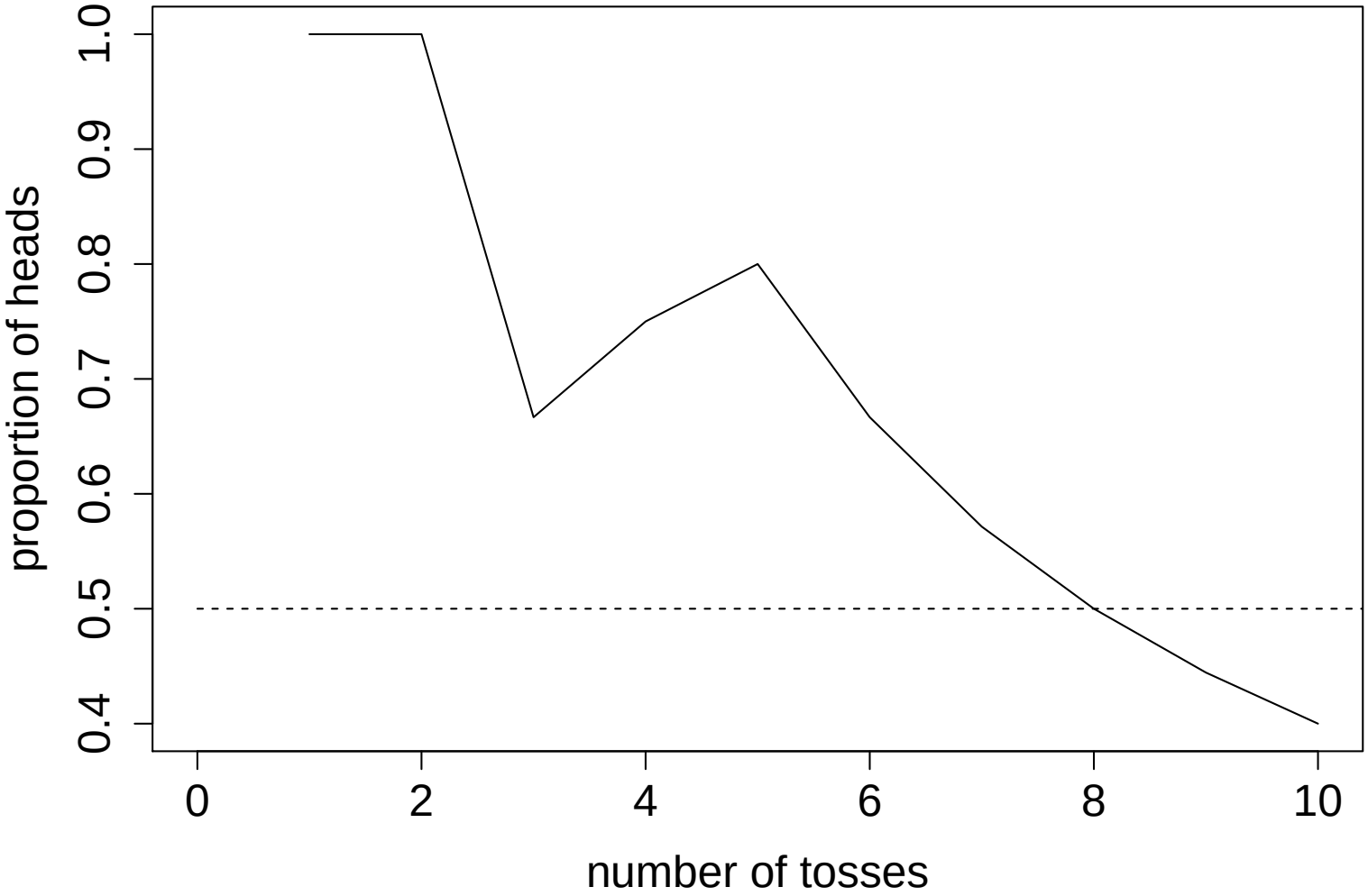
  Given the data and the model, what scientific conclusions can be drawn?

Review: what have we learnt?

1. **Graphical presentation of data is almost always useful.**
2. **Statistical models should:**
   - **respect the data;**
   - **respect the underlying science.**
3. **Transforming the data can help to achieve both goals.**
4. **The results of a statistical analysis should always be interpreted in relation to the original scientific question.**
5. **The fundamental role of statistical method is to enable scientists to answer their questions as precisely as possible.**
6. **The fundamental ingredients of statistical method are: design; modelling; inference.**

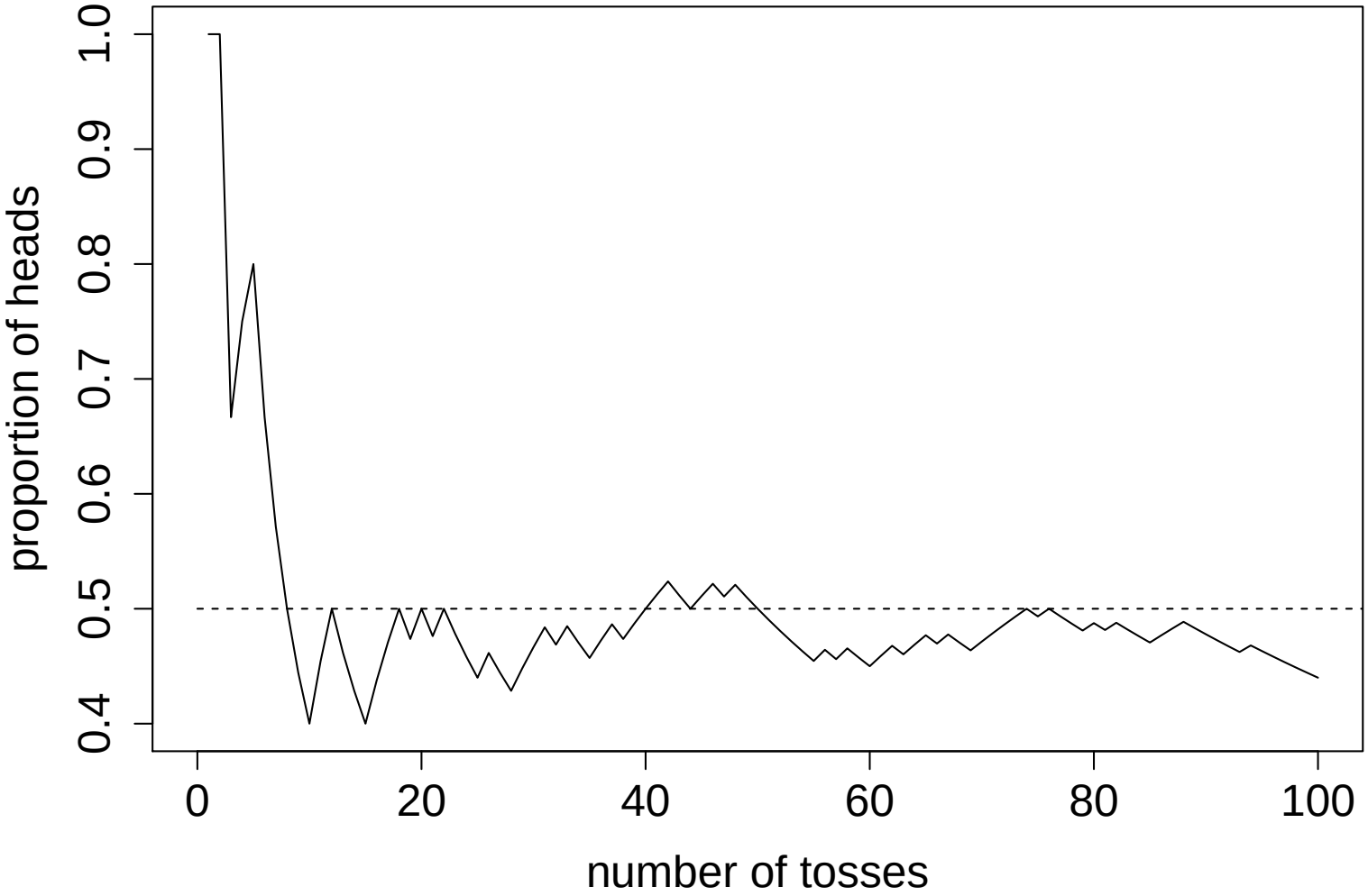
Uncertainty

Tossing a coin



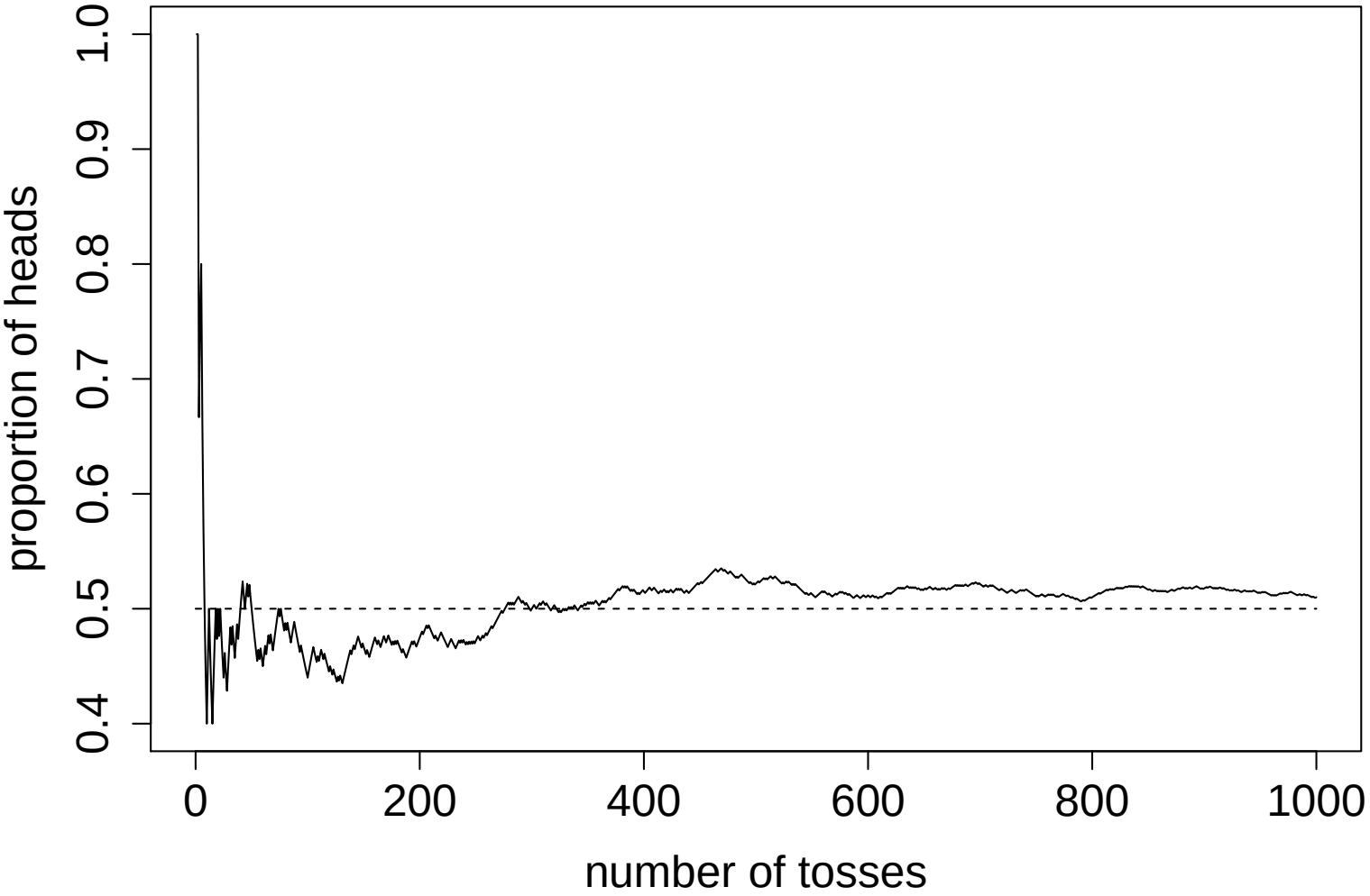
Is the coin fair?

Tossing a coin



Is the coin fair?

Tossing a coin



Is the coin fair?

Variation

Variation in experimental results can arise in two qualitatively different ways:

- **systematic: a change in the experimental conditions produces a different result**
- **stochastic: replication under identical conditions produces a different result**

Even a well-designed experiment can therefore lead to uncertainty in how to interpret the results, but:

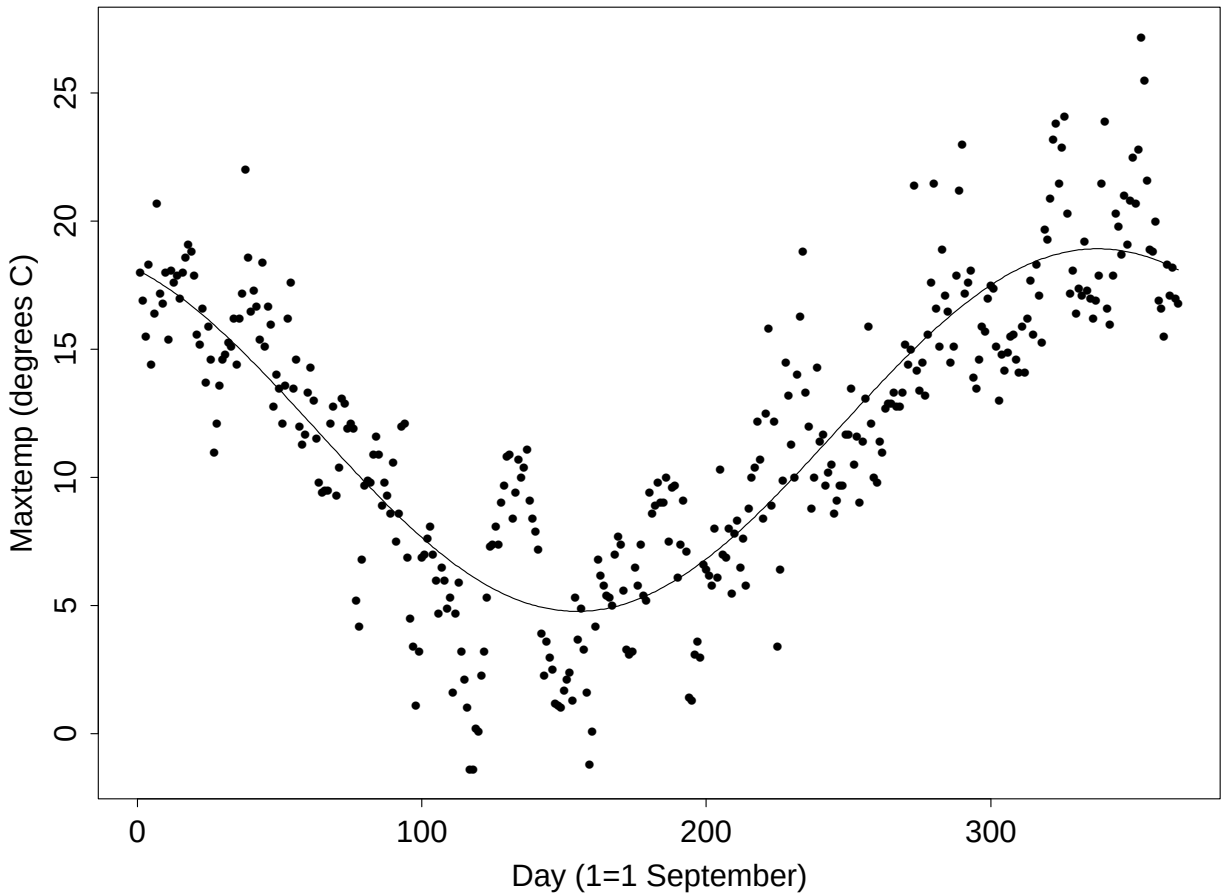
- **stochastic does not necessarily mean completely random**
- **statisticians use probability to measure uncertainty**

Example: weather forecasting

How uncertain are we about the weather later today? tomorrow? next week? next year?

A meteorological time series

- **maximum daily temperatures (degrees C) at Bailrigg (Lancaster) field-station, September 1995 to August 1996**
- **note that an unusually cold Christmas 1995 was followed by a mild period in January-February**



Probability: three definitions

Mathematical

A branch of pure mathematics, invented by the Russian mathematician Andrey Kolmogorov (1903 – 1987) to formalise the everyday notion of uncertainty.

Frequentist

Probability as the limit of a proportion: eg genetic variation,...

Subjectivist

Probability as personal belief: eg personal decision-making

Kolmogorov's axioms

W : all possible outcomes of a study
$A, B, \ldots$: any particular outcomes
$\mathrm{P}(\cdot) =$ "the probability of"

1. $\mathrm{P}(A) \geq 0$
2. $\mathrm{P}(W) = 1$
3. if A, B are mutually exclusive, $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

And some consequences

1. **Addition law**

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

2. **Multiplication law**

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A),$$

where $\mathrm{P}(B|A)$ is the probability of B given that A has occurred.

3. **Bayes' Theorem**

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\sum_j \mathrm{P}(B|A_j)\mathrm{P}(A_j)},$$

where $A_1, A_2, \ldots$ represent all possible different outcomes

Grimmett and Stirzaker (2020)

A topical example: Covid testing

Test	Disease status D	$\bar{D}$
$+$	$p \times \mathrm{Se}$	$(1-p) \times (1-\mathrm{Sp})$
$-$	$p * (1-\mathrm{Se})$	$(1-p) \times \mathrm{Sp}$
	p	$1-p$

How likely is it that a positive test result is correct?

Lateral flow test properties:

- **Sensitivity** $\mathrm{Se} \approx 0.8$
- **Specificity** $\mathrm{Sp} \approx 0.99$

p = **prevalence of disease**

$$\mathrm{P}(D|+) = \frac{\mathrm{P}(D \cap +)}{\mathrm{P}(+)}$$

$$= \frac{p \times \mathrm{Se}}{p \times \mathrm{Se} + (1-p) \times (1-\mathrm{Sp})}$$

$$\mathrm{Se} = 0.8, \mathrm{Sp} = 0.99, p = 0.01 \Rightarrow \mathrm{P}(D|+) = 0.447$$

Review: what have we learnt?

1. **Variation** in the results of an investigation can be of two kinds:
   1. **systematic variation** arises as a direct result of known factors – typically a change in the inputs to the investigation
   2. **stochastic variation** arises when all known factors cannot explain the observed variation
2. The mathematical theory of **probability** formalises and quantifies the intuitive idea of **uncertainty**
3. **Stochastic** is not necessarily **completely random**: eg coin-tossing vs weather forecasting

Design

Asthma

- What is it?
- How do you treat it?
- How do you know if the treatment works?

A Simple Comparative Trial

- **Formoterol (F) and Salbutamol (S) are two drugs used to treat chronic asthma.**
- **trial was carried out to compare efficacy of F and S:**
  - **asthmatic children recruited;**
  - **Peak Expiratory Flow (PEF) measured after administration of F or S**

PEF is the *(primary) outcome*, F and S are the treatments of interest, time is a potential confounder.

Graff-Lonnevig and Browaldh (1990). Twelve hours bronchodilating effect of inhaled Formoterol in children with asthma: a double-blind cross-over study versus Salbutamol. *Clinical and Experimental Allergy*, **20, 429–432.**

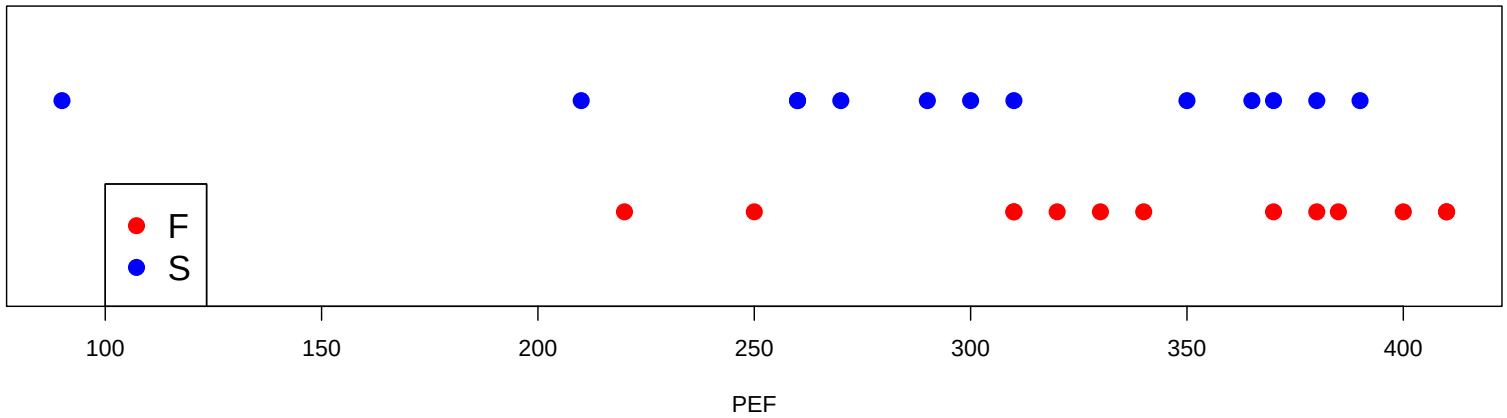
Data from the Asthma Trial

Formoterol (sorted)

220 250 310 310 320 330 340 370 380 385 400 410 410

Salbutamol (sorted)

90 210 260 260 270 290 300 310 350 365 370 380 390



- one very small S result
- considerable overlap between F and S results
- comparison between F and S inconclusive?

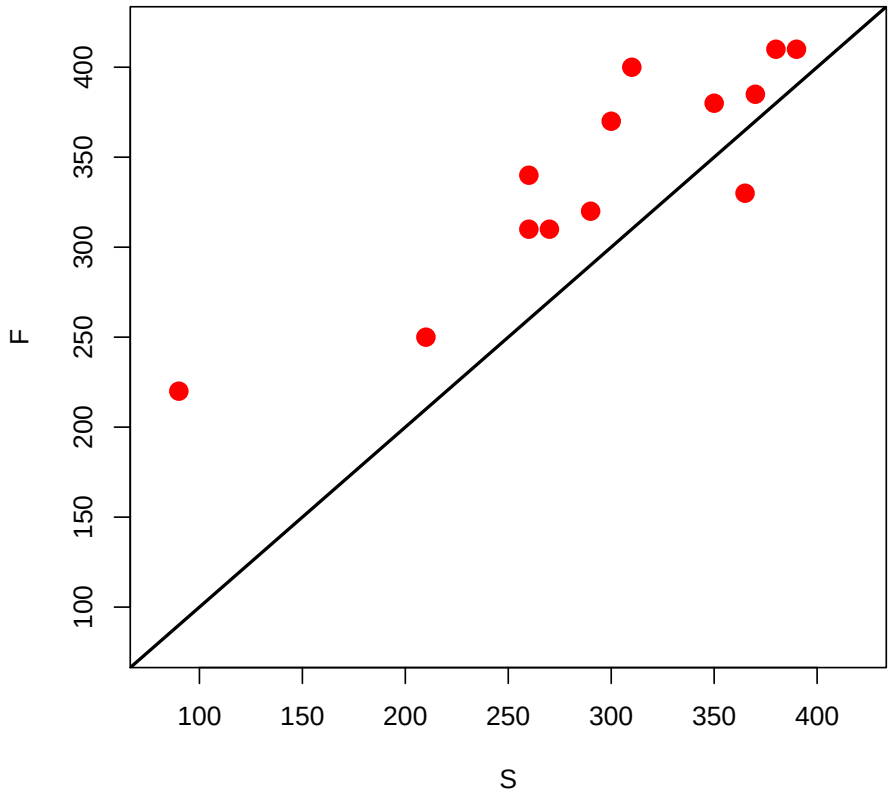
Design-based analysis

- **Difference between average PEF under F and under S is 45.4 in favour of F**
- **Hypothesise that both treatments are equally effective**
  - **what would the data have looked like?**
  - **and what do they actually look like?**

Data from the Asthma Trial: continued

Child	Drug F	Drug S	Child	Drug F	Drug S
1	310	270	8	385	370
2	310	260	9	400	310
3	370	300	10	410	380
4	410	390	11	320	290
5	250	210	12	340	260
6	380	350	13	220	90
7	330	365			

A Picture Paints a Thousand Words ...



F and S results are (positively) correlated... why?

The full story

- **Formoterol (F) and Salbutamol (S) are two drugs used to treat chronic asthma.**
- **trial was carried out to compare efficacy of F and S:**
  - **13 asthmatic children recruited;**
  - **Peak Expiratory Flow (PEF) measured after administration of F;**
  - **PEF measured after administration of S;**
  - **one week between two measurements (ordering chosen at random)**

PEF is the *(primary) outcome*, F and S are the treatments of interest, time is a (potential) confounder.

Design-based analysis

40 50 70 20 40 30 −35 15 90 30 30 80 130

- **Difference between average PEF under F and under S is 45.4 in favour of F**
- **Hypothesise that both treatments are equally effective**
  - **what would the data have looked like?**
  - **and what do they actually look like?**

What does this example tell us?

Study design is important.

Decide what comparisons are of interest

- **F vs S**

Design the experiment to eliminate extraneous sources of variation

- **pairing eliminates variation between children**
- **randomization eliminates (on average) the time effect**

Going beyond hypothesis tests

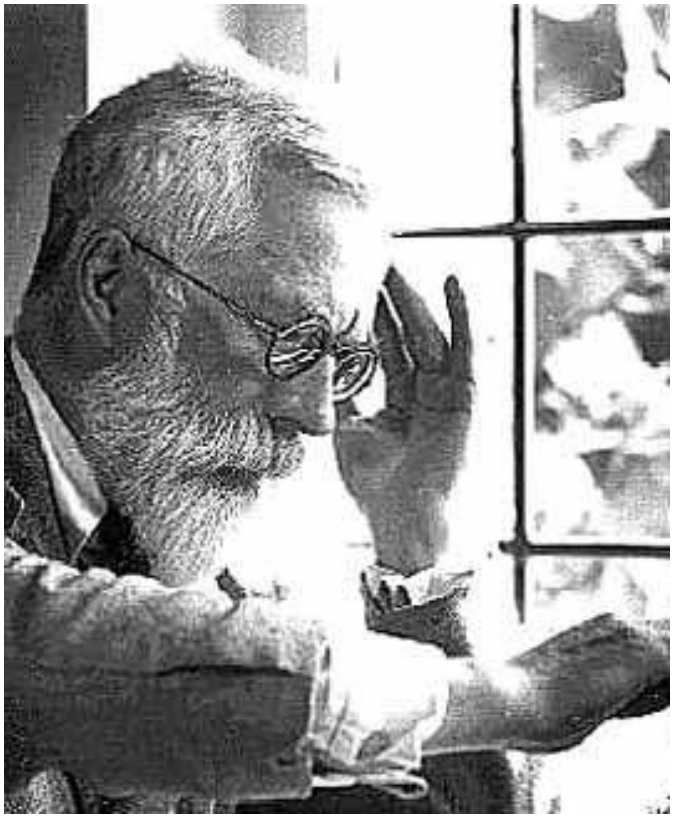
- looks like F beats S
- but by enough to be useful? 45.4litres/min? or 45.4± ?

$n = 13$	how much data do we have?
$\bar{d} = 45.4$	how big is the average experimental effect?
$SD = 40.6$	how variable is it?
$SE = SD/\sqrt{n} = 11.3$	how precisely have we estimated it?

A convenient rule of thumb:

$$\bar{d} \pm 2 \times SE$$

Two Statistical Pioneers



Sir Ronald Aylmer Fisher (1890–1962): statistician and geneticist

Sir Austin Bradford Hill (1897–1991): medical statistician

Principles of experimental design: blocking

A block is a group of experimental units which are thought to be relatively homogeneous, i.e. which will give relatively similar results.

Examples:

- **paired PEF values**
- **siblings**
- **communities**

Blocking

Whenever possible, design your experiment so that comparisons of interest can be made within the same block.

Principles of experimental design: randomisation

How should you choose which units receive which experimental treatment?

Randomisation:

Within any block, allocate treatments to experimental units at random;

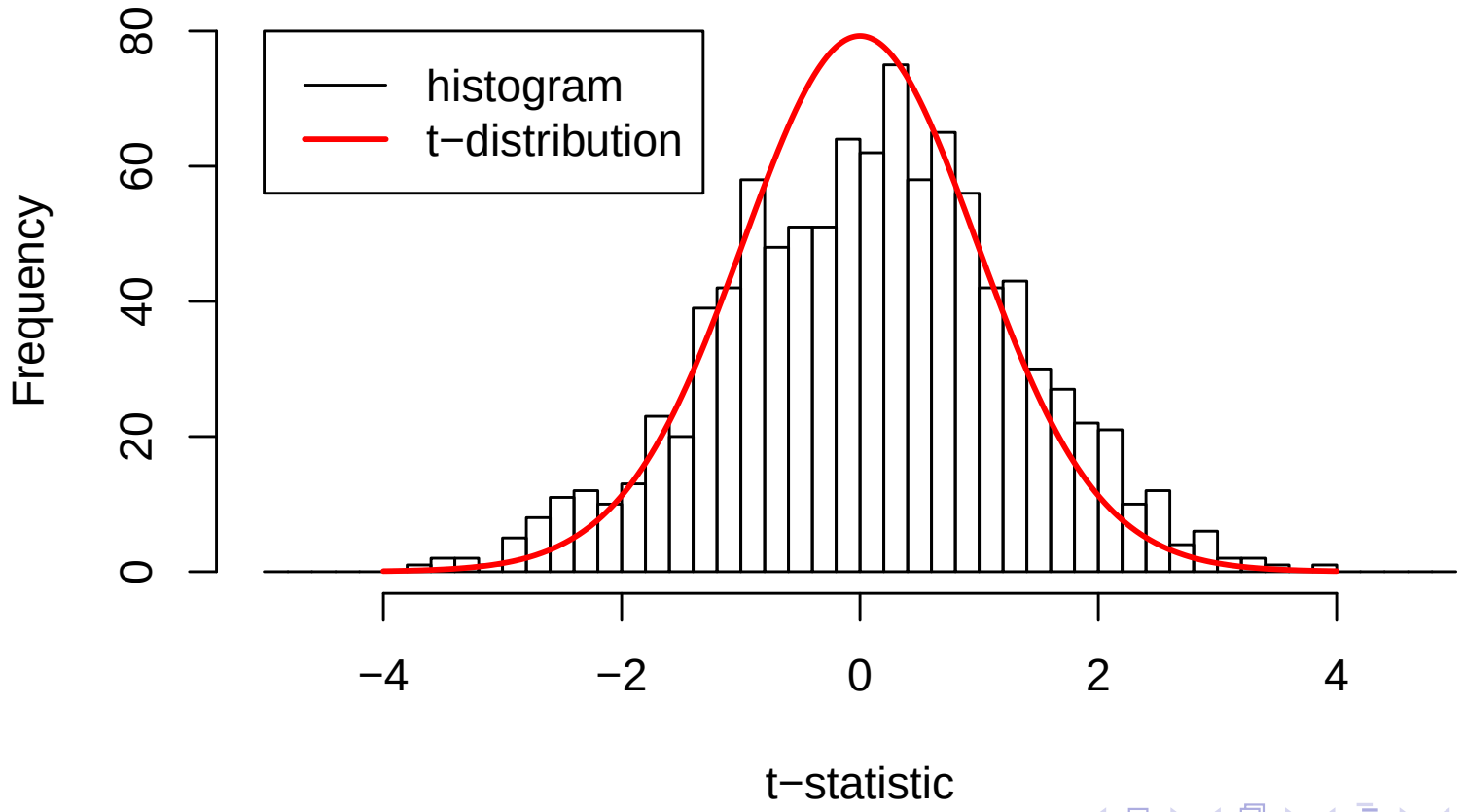
- **to avoid conscious or unconscious bias in the allocation of experimental treatments to experimental units;**
- **and (sometimes – but more often than you might think) to ensure validity of statistical inferences**

Box, Hunter and Hunter (1978)

Don't do t-tests on non-Normal data?

$$t = \bar{d}/SE = 45.4/11.2 = 4.05 \quad \mathrm{P}(|t_{12}| > 4.05 = 0.0016$$

Design-based sampling distribution of the *t*-statistic



Identifying sources of variation

Before you run an experiment:

- **list the possible sources of variation in the results;**
- **design the experiment to eliminate extraneous sources of variation from comparisons of interest;**
- **if it is not possible to eliminate an extraneous source of variation, use random allocation to avoid bias**

Example: in the paired experiment

- **pairing eliminated extraneous variation between children;**
- **crossover eliminated extraneous variation between time-periods**
- **randomisation eliminated any possible residual bias**

Designing an agricultural field trial

The research farm at Rothamsted, Hertfordshire, UK.

Background

Agricultural field trials are conducted to compare the yields of different varieties of crop-plants under realistic conditions

Experimental units are contiguous plots of land, typically long, narrow strips within a square or rectangular field.

Design questions include:

- **how to orient the strips (eg North-South or East-West)?**
- **which treatments to apply to which strips?**

Example

Aim: compare yields of four varieties of wheat

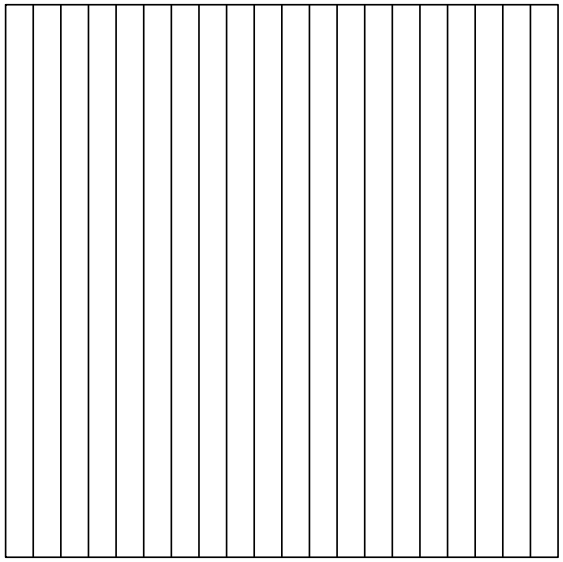
Experimental material:

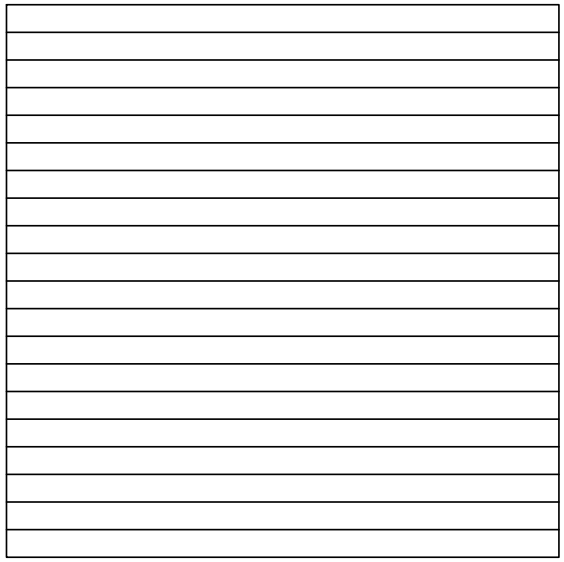
- experiment to be run on a 100 metre by 100 metre square field.
- experimental units to be 100 metre by 5 metre strips, hence 20 units in all.

Context: there is a suspected North-South fertility gradient over the field.

Two possible layouts of the experiment

North





South

Points for discussion

1. What statistic(s) might we use to compare the yields of the four varieties of wheat?

2. Would you prefer to have the strips running North-South or East-West?

3. How would you allocate varieties, A, B, C, D say, amongst the 20 strips?

4. How might you alter your design if the precise objective was to find which varieties give the largest yields:

- **on fertile ground?**
- **and on infertile ground?**

Review: what have we learnt?

1. **Careful design of an investigation can:**
   1. **simplify the analysis of the resulting data**
   2. **enable more precise conclusions to be drawn**
2. **Two fundamental principles of good statistical design are:**
   1. **blocking: minimise the adverse effects of known sources of extraneous variation by comparing like-with-like**
   2. **randomisation: minimise the adverse effects of unknown sources of extraneous variation and/or subconscious bias on the investigator's part by random allocation of experimental units to treatments**
3. **Randomisation also allows simple questions to be answered without having to assume that the data follow a particular probability distribution.**

Inference

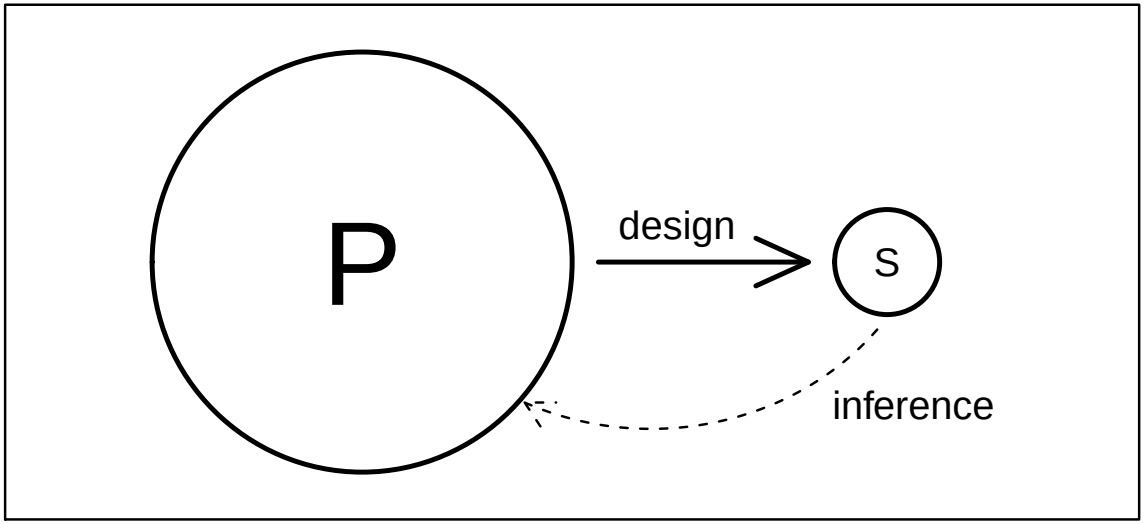
Key ideas in statistical inference

Populations and samples

- **population**: the set of (biological, environmental,...) material to which you hope your work is relevant
- **sample**: the (much smaller) set of material on which you actually base your work

Design and inference

- **design**: how you choose your sample
- **inference**: how you use results from your sample to reach conclusions about the population



How to generate your sample:

1. **random sampling from a finite population**
   Population size N, sample size n, each member of the population has a **known** probability, p_i say, of being included in the sample
   1. **simple** random sampling: all p_i equal
   2. **stratified** random sampling: all p_i equal within pre-defined sub-populations
2. **quasi-random sampling from an infinite population**
   This is the implicit assumption in most lab-based studies
3. **convenience sampling**
   Use whatever is closest to hand...not recommended

Inferential goals

Parameter estimation

- **parameter**: an unknown constant whose value is scientifically interesting
- **point estimate**: a best guess at the true value
- **interval estimate**: a range of values which is, in some sense, likely to include the true value of a parameter (again, informed by your data)

Hypothesis testing

- **hypothesis**: a statement about a parameter
- **statistical test**: a way of assessing whether your data are reasonably consistent with a pre-specified hypothesis

Prediction

- a probability statement about an unobserved outcome

Interval estimation

A point estimate is of little value without some indication of its precision

Interval estimation is a compromise between:

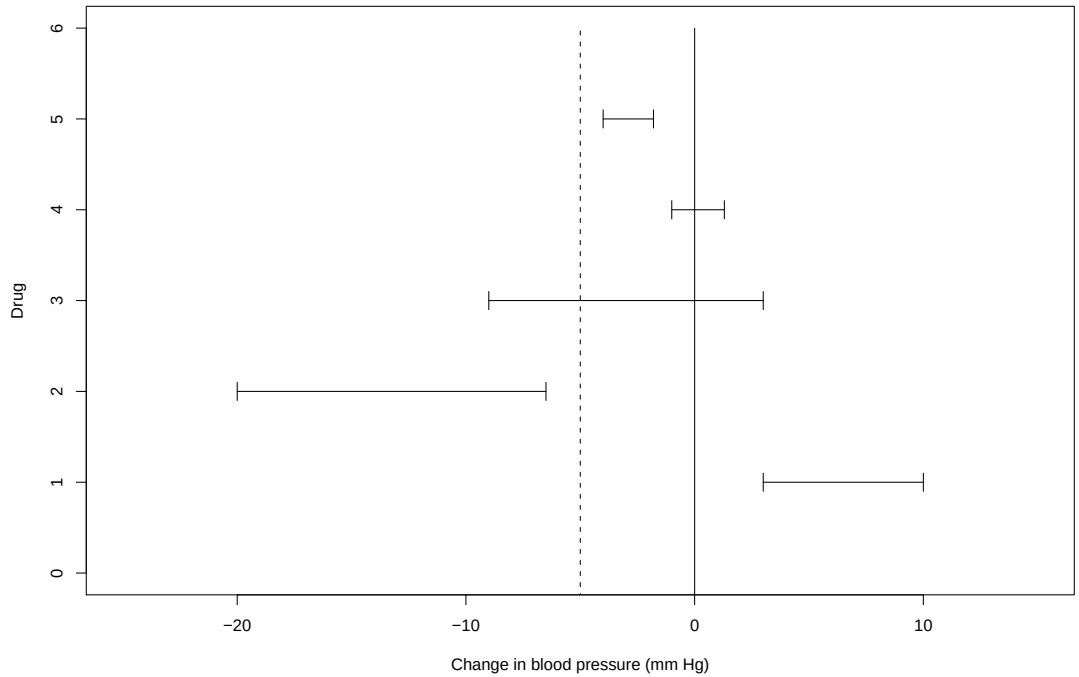
- **the width of your interval;**
- **the likelihood that your interval will include the right answer.**

A confidence interval is defined as follows:

- **choose an acceptable level of confidence, say $100p\%$ (conventionally $p = 0.95$, but it's your choice)**
- **construct an interval so that $100p\%$ of the time, it will include the true value of the parameter**

Exercise: all other things being equal, if you increase p, will the resulting confidence interval become wider or narrower?

Statistical significance is not the same as clinical/practical significance: comparing five anti-hypertensive drugs



- **which drugs give a significant reduction in blood pressure?**
- **which drugs give a useful reduction in blood pressure?**
- **which drugs need further investigation?**

Estimating a mean

Probably the most widely used formula for a confidence interval is

$$\bar{x} \pm 2\sqrt{s^2/n}$$

Data independently replicated under identical conditions

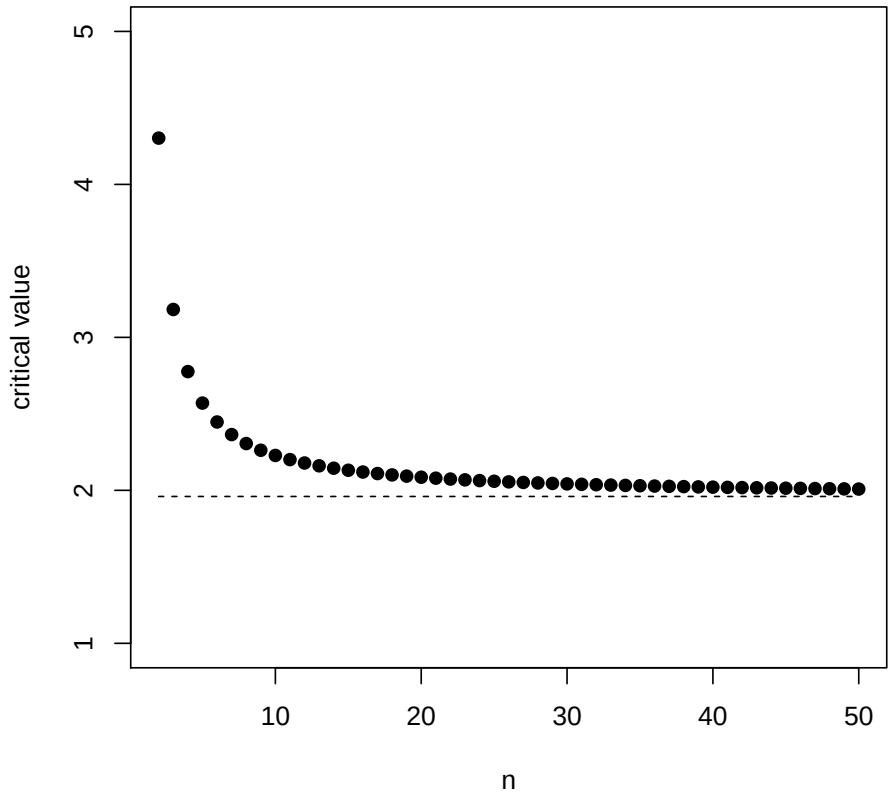
$$x_1, x_2, \ldots, x_n$$

- $\bar{x}$ **is the sample mean**
- s^2 **is the sample variance**
- n **is the sample size**
- **or replace** ± 2 **by a number depending on** n **... see next slide**

Estimating a mean (2)

Strictly, you should use $\pm c_n$, where c_n depends on n, but is approximately 2 for reasonably large n

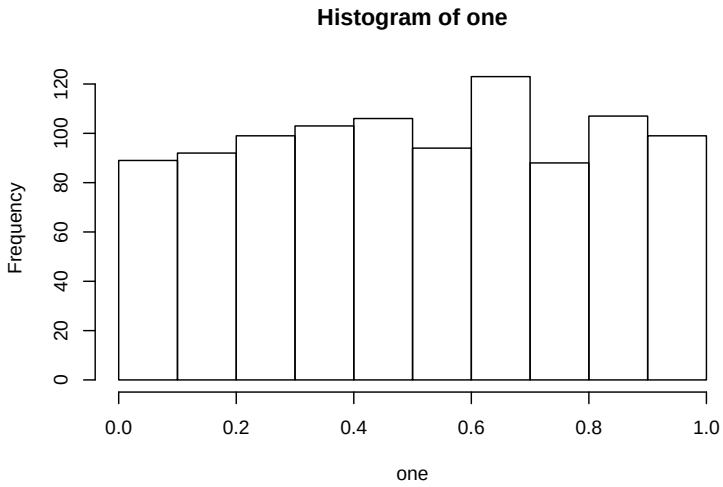
n	c_n
5	2.78
10	2.26
20	2.09
50	2.01
∞	1.96

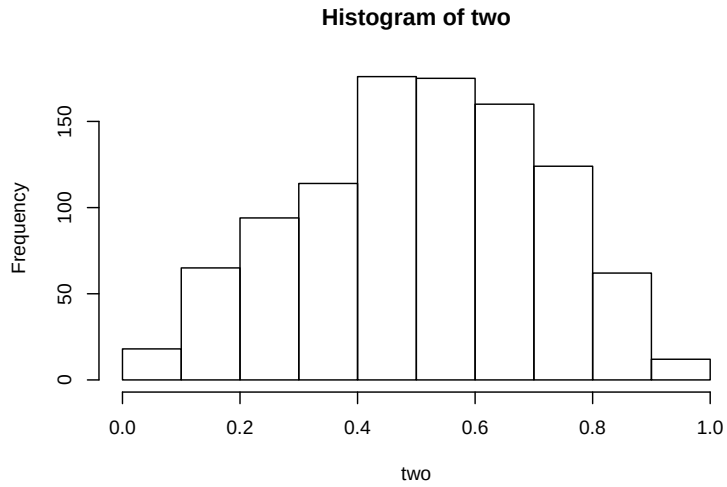


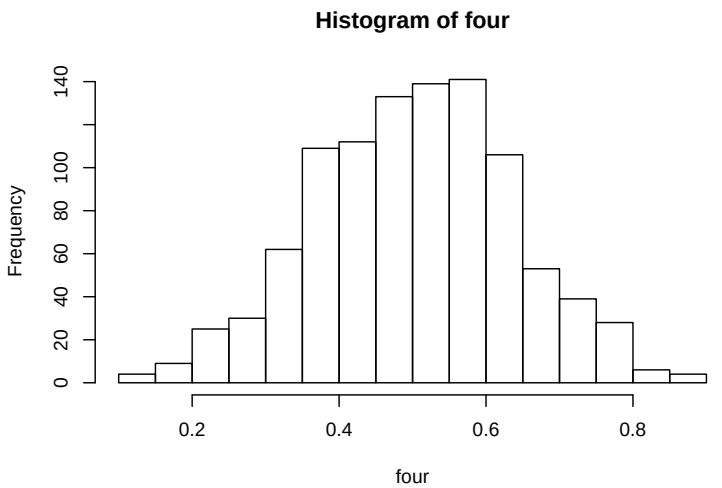
Where does the formula for the confidence interval come from?

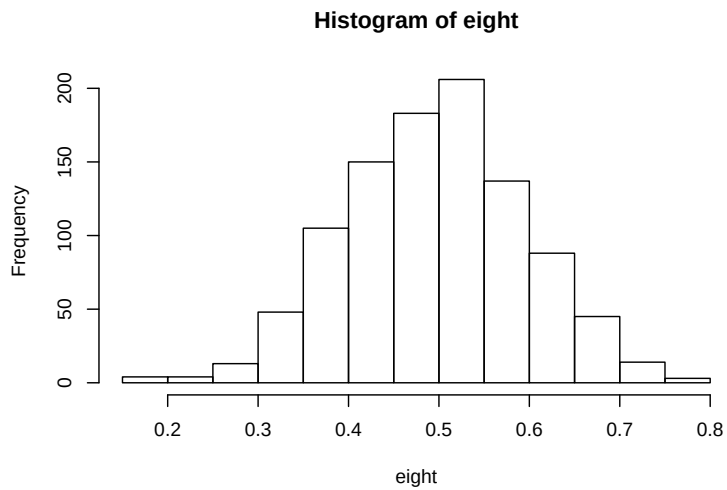
Estimating a mean (3)

The distribution of the sample mean changes with sample size, *n*:



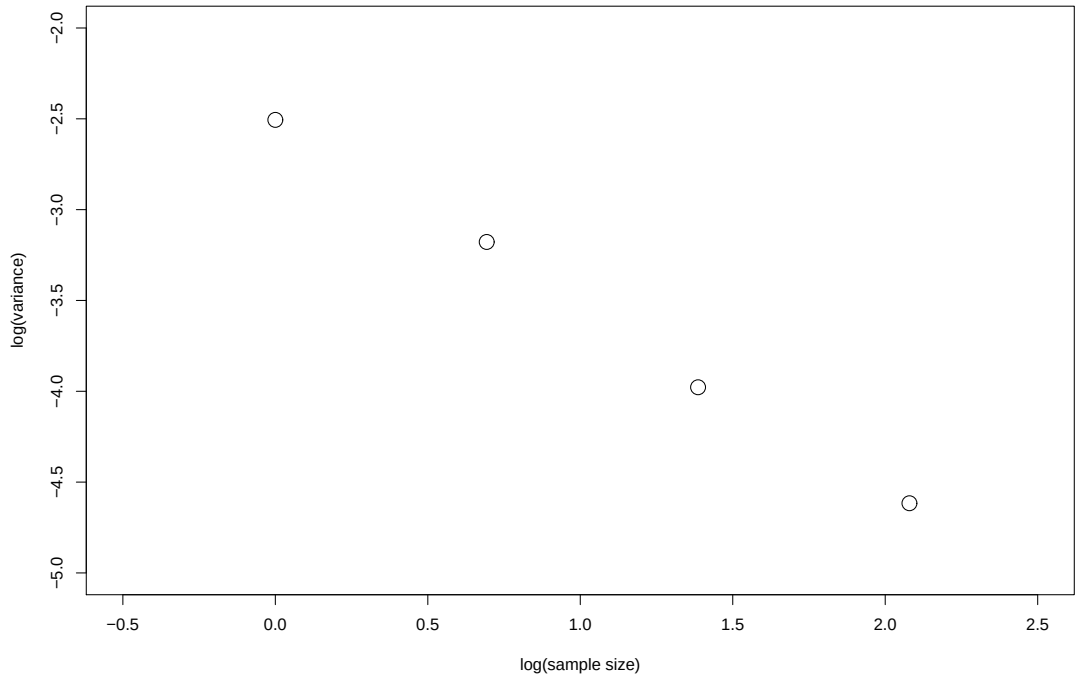






Estimating a mean (4)

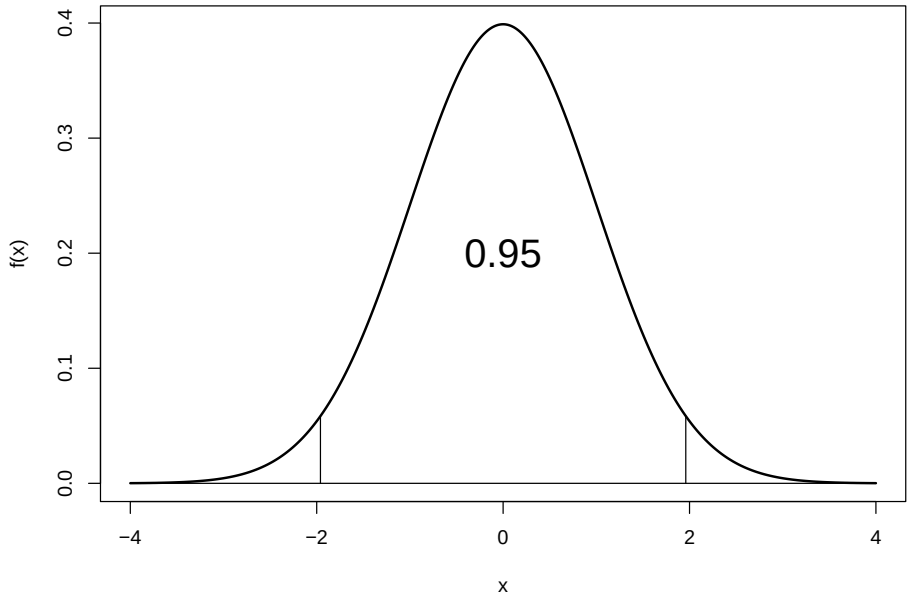
The variance, V, of the sample mean changes with sample size, n:



$$\log(V) = a - \log(n) \quad \Rightarrow \quad V = A \times n^{-1}$$
$$\Rightarrow \quad SE \propto \times n^{-0.5}$$

The Central Limit Theorem

Suppose outcomes $Y_1, \ldots, Y_n$ are an independent random sample from any distribution with mean μ and variance σ^2, and write $\bar{Y}$ for the sample mean, $\bar{Y} = \sum_{i=1}^{n} Y_i$.



Theorem In the limit, as $n \to \infty$,

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0, 1).$$

Or in words...sample means are approximately Normally distributed

Also called the Gaussian distribution, after the German mathematician Carl Friedrich Gauss (1777 – 1855).

Likelihood-based inference

A statistical model is a specification of the joint distribution of a set of random variables, $Y = (Y_1, ..., Y_n)$, indexed by a set of parameters, $\theta = (\theta_1, ..., \theta_p)$; a shorthand notation for this is $[Y; \theta]$.

The likelihood function is $L(\theta) = [y; \theta]$, where y is the observed value of Y

The maximum likelihood estimator, $\hat{\theta}$, maximises $L(\theta)$

For n large, $[\hat{\theta}] = \mathrm{MVN}(\theta, V(\theta))$, where $V(\theta) = \left[-\frac{\partial^2 L(\theta)}{\partial \theta^2}\right]^{-1}$

Bayesian estimation requires you to specify a *prior distribution*, $[\theta]$, from which you can deduce a *posterior distribution*,

$$[\theta|y] = \frac{L(\theta)[\theta]}{\int L(\theta)[\theta]d\theta}$$

Pawitan (2001), Lee (2012)

Conclusions

- sample means are approximately Normally distributed (symmetric, bell-shaped histogram)
- larger samples lead to more precise estimates
- the variance of an estimate is inversely proportional to the sample size, n
- the standard error of an estimate is therefore inversely proportional to $\sqrt{n}$
- law of diminishing returns – doubling the sample size does not double the precision of your estimate
- likelihood-based methods provide a generally applicable, efficient and principled approach to model-based inference

Review: what have we learnt?

1. **How you choose your sample is important**
2. **Why statistical significance is not the same thing as clinical/practical significance**
3. **How to estimate a mean from an independently replicated sample**
4. **The $\sqrt{n}$ law of statistical precision**
5. **The Central Limit Theorem...sample means are approximately Normally distributed**

Modelling

Statistical modelling principles

- models are devices to answer questions, and should:
  - be not demonstrably inconsistent with the data;
  - incorporate the underlying science, where this is well understood
  - be as simple as possible, within the above constraints

"Too many notes, Mozart"

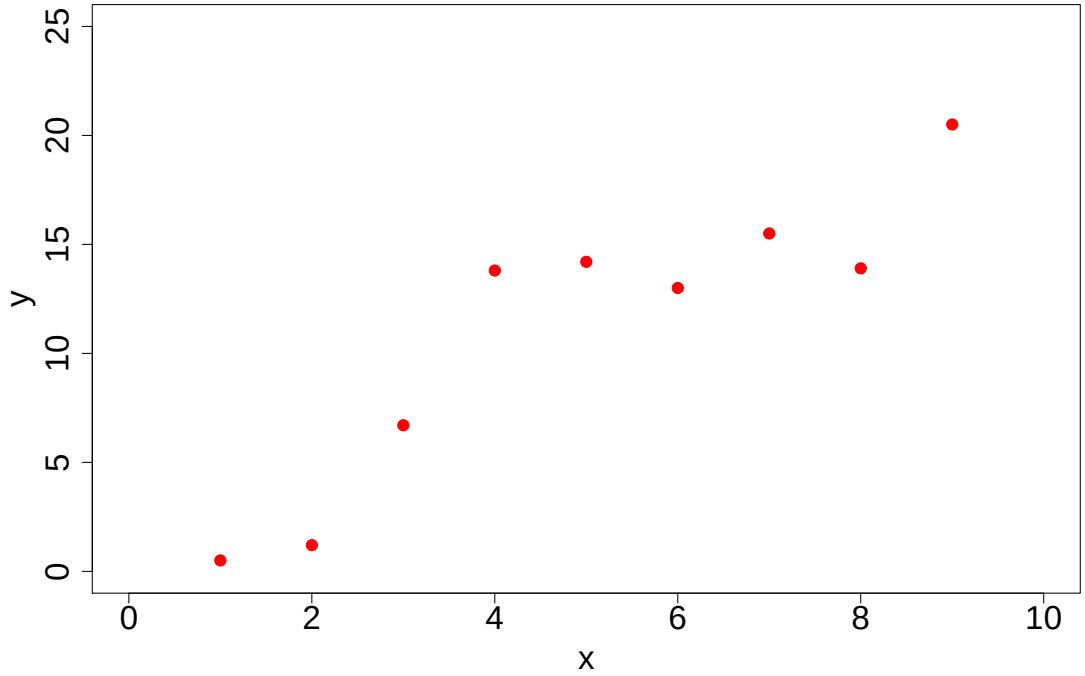
Emperor Joseph II

"Only as many as there needed to be"

Mozart (apochryphal?)

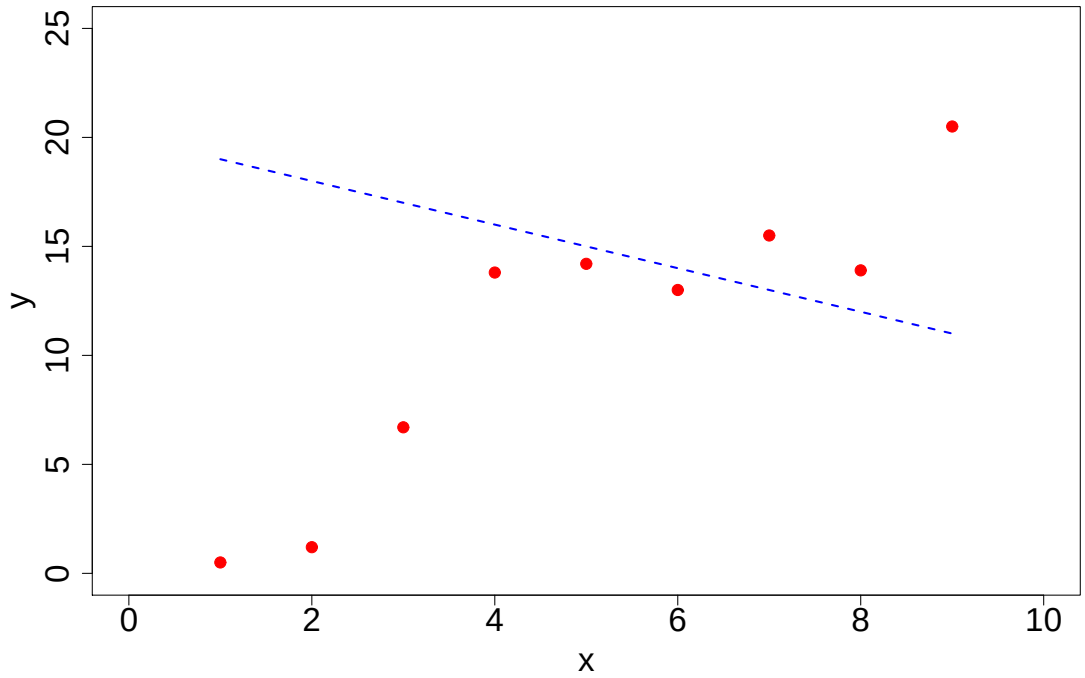
Empirical or mechanistic approach to model-building?

An empirical model: simple linear regression



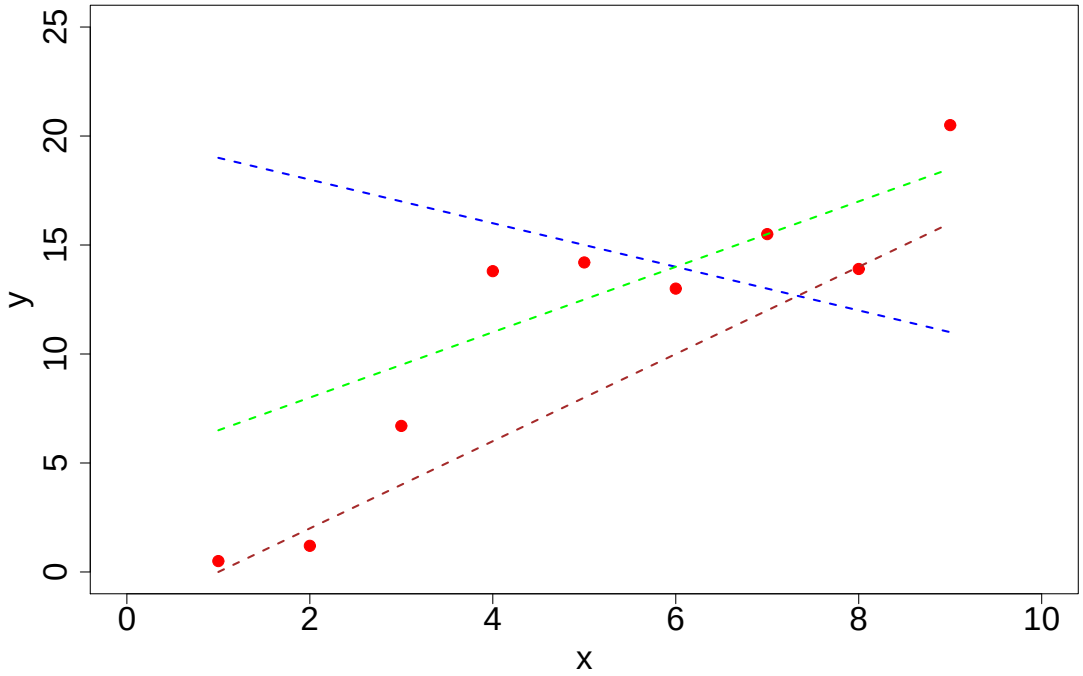
- **data are pairs of values x and y (input and output)**
-
-

An empirical model: simple linear regression



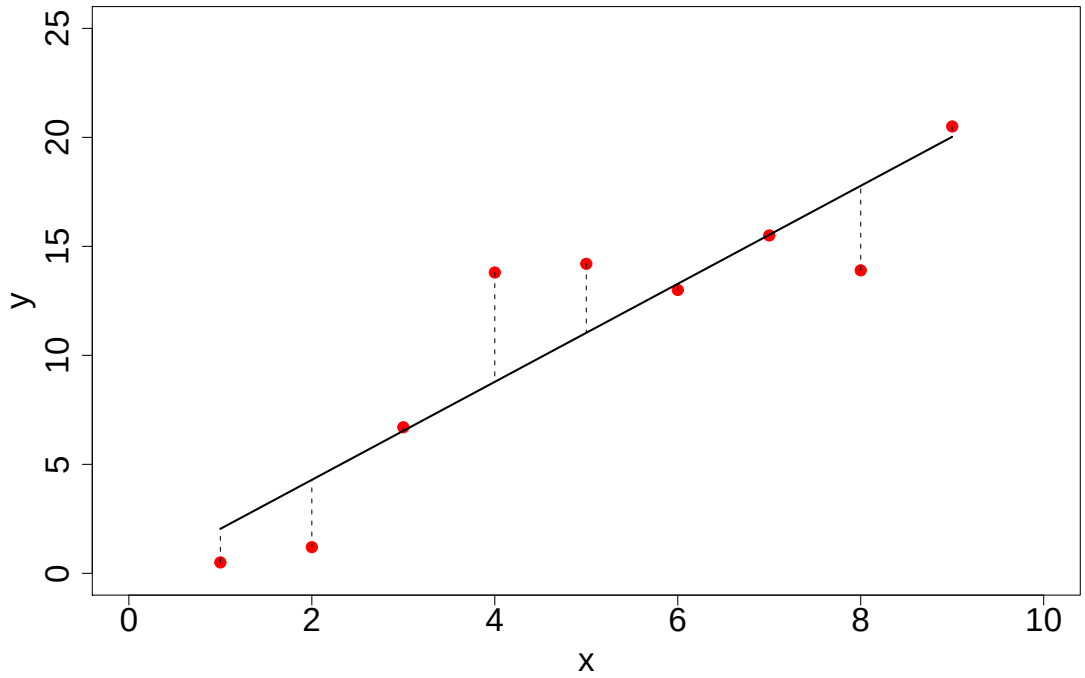
- **data are pairs of values x and y (input and output)**
- **the best-fitting line?**
-

An empirical model: simple linear regression



- **data are pairs of values x and y (input and output)**
- **some lines are "obviously" better fits to the data than others, but which is the "'best-fitting" line?**
-

An empirical model: simple linear regression



- **data are pairs of values x and y (input and output)**
- **dashed vertical lines are residuals**
- **the best-fitting line makes the sum of squared residuals as small as possible**

Exercise: **why are residuals measured vertically?**

Plant growth and pollution

- **Glyphosate is a powerful weed-killer, its presence in the water-supply is potentially harmful to irrigated crops.**
- **Experiment conducted to investigate how average root-length of batches of 15 safflower plants is affected by glyphosate added to distilled or tap water**

Glyphosate data

x (ppm)	0.000	0.000	0.053
y (distilled)	107.0	110.9	106.2
y (tap)	111.0	168.3	105.7
x (ppm)	0.106	0.211	0.423
y (distilled)	97.3	105.9	88.5
y (tap)	116.7	143.7	84.7
x (ppm)	0.845	1.609	3.380
y (distilled)	74.4	46.2	30.0
y (tap)	59.3	36.7	38.0

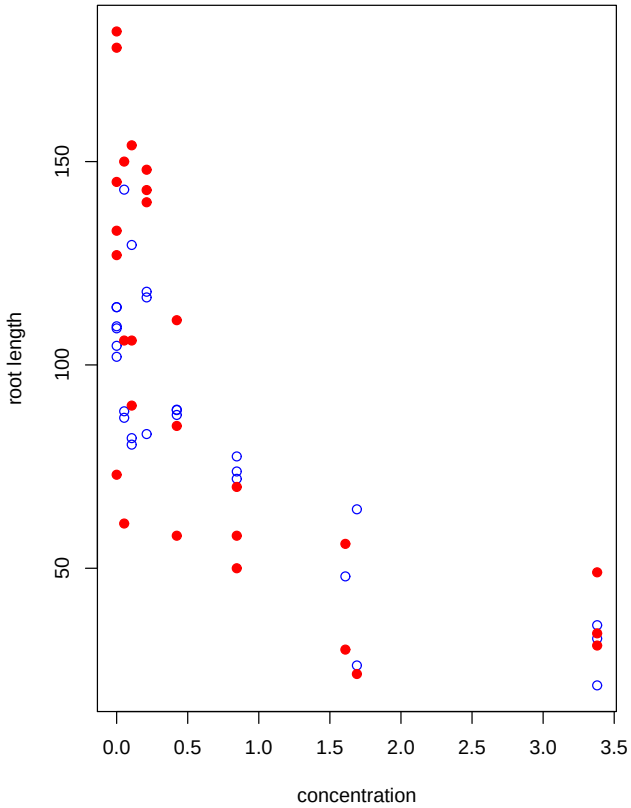
Scientific and statistical objectives

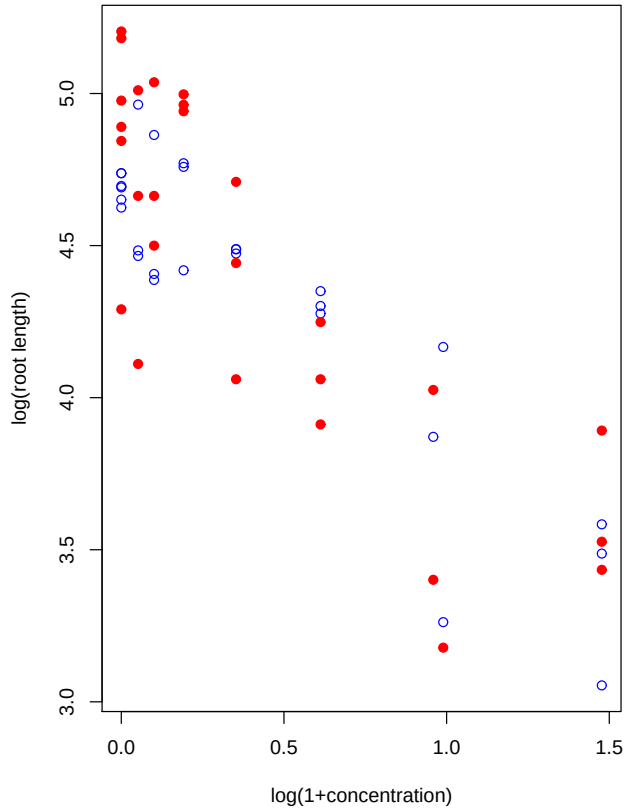
- what can these data tell us about the effect of small concentrations of glyphosate on plant growth?
- how could we build a statistical model to describe the relationship between glyphosate concentration and root-length?

Where do models come from?

- in the gravity experiment, the linear regression model had a mechanistic justification (physics plus physiology)
- In the glyphosate experiment, there is no scientific law to guide us.
- but we may still be able to use the linear regression model to describe the empirical relationship between glyphosate concentration and root-length.

Plotting the glyphosate data





Notes on data-transformations

- other transformations of the data could have been used
- choice could be determined by empirical and/or scientific considerations

Example.

Suppose x and y follow a power law model,

$$y = ax^b$$

Then, log-log transformation produces a linear model,

$$Y = \alpha + \beta X$$

where $Y = \log y$, $X = \log x$, $\alpha = \log a$ and $\beta = b$.

A linear model for the glyphosate data

$y = \log(\text{root length})$ (response)

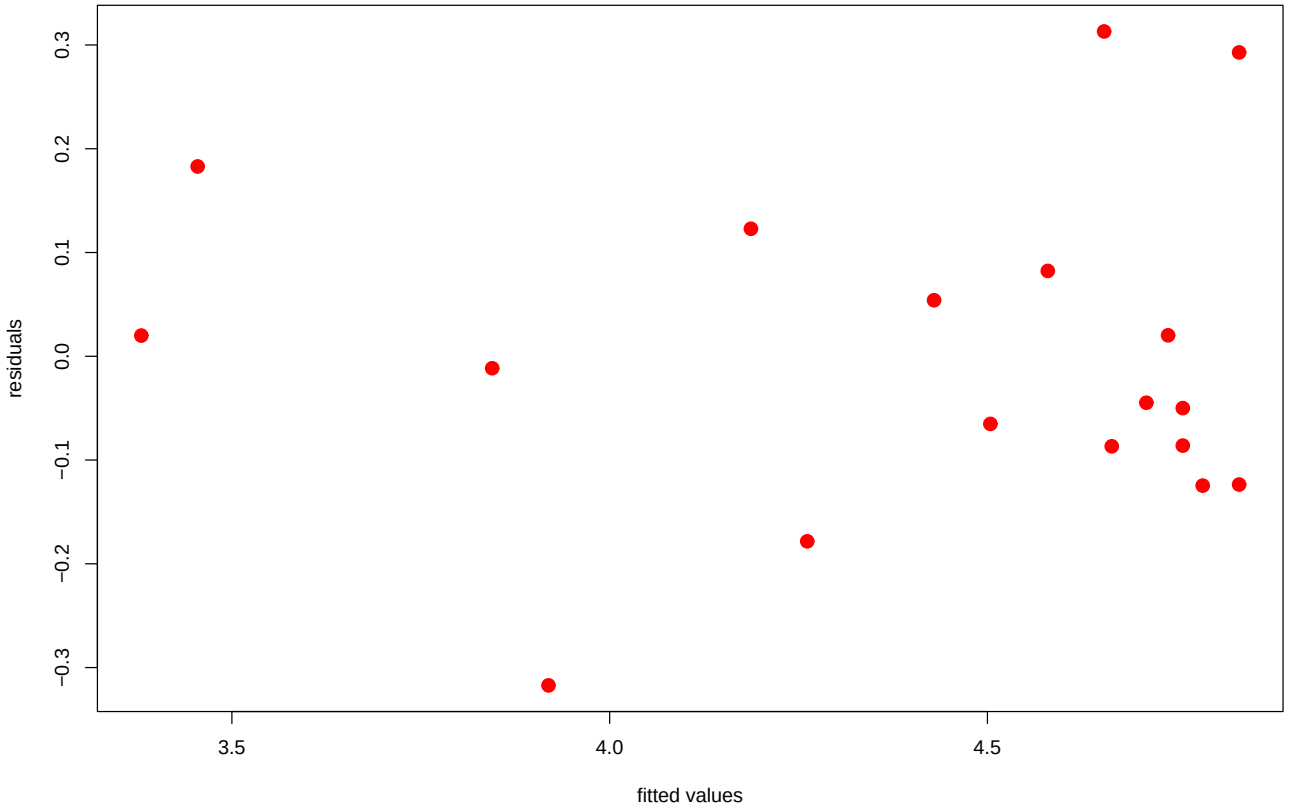
$x = \log(1 + \text{glyphosate})$ (explanatory variable)

$w = 0/1 = \text{distilled/tap water}$ (factor)

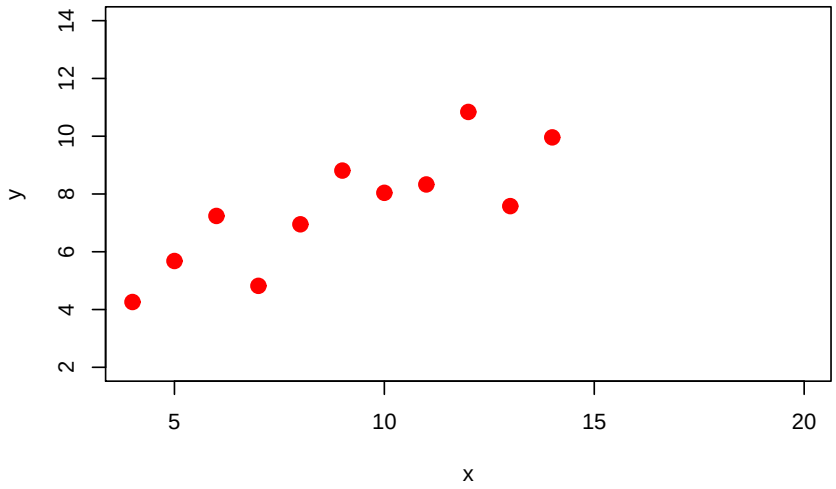
$$y = \{\alpha_0 + \alpha_1 w\} + \beta x + z$$

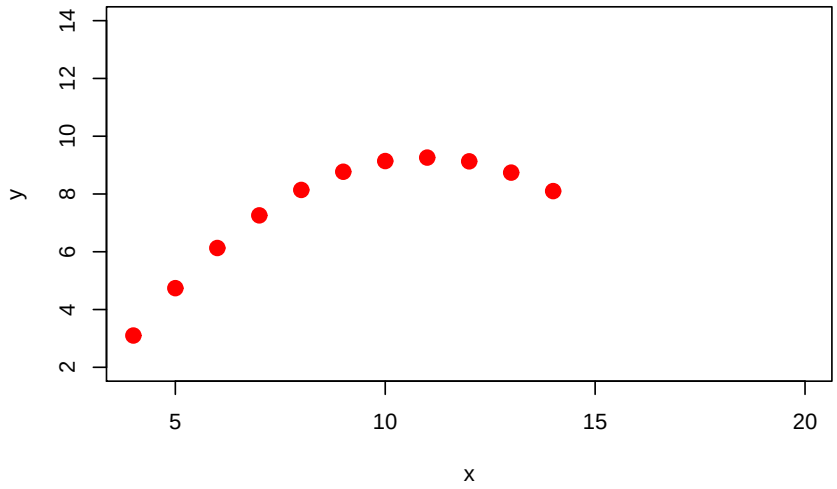
- parallel straight-line relationships for distilled and for tap-water
- $\alpha_1 = 0$ if source of water does not affect average root-length
- β measures effect of glyphosate on plant-growth (on transformed scale)

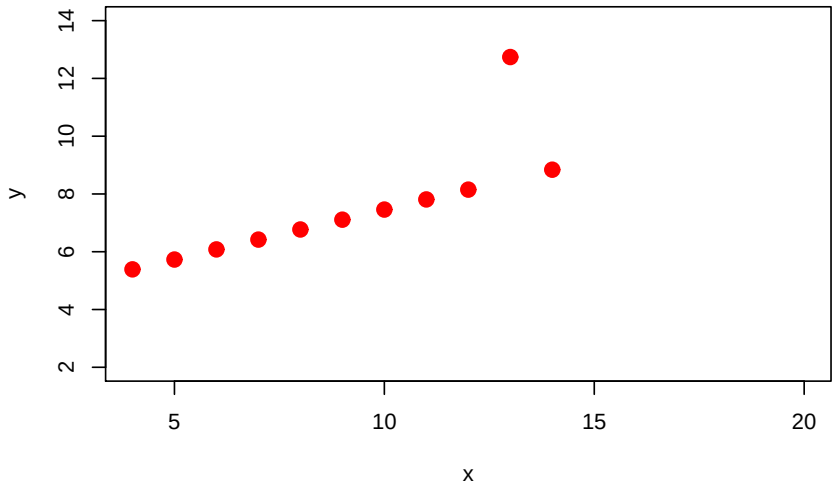
Residuals vs fitted values plot

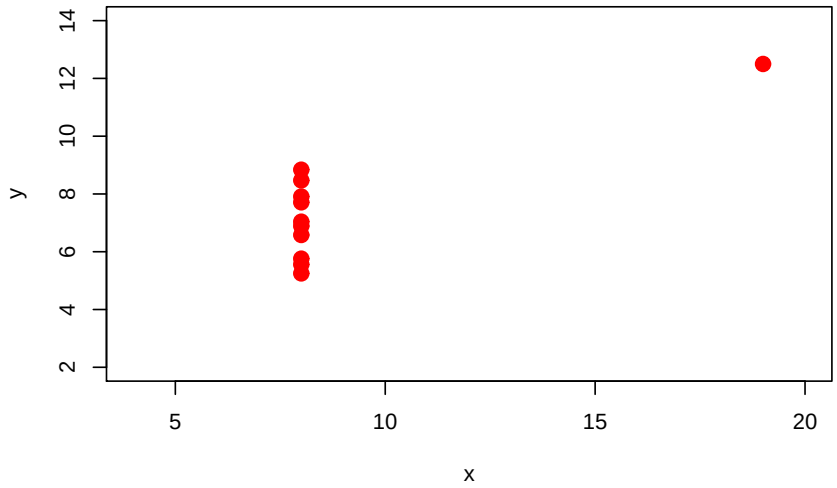


Anscombe's quartet

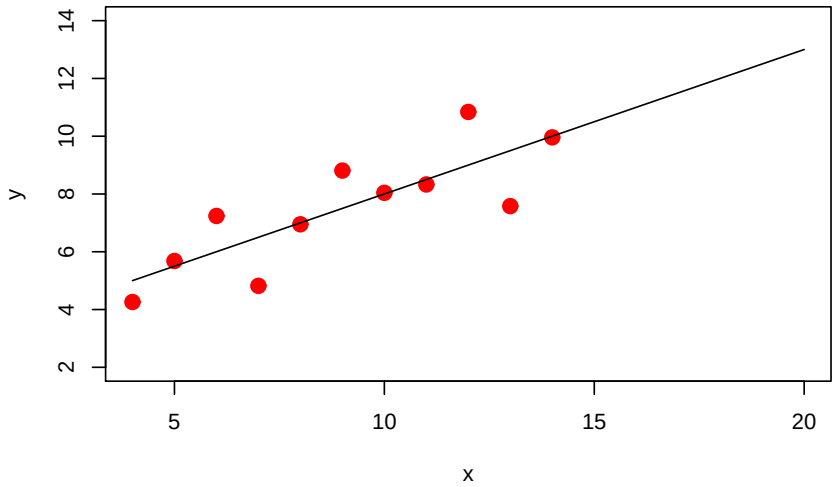


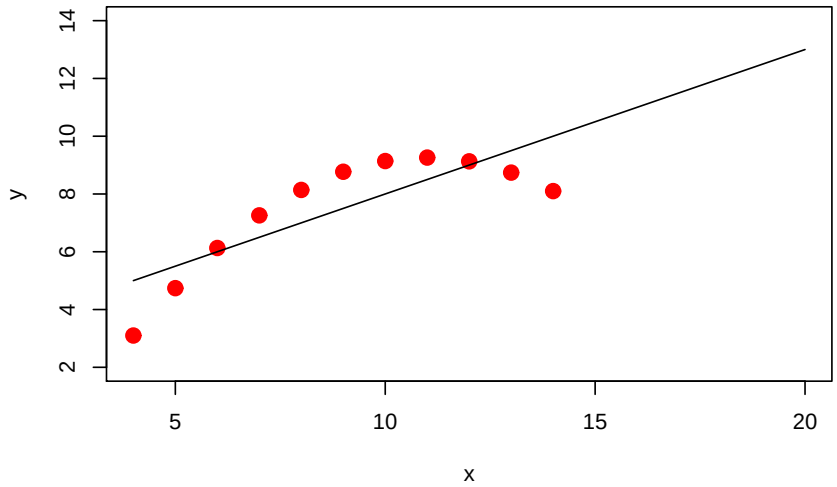


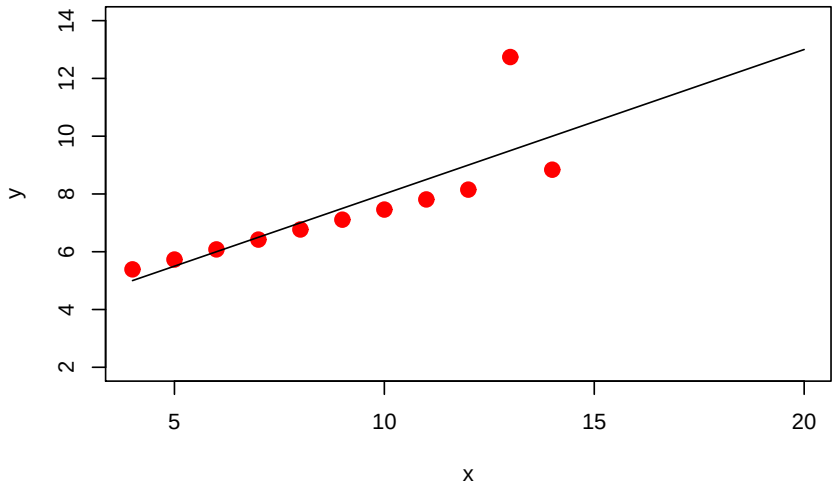


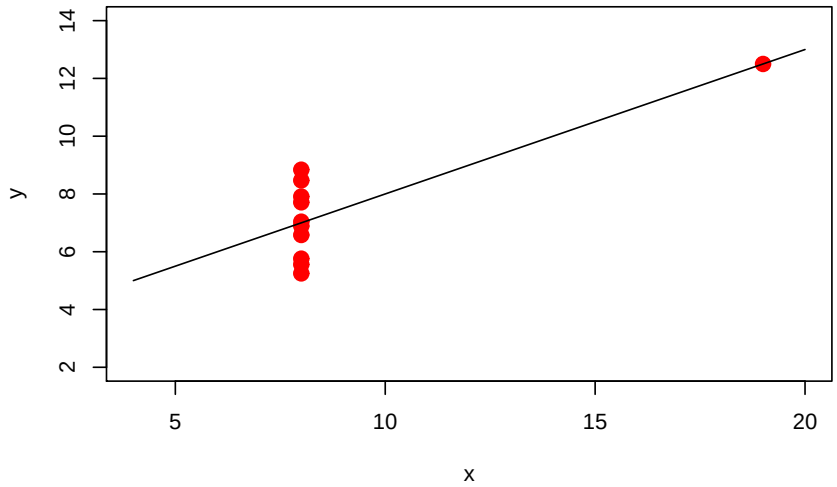


Residuals: Anscombe's quartet









Residuals

DATA = FITTED VALUE + RESIDUAL

Anscombe's quartet

All four data-sets have:

- **the same best-fitting straight line, $y = \hat{\alpha} + \hat{\beta}x$**
- **the same standard errors for $\hat{\alpha}$ and $\hat{\beta}$**
- **the same residual sum of squares**

R code for linear regression

```
Call:
lm(formula = y1   x123)

Coefficients:
                Estimate      SE   t value
 (Intercept)     3.0001   1.1247     2.667
 x123            0.5001   0.1179     4.241

Residual standard error:  1.237 on 9
degrees of freedom
```

But the residuals themselves tell four different stories.

	res1	res2	res3	res4
1	-0.740	-1.901	0.389	0.000
2	0.179	-0.761	0.229	-0.111
3	1.239	0.129	0.079	-1.751
4	-1.681	0.759	-0.081	0.909
5	-0.051	1.139	-0.230	-1.241
6	1.309	1.269	-0.390	1.839
7	0.039	1.139	-0.540	-0.421
8	-0.171	0.759	-0.689	1.469
9	1.839	0.129	-0.849	-1.441
10	-1.921	-0.761	3.241	0.709
11	-0.041	-1.901	-1.159	0.039

Why are the four sets of residuals different?

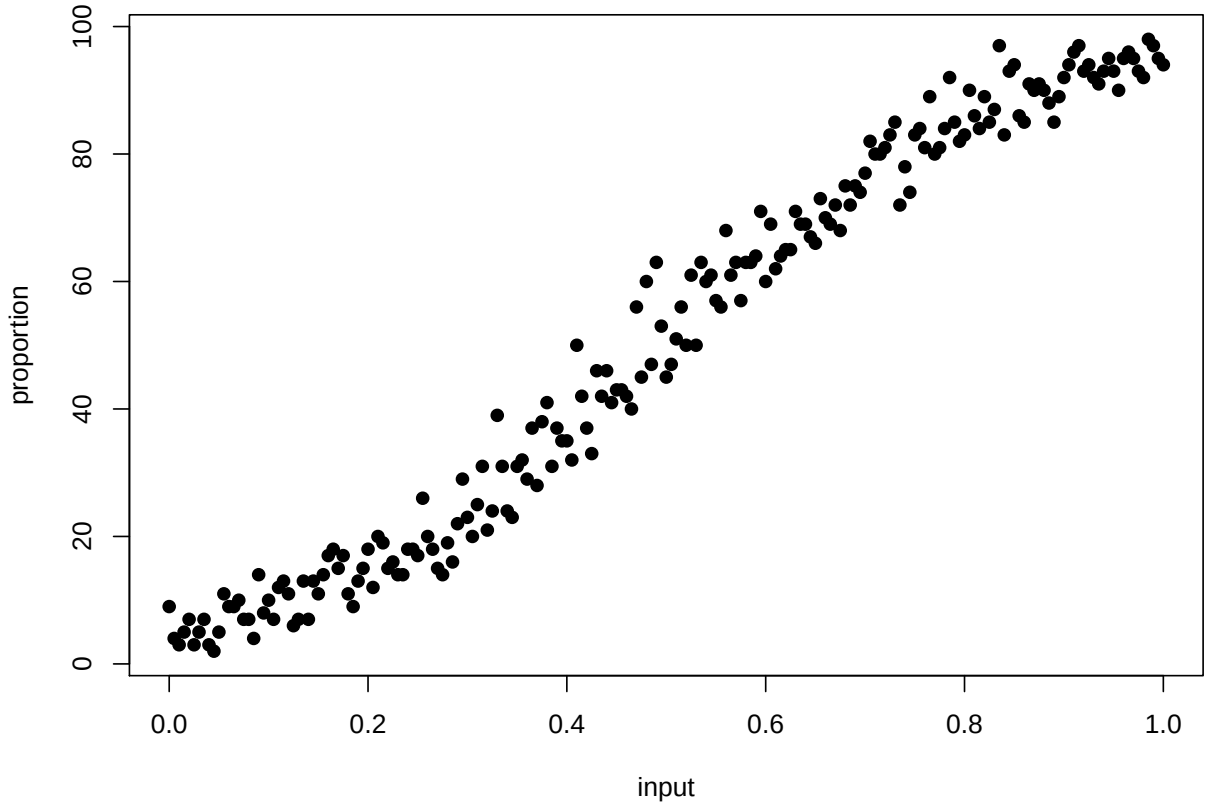
- obvious if you have only one explanatory variable
- less obvious when you have many

Analysing residuals

- check that their average value is (close to) zero
- plot them against fitted values
- plot them against explanatory variables in the model
- plot them against explanatory variables not in the model (for example, residuals against time-order)

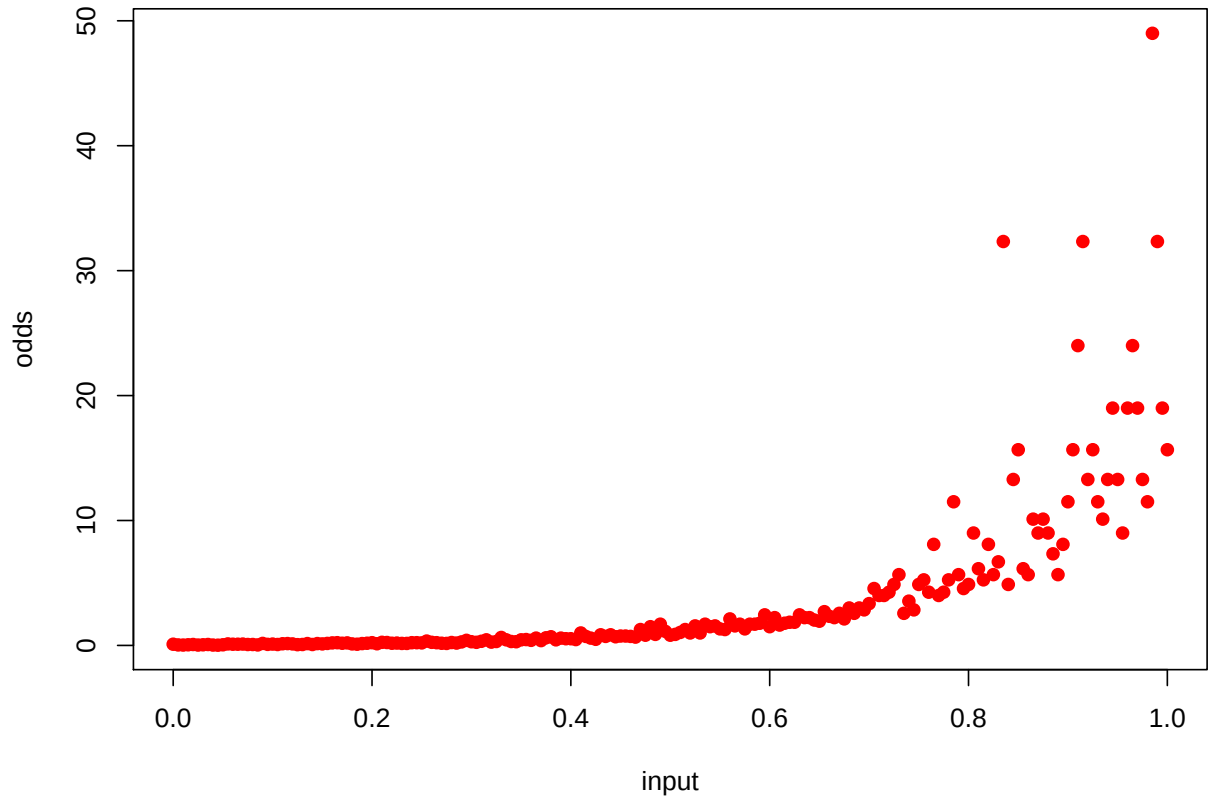
What if our response is a proportion?

- **a linear regression may fit over a restricted range of the input variable**
- **but it usually fails when the observed proportions cover most of the range from zero to one.**



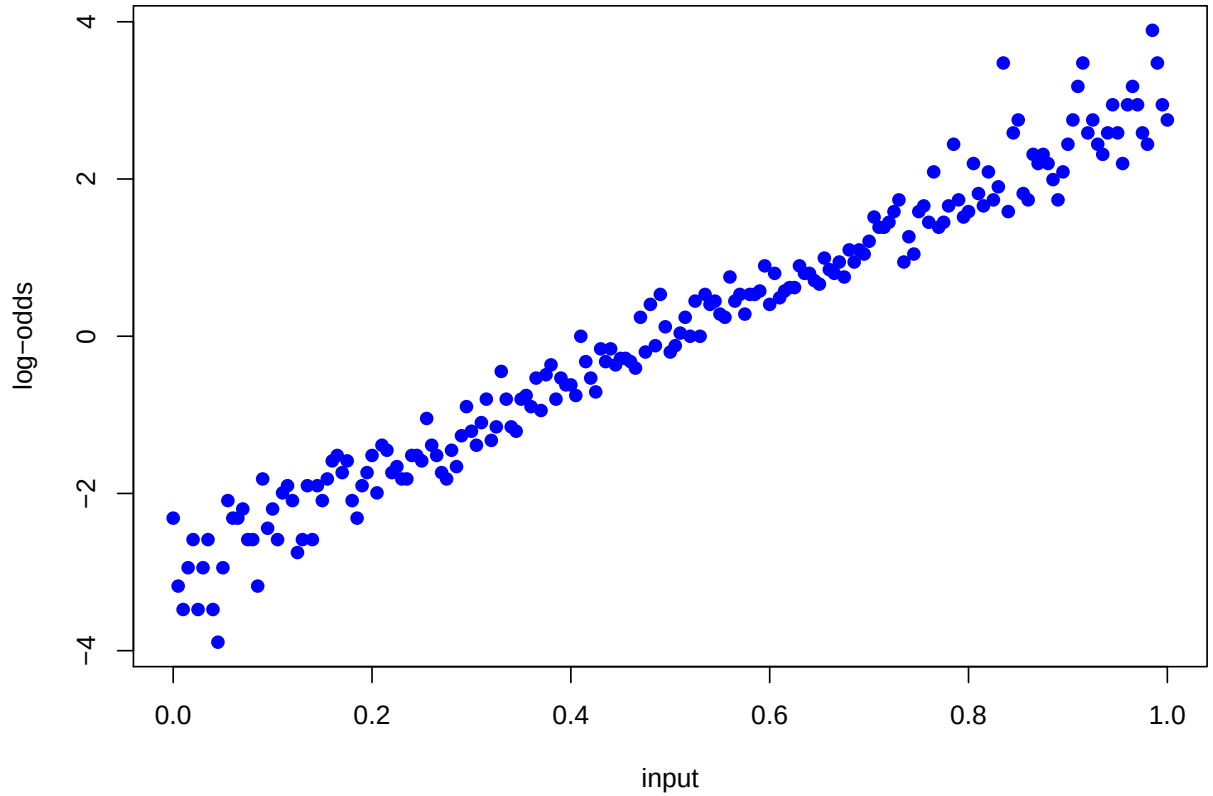
What if our response is a proportion?

- **a linear regression may fit over a restricted range of the input variable**
- **but it usually fails when the observed proportions cover most of the range from zero to one.**



What if our response is a proportion?

- **a linear regression may fit over a restricted range of the input variable**
- **but it usually fails when the observed proportions cover most of the range from zero to one.**



Generalising the linear model

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

Or equivalently:

- $\mu = \mathrm{E}[Y] = \alpha + \beta x$
- $Y \sim \mathrm{N}(\mu, \sigma^2)$

The generalisation:

- $h(\mu) = \alpha + \beta x$ **(link function)**
- $Y \sim f(\mu, \ldots)$ **(error distribution)**

Generalised linear models

- $h(\mu) = \alpha + \beta x$ (Link function)
- $Y \sim f(\mu, ...)$ (error distribution)
- choice of link function makes linear dependence on explanatory variables less restrictive
- choice of error distribution adds flexibility
- puts a very wide range of statistical methods under a common framework
- encourages open thinking
  (problem-driven rather than recipe-driven)

McCullagh and Nelder (1989)

Residuals re-visited

Linear model

$$\mu_i = \mathrm{E}[Y_i] = \alpha + \beta x_i \qquad \mathrm{Var}(Y) = \sigma^2$$

Generalised linear model

$$\mu_i = \mathrm{E}[Y_i] = h^{-1}(\alpha + \beta x_i) \qquad \mathrm{Var}(Y) = v(\mu_i)$$

Residuals and standardised residuals

	Linear	Generalised linear
residual	$y_i - \hat{\mu}_i$	$y_i - \hat{\mu}_i$
standardised residual	$(y_i - \hat{\mu}_i)/\hat{\sigma}$	$(y_i - \hat{\mu}_i)/\sqrt{v(\hat{\mu}_i}$

The statistical modelling cycle

1. **identify the question**
2. **design the study**
3. **collect the data**
4. **build the model**
    1. **exploration**
    2. **fitting**
    3. **diagnostic checking**
    4. **repeat (a) to (c) as necessary**
5. **answer the question**

Examples of generalised linear models

Open-ended count response

- Default choice is Poisson log-linear model

$$\log(\mu) = \alpha + \beta x \qquad Y \sim \text{Poiss}(\mu)$$

- Data-sets often show evidence of extra-Poisson variation

$$\log(\mu) = \alpha + \beta x \qquad \text{Var}(Y) = \phi\mu : \phi > 1$$

Examples of generalised linear models (2)

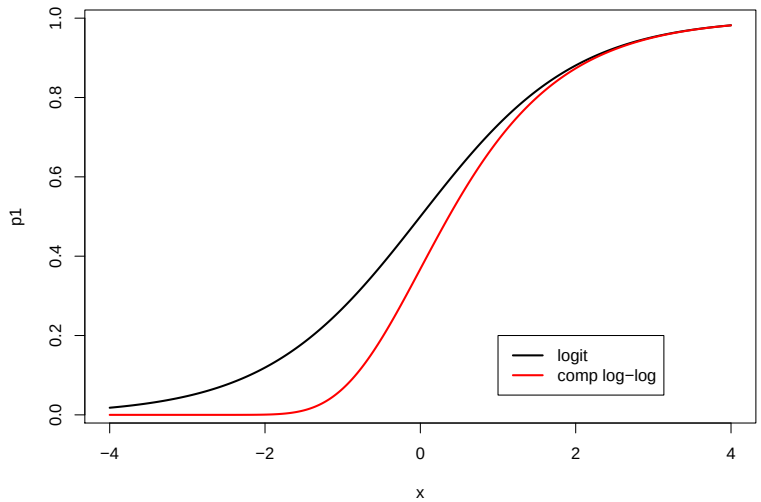
Binary or closed-count response (number out of n)

- **(Binomial) logistic model**

$$\log\{\mu/(1-\mu)\} = \alpha + \beta x \qquad \mu = P(Y = 1)$$

- **Complementary log-log**

$$\log\{-\log(\mu)\} = \alpha + \beta x \qquad \mu = P(Y = 1)$$



- **Extra-binomial variation can also arise, and can be handled in the same way as extra-Poisson variation**

Examples of generalised linear models (3)

Survival analysis (life-times)

- **Exponential log-linear model**

$$\log(\mu) = \alpha + \beta x \qquad \mu = P(Y = 1)$$

$$f(y) = \mu^{-1} \exp(-y/\mu)$$

Rarely used in practice, because of the problem of **censoring**

- Standard approach is semi-parametric **proportional hazards modelling**

Cox (1972)

Review: what have we learnt?

1. **statistical models are devices to answer questions**
2. **statistical models can be empirical or mechanistic (or a mix of the two)**
3. **the linear regression model is surprisingly flexible**
4. **but for count data, generalised linear models are usually preferable**
5. **residual diagnostic checks on model fit are an important part of the statistical modelling cycle**

Random effects

Linear regression re-visited

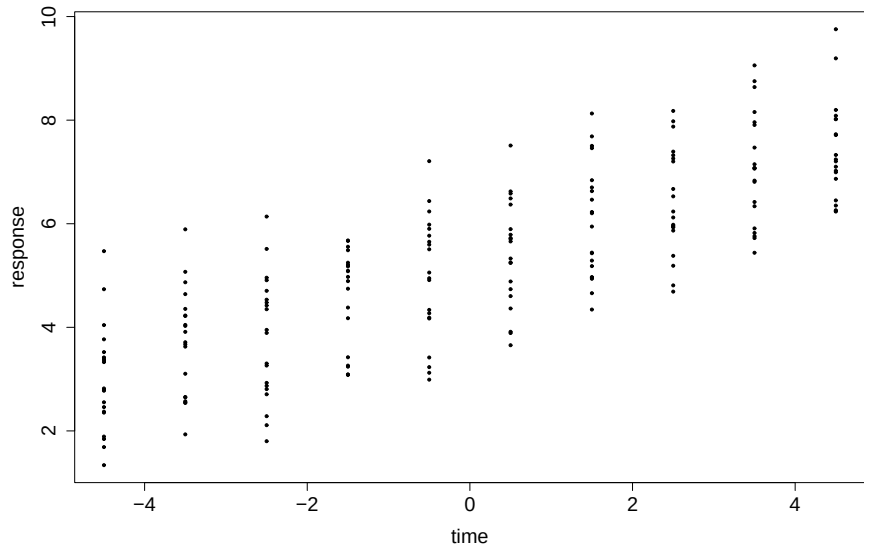
- **input variable x**

  factor, covariate, explanatory variable, ...

- **output variable y**

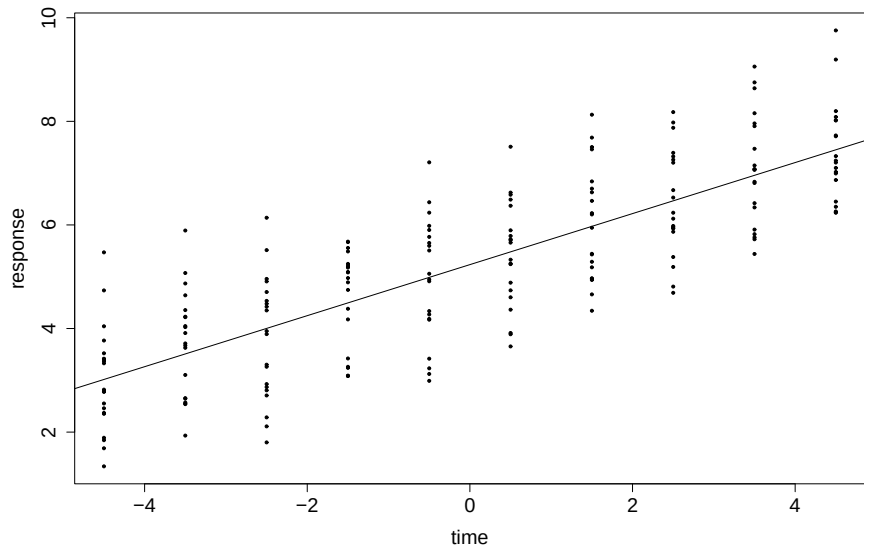
  response, end-point, primary outcome,...

A synthetic example



- **relationship between x and y can be captured approximately by a straight line**
- **scatter about the line is approximately the same at all values of x**

Interpreting the linear regression model

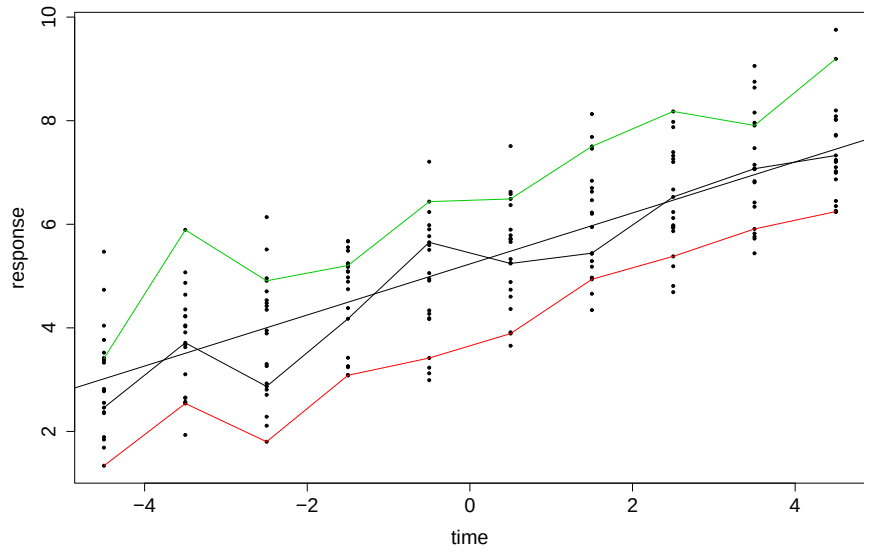


$$Y_i = \alpha + \beta x_i + z_i$$

How precisely can we:

- **estimate the parameters α and β?**
- **estimate the straight-line relationship?**
- **predict a future value of y?**

The data are longitudinal



- **simple linear regression software assumes that data are uncorrelated**
- **in longitudinal studies, with repeated measurements on each subject, this is rarely true**
- **reported standard errors and p-values are then not correct**

Diggle, Heagerty, Liang and Zeger (2002)

Standard fitting method: least squares

```
> fit1<-lm(y~1+x)
> summary(fit1)

... plus lots of stuff you don't want to know

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.26789    0.10274   31.81   <2e-16 ***
x            0.39286    0.01924   20.41   <2e-16 ***

Residual standard error: 0.7817 on 198 degrees of freedom
Multiple R-squared: 0.6779,     Adjusted R-squared: 0.6763
F-statistic: 416.7 on 1 and 198 DF,  p-value: < 2.2e-16
```

Correct fitting method: maximum likelihood with random effect

```
> library(nlme)
> fit2<-lme(y~1+x,random=~1|id)
> summary(fit2)
Linear mixed-effects model fit by REML
...
Random effects:
 Formula: ~1 | id
        (Intercept)  Residual
StdDev:   0.7477531 0.2730349

Fixed effects: y ~ 1+x
               Value  Std.Error  DF  t-value p-value
(Intercept) 3.267887 0.17100989 179 19.10935       0
x           0.392856 0.00672165 179 58.44637       0

Number of Observations: 200
Number of Groups: 20
```

Random effects

- **Random effects can be thought of as missing information on individual subjects that, were it available, would be included in the statistical model**
- **we model the missing information as a random sample from a distribution (usually, we assume a Normal distribution)**
- **This induces correlation amongst repeated measurements on the same subjects**

Example: some subjects are intrinsically high responders, others intrinsically low responders

$\Rightarrow$ replace fixed intercept α by a random intercept, $\alpha + U_i$

Random effects or fixed effects?

For our synthetic example, write:

$$\begin{aligned} Y_{ij} &= j^{th} \text{ response from } i^{th} \text{ subject: } i = 1, \ldots, n \\ x_{ij} &= \text{corresponding value of explanatory variable} \end{aligned}$$

A random effects model

- $Y_{ij} = \alpha + \beta x_{ij} + U_i + Z_{ij}$
- U_i = random effect for subject i
- Z_{ij} = residual
- all U_i and all Z_{ij} mutually independent

Model implies that different responses on the same subject are positively correlated:

$$\rho = \frac{\mathrm{Var}(U)}{\mathrm{Var}(U) + \mathrm{Var}(Z)}$$

Random effects or fixed effects?

For our synthetic example, write:

$$\begin{aligned} Y_{ij} &= j^{th} \text{ response from } i^{th} \text{ subject: } i = 1, \ldots, n \\ x_{ij} &= \text{corresponding value of explanatory variable} \end{aligned}$$

A fixed effects model

- $Y_{ij} = \alpha_i + \beta x_{ij} + Z_{ij}$
- $\alpha_i =$ intercept for subject i
- $Z_{ij} =$ residual
- all Z_{ij} mutually independent

Model implies that different responses on the same subject are uncorrelated

Random or fixed effects: which model is correct?

- **They both are**
- **The choice between them depends primarily on why you are analysing the data**
  - **to adjust for the heterogentiy amonst the particular subjects in your data**
  - **to undertand the heterogeneity amongst members of the population from which your subjects were drawn**
- **These two cases call for fixed effects and random effects models, respectively**
- **A second consideration is statistical efficiency: for large n, the fixed effects model has many more parameters than the random effects model, which necessarily implies that the parameter estimates are less precise**

A philosophical objection to fixed effects?

Y_{ij} = mark for student i on exam paper $j = 1, \ldots, p$

Question: what overall mark should you give to student i?

Fixed effects model: $Y_{ij} = \alpha_i + Z_{ij} \quad Z_{ij} \sim \mathrm{N}(0, \tau^2)$, independent

Answer: $\hat{\alpha}_i = \bar{Y}_i$, the observed average mark for student i

Random effects model: $Y_{ij} = A_i + Z_{ij}$

- $Z_{ij} \sim \mathrm{N}(0, \tau^2)$, independent
- $A_i \sim \mathrm{N}(\alpha, \sigma^2)$, independent

Answer: $\hat{A}_i = c \times \bar{y}_i + (1 - c) \times \bar{y}, \quad c = p/(p + \tau^2/\sigma^2)$

Observed average for student i is "shrunk" towards the class average

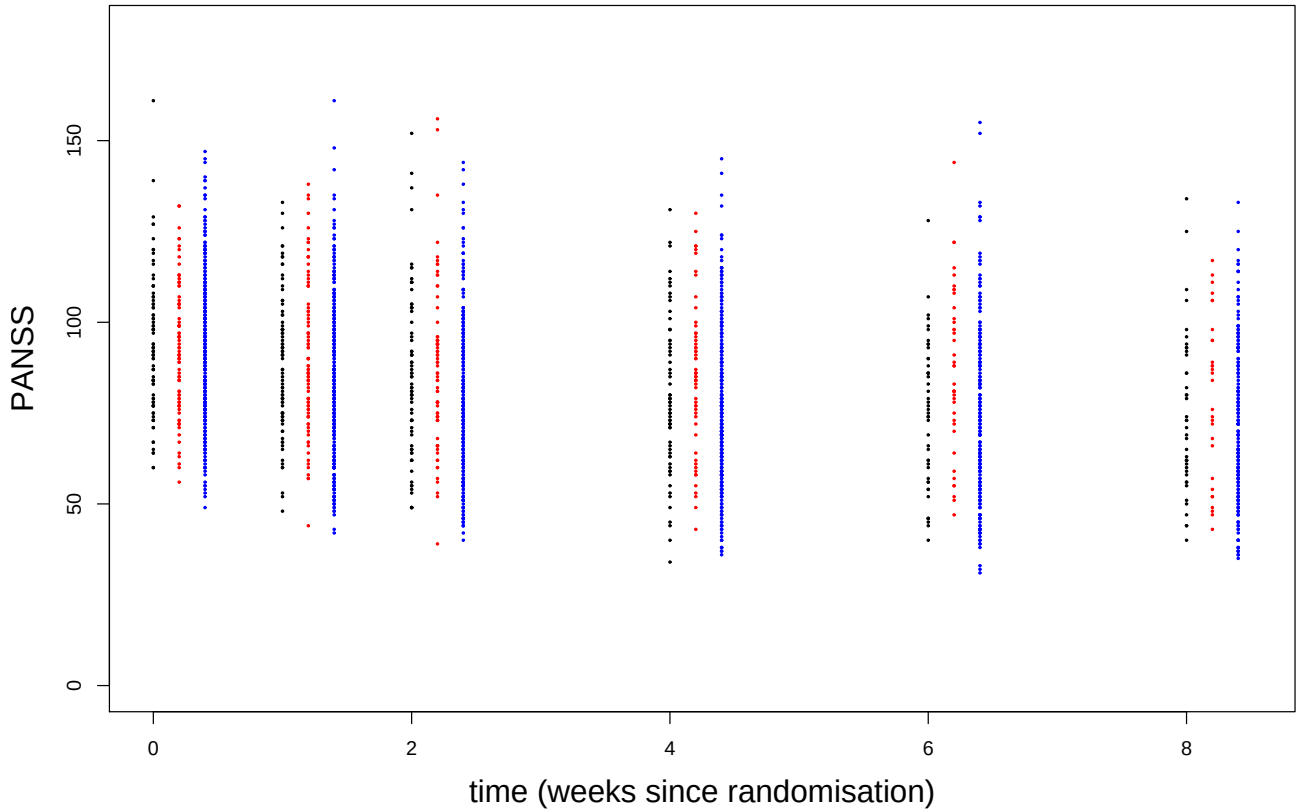
An RCT of drug therapies for schizophrenia

- **Randomised clinical trial of drug therapies**
- **Three treatments:**

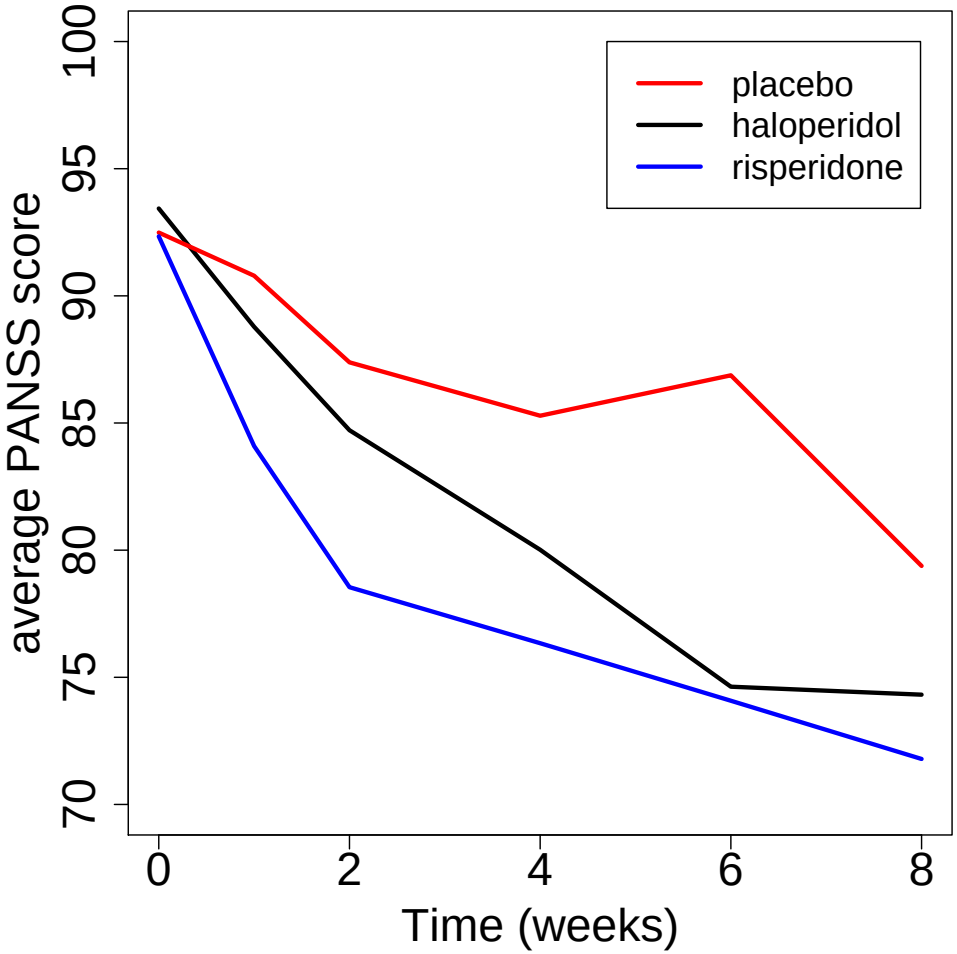
  haloperidol (standard); placebo; risperidone (novel)
- **Dropout due to "inadequate response to treatment"**

Treatment	Number of non-dropouts at week					
	0	1	2	4	6	8
haloperidol	85	83	74	64	46	41
placebo	88	86	70	56	40	29
risperidone	345	340	307	276	229	199
total	518	509	451	396	315	269

The schizophrenia trial data



A summary of the schizophrenia trial data



A model for the schizophrenia trial data

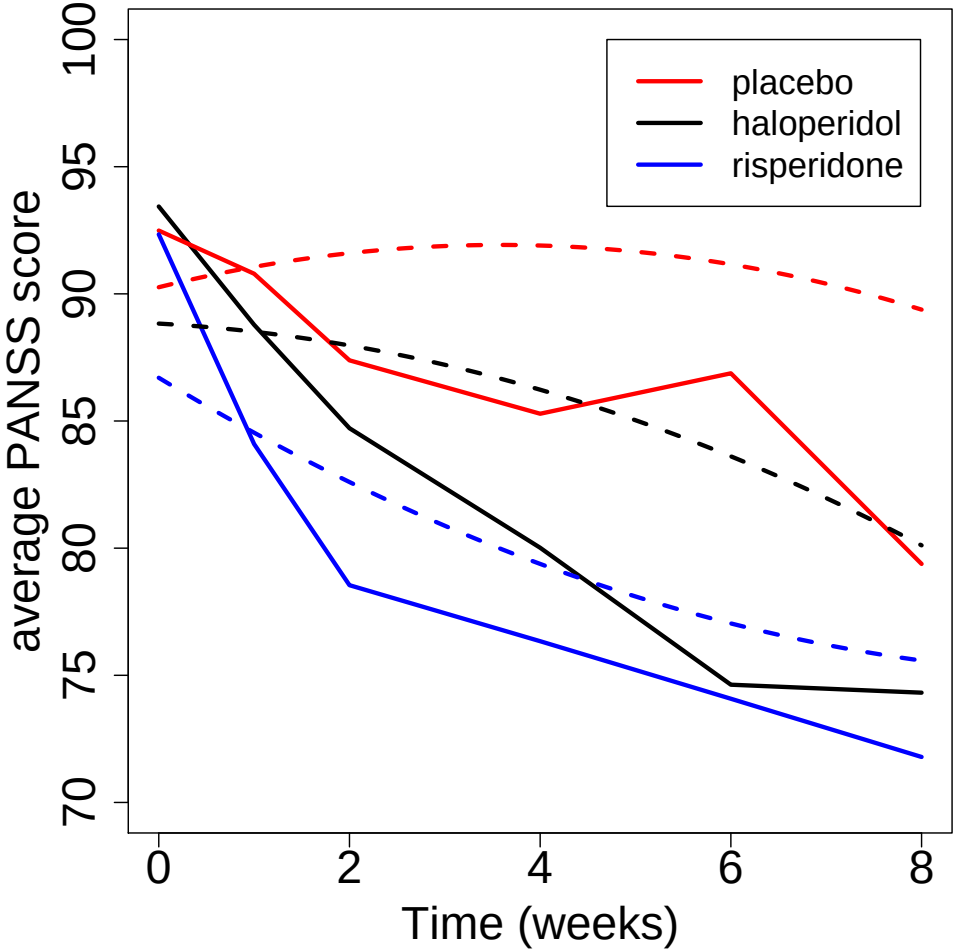
Mean response depends on treatment and time

Two random effects:

- **between subjects (high or low responders)**
- **between times within subjects (good and bad days)**

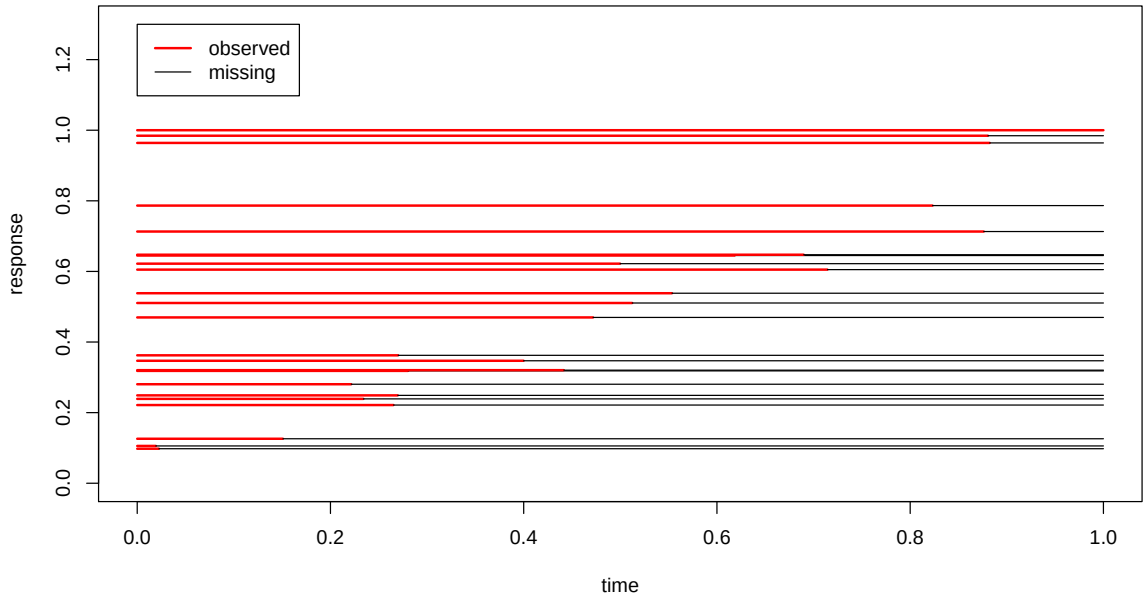
Method of analysis allows for dropouts: maximum likelihood with random effects

PANSS mean response profiles



What's going on?

- **dropout is selective (high responders more likely to leave)**
- **but the data are correlated**
- **and this allows the model to infer what you would have seen, had there not been any dropouts**
- **which may or may not be what you want**



Diggle, Heagerty,Liang and Zeger (2002)

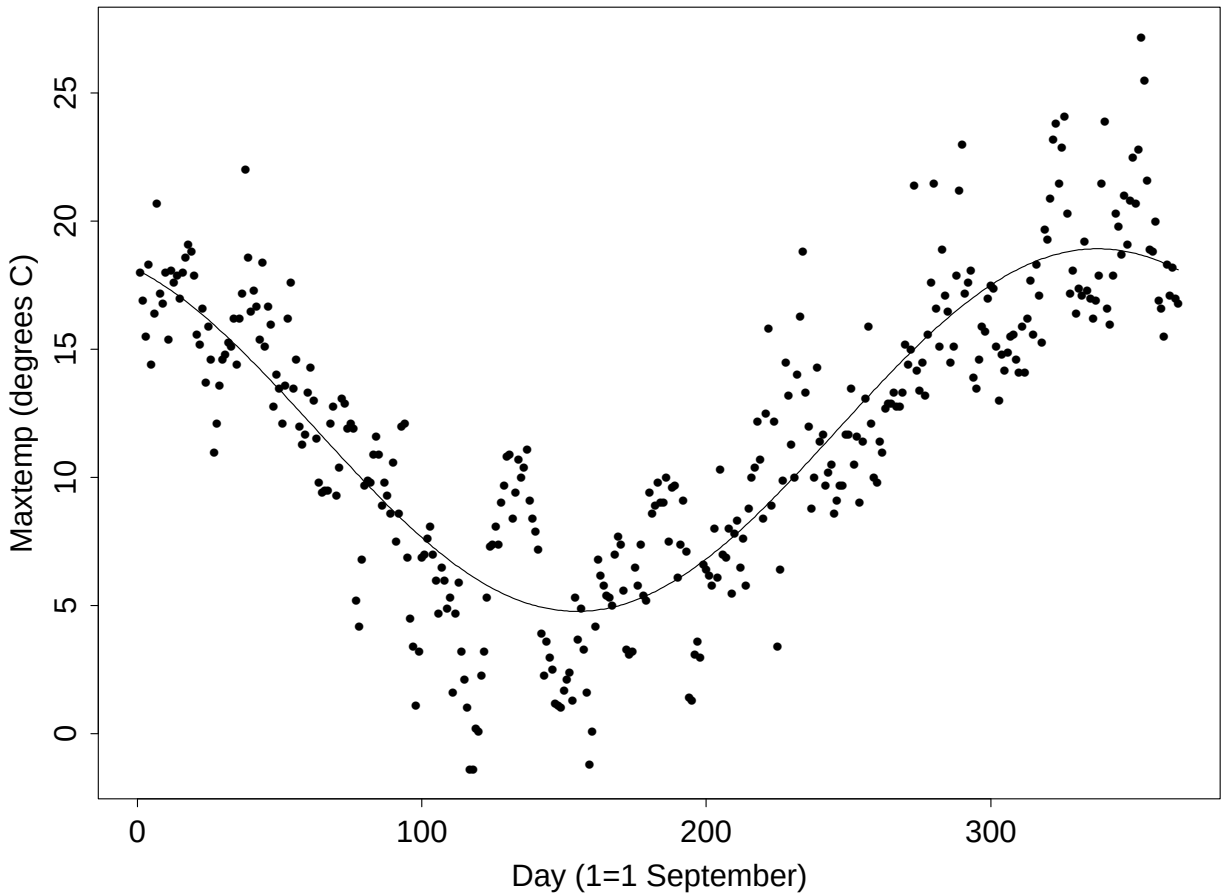
Review: what have we learnt?

1. **Fixed effects describe the variation in average responses of groups of subjects according to their measured characteristics (age, sex, treatment,...)**
2. **Random effects describe variation in subject-specific responses according to their unmeasured characteristics**
3. **Both kinds of model can easily be fitted using the open-source software R, or in various proprietary packages**
4. **Dropout in longitudinal studies can have surprising consequences:**
5. **Random effects and parameters are different things**
   - **parameters don't change if you re-run an experiment**
   - **random effects do**

Time

A meteorological time series

- **maximum daily temperatures (degrees C) at Bailrigg (Lancaster) field-station, September 1995 to August 1996**
- **note that an unusually cold Christmas 1995 was followed by a mild period in January-February**



Points for discussion

- **what are the main features of the data?**
- **how did I fit the smooth curve to the data?**
- **what features are and are not explained by the fitted curve?**

A harmonic regression model

$$\begin{aligned} Y(t) &= \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual} \\ &= \mu + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p) + \text{residual} \end{aligned}$$

- μ = overall mean value (of time series $Y(t)$)
- p = period
- α = amplitude
- ϕ = phase

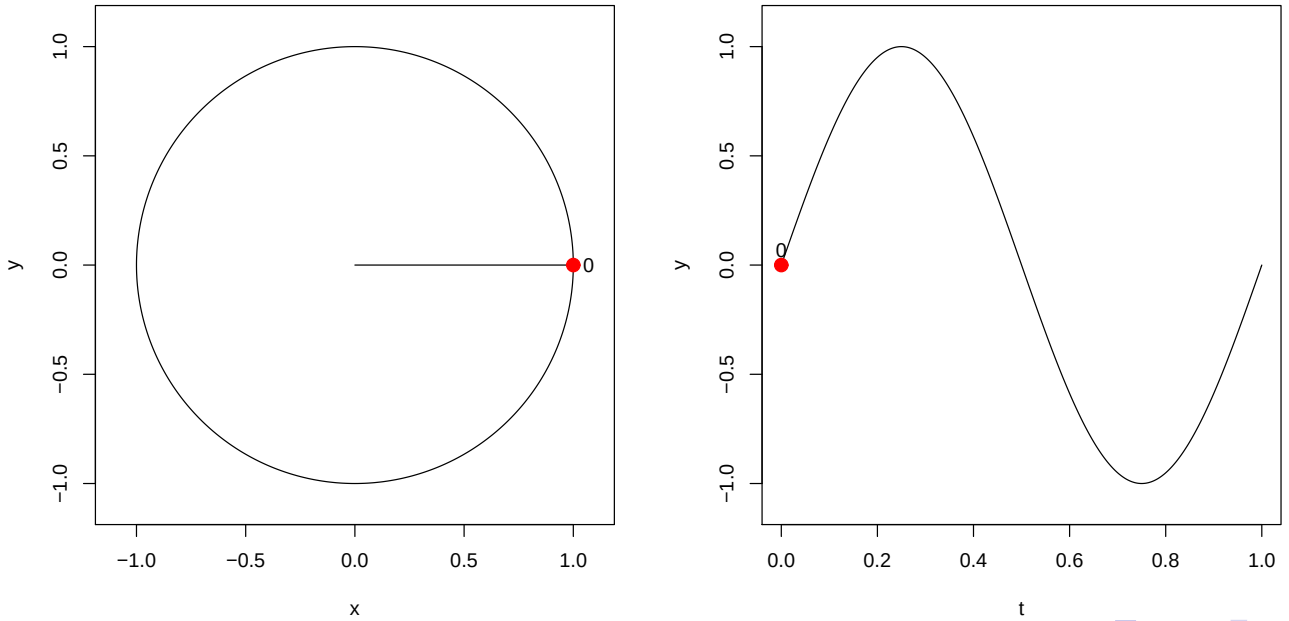
Usually, the period is known, but the mean, amplitude and phase are not

Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle

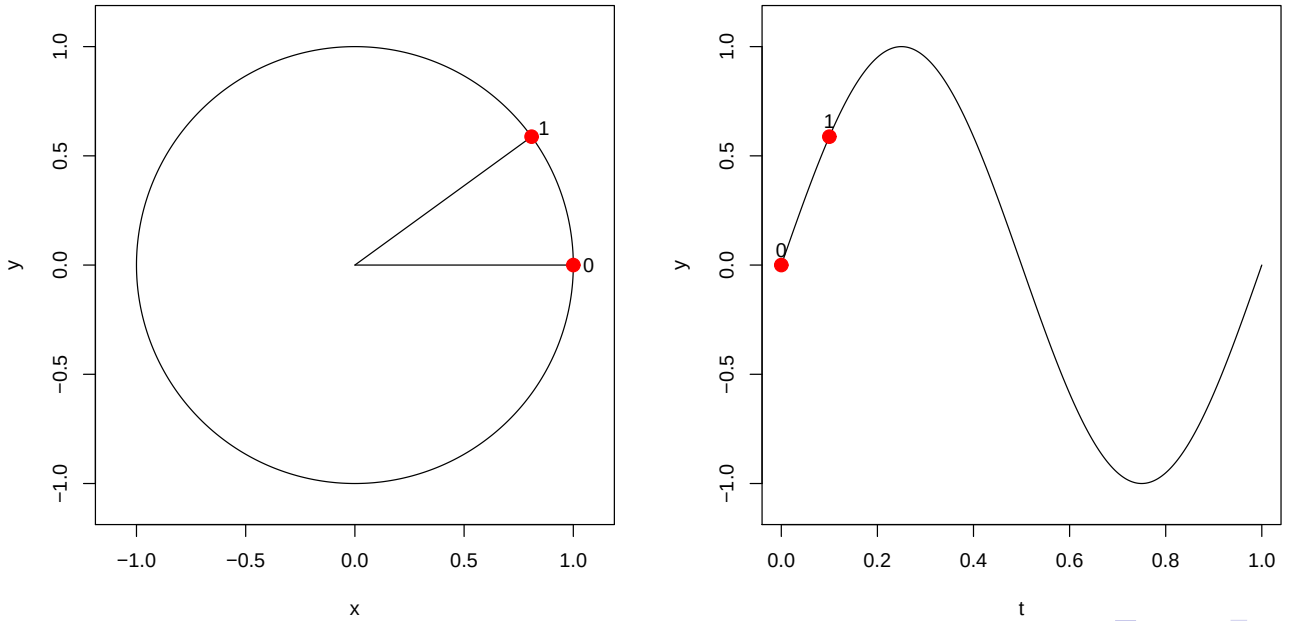


Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle

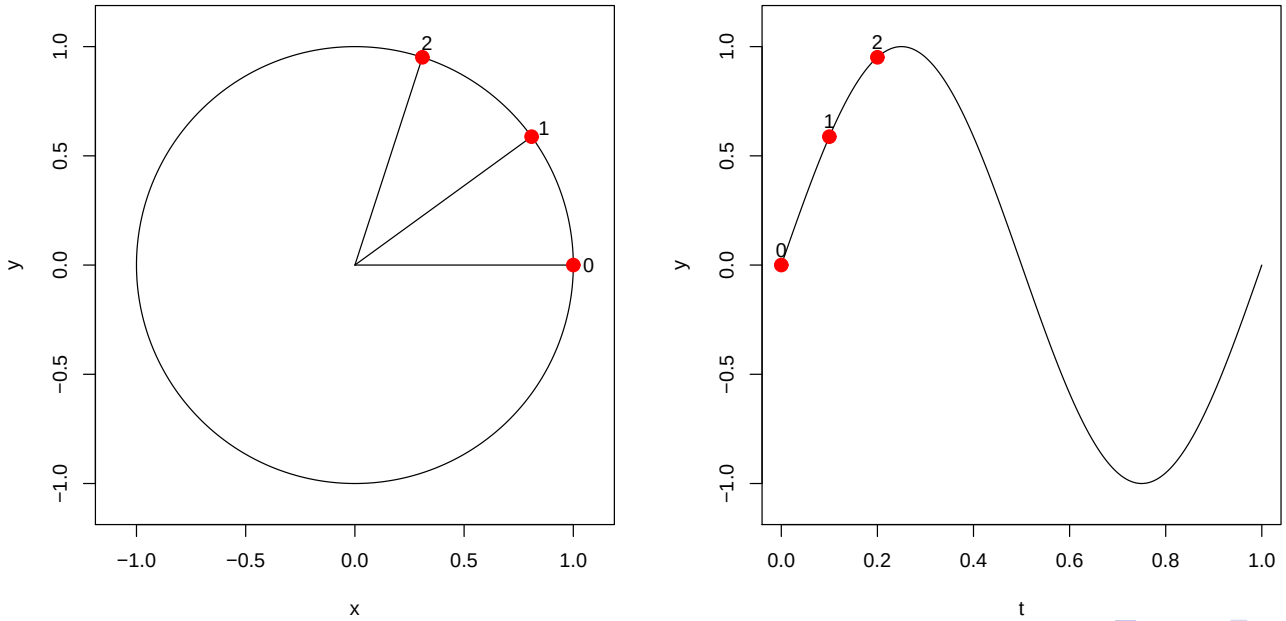


Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle

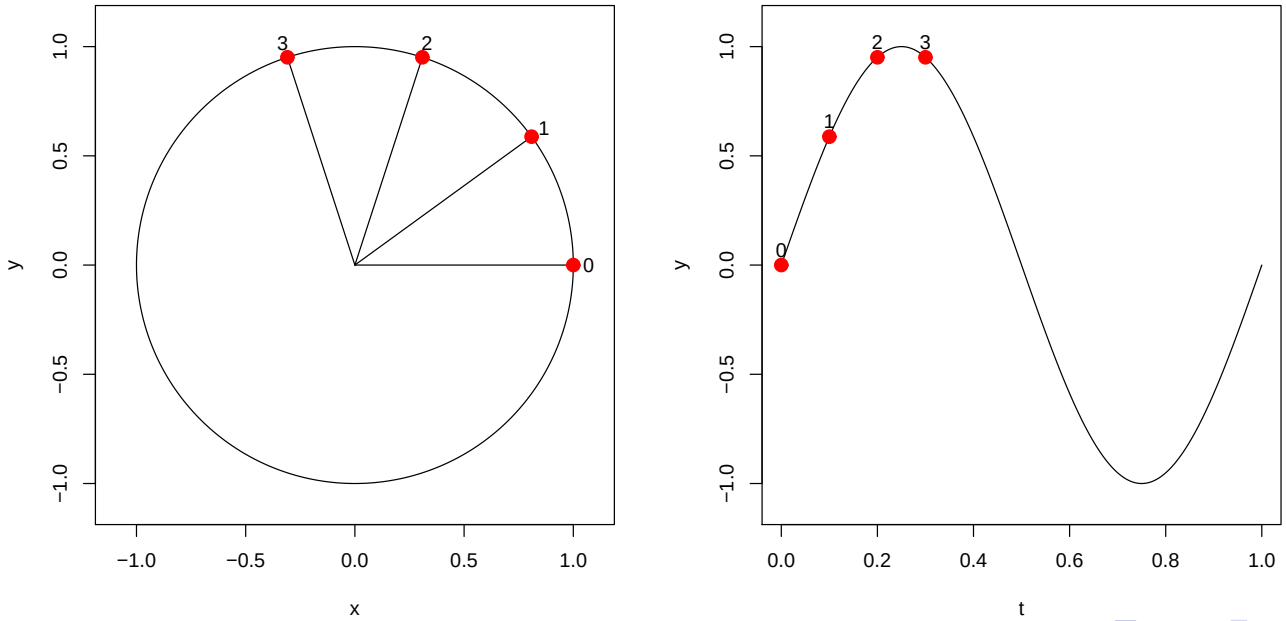


Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle

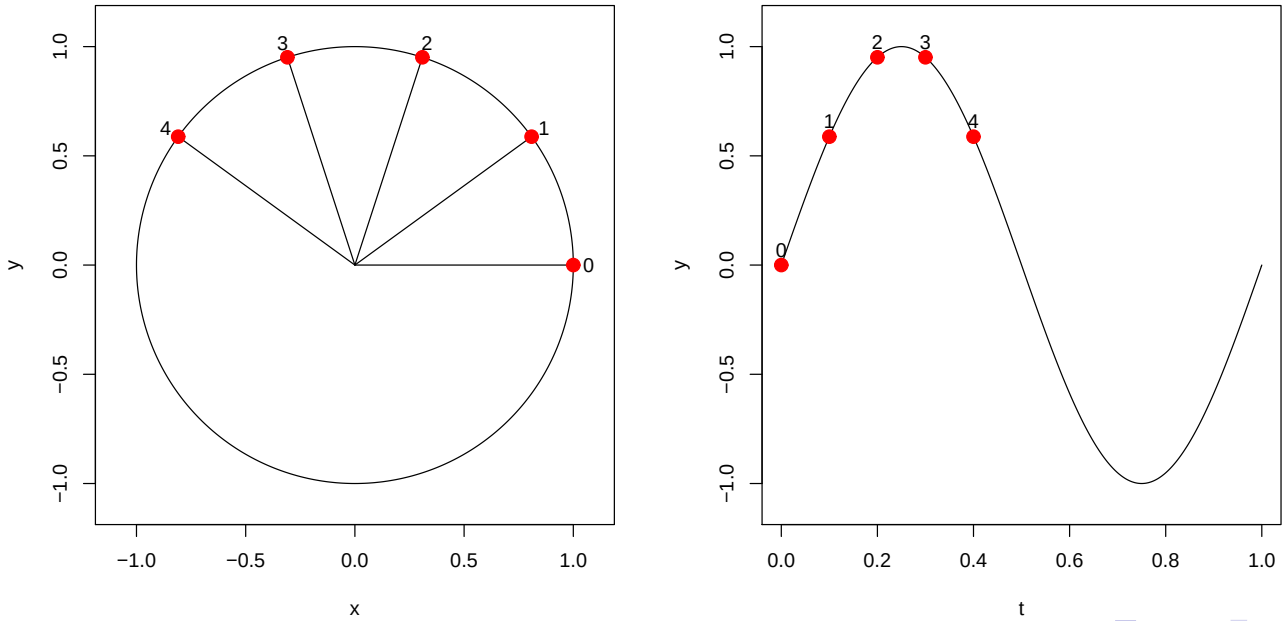


Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle

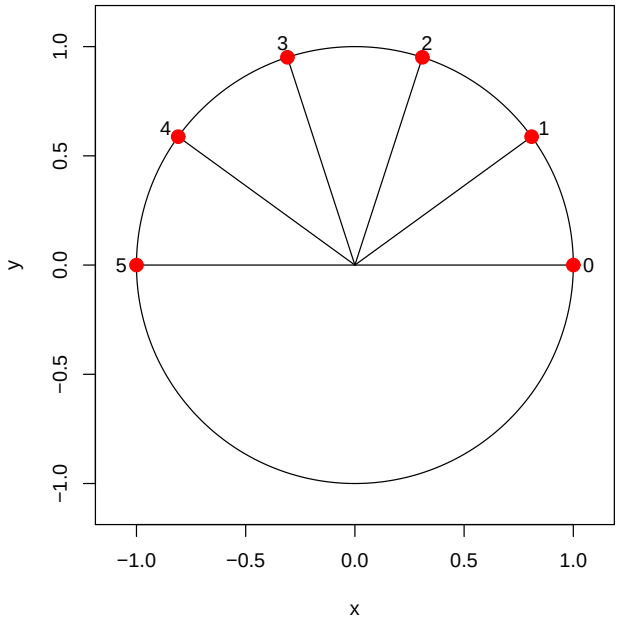


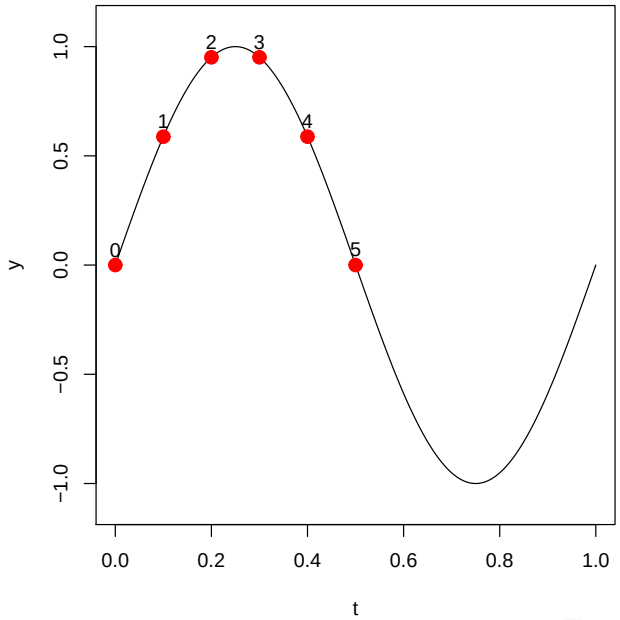
Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle



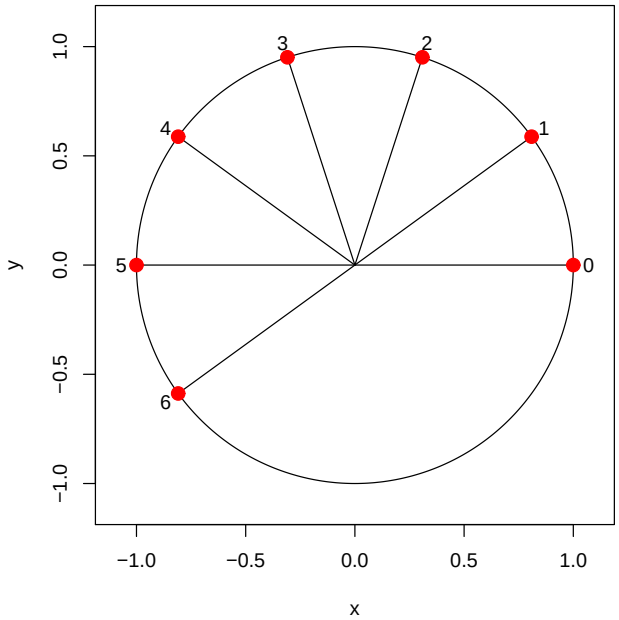


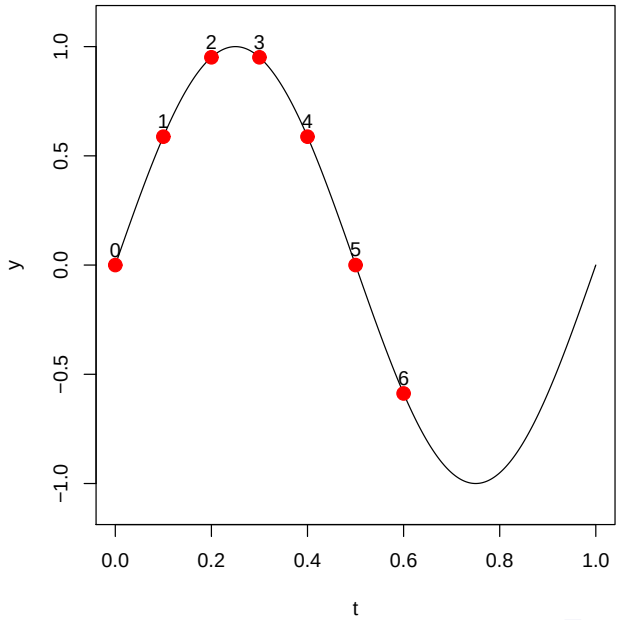
Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle



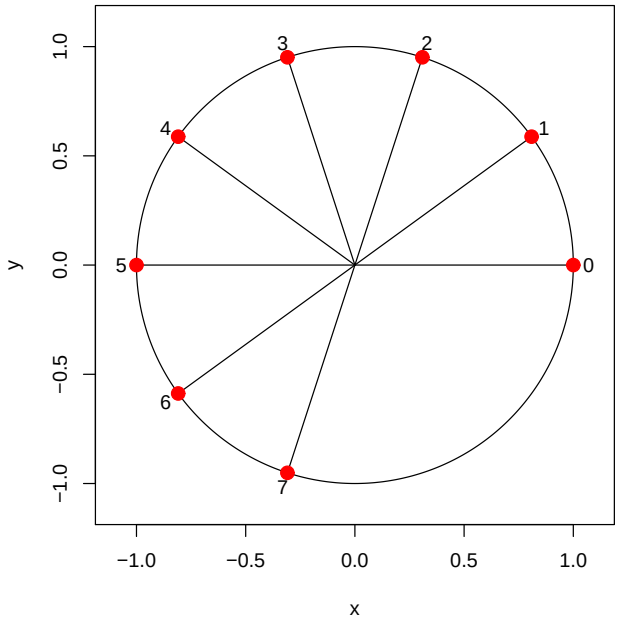


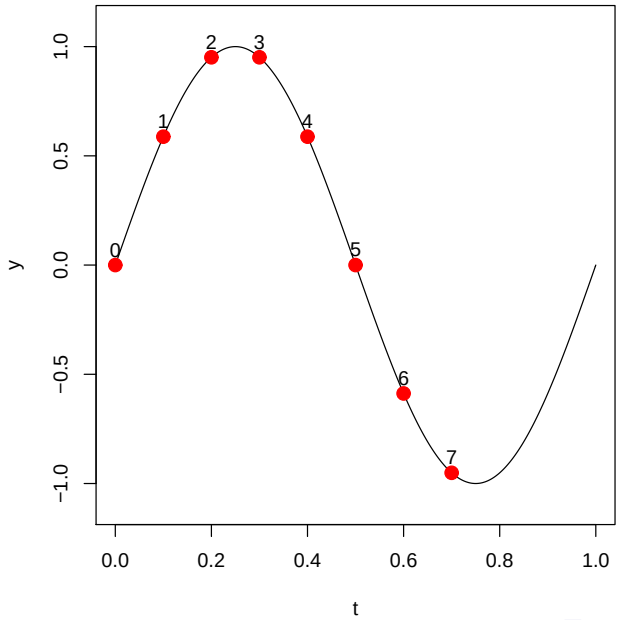
Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle



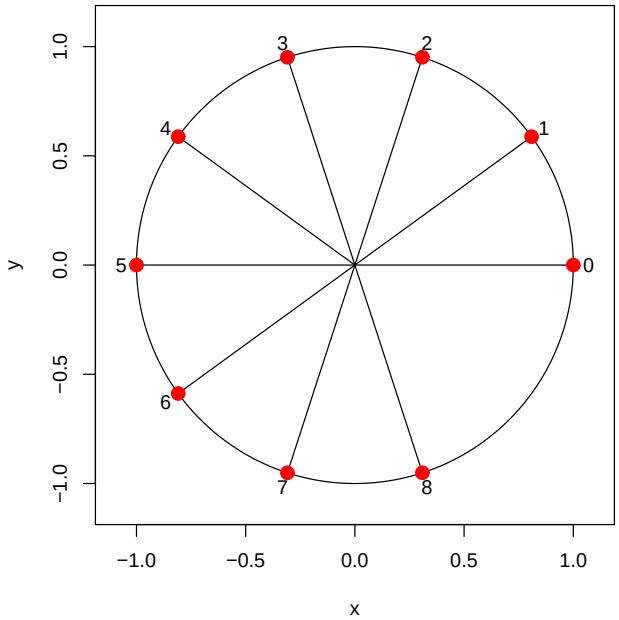


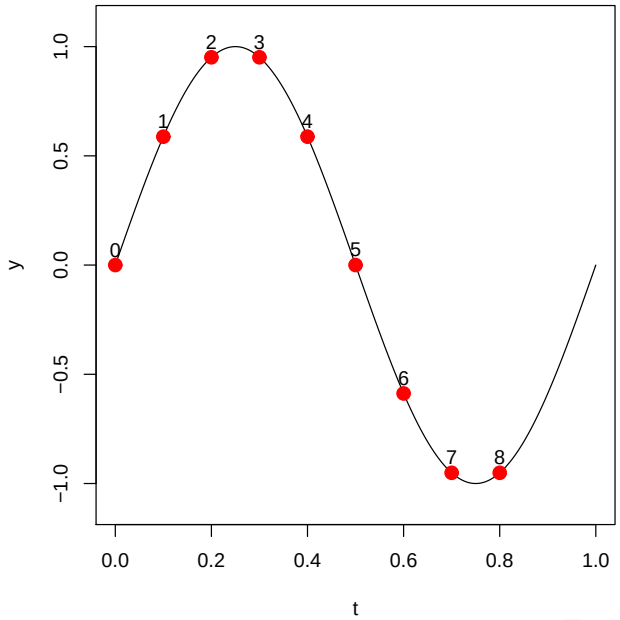
Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle



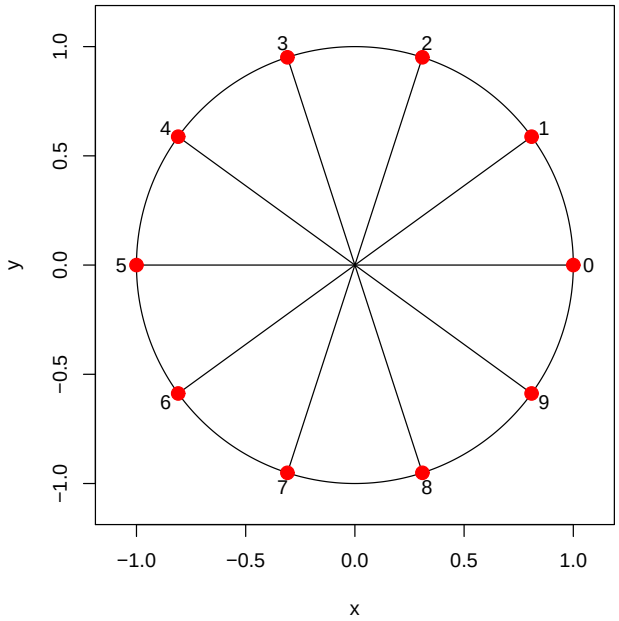


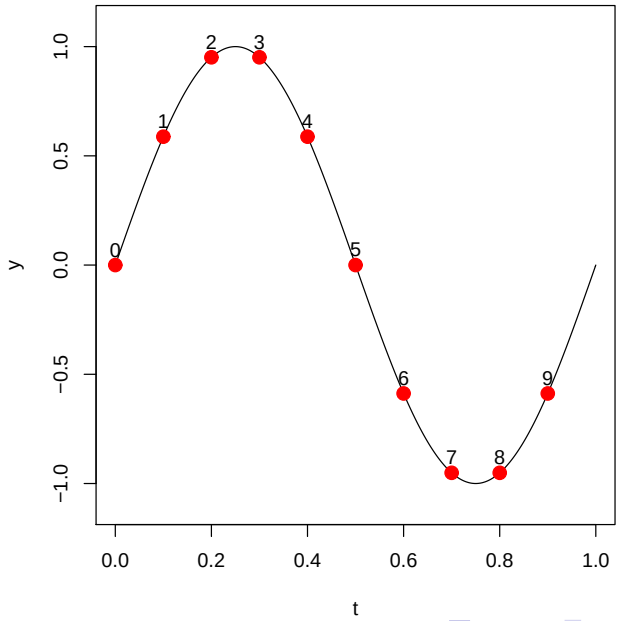
Why does the model work?

Use the first form of the model,

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

Now imagine tracking the vertical displacement of a particle moving at constant speed around a circle





Lifting, stretching, shifting

$$Y(t) = \mu + \alpha \cos(2\pi t/p + \phi) + \text{residual}$$

- **Lifting**

  μ: adjust to match observed and modelled average value

- **Stretching**

  α: adjust to match observed and modelled range

- **Shifting**

  ϕ: adjust to match observed and modelled nadir

Fitting the model

Use the second form of the model,

$$Y(t) = \mu + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p) + \text{residual}$$

Note that the following quantities are known, i.e. they can be calculated without having to estimate anything

- $x_1(t) = \cos(2\pi t/p)$
- $x_2(t) = \sin(2\pi t/p)$

Re-write the model as

$$Y = \mu + \beta_1 x_1 + \beta_2 x_2$$

After fitting (see next page), amplitude and phase can be recovered using

$$\alpha = \sqrt{\beta_1^2 + \beta_2^2} \qquad \phi = \tan^{-1}(\beta_2/\beta_1)$$

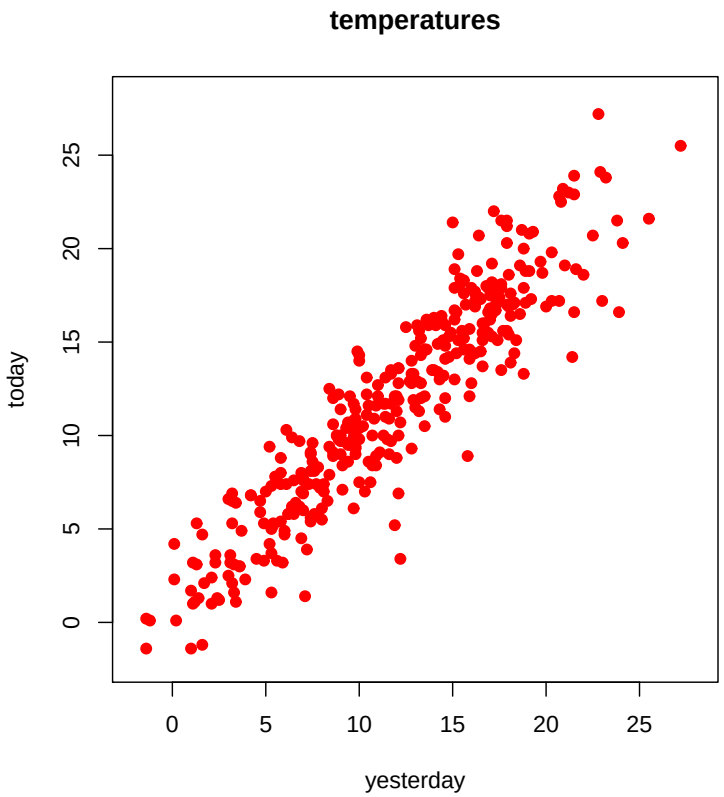
Using the `lm()` function to fit the model

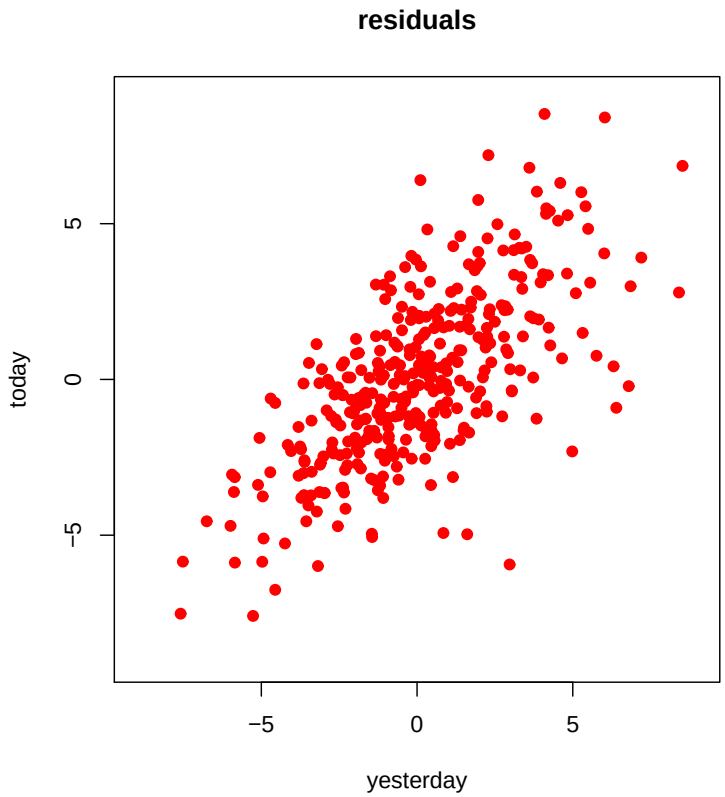
```
data<-read.table("maxtemp.dat")
y<-data[,4]
day<-1:366
x1<-cos(2*pi*day/366)
x2<-sin(2*pi*day/366)
fit<-lm(y~x1+x2)
summary(fit)
```

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5921 -1.8240 -0.1475  1.7140  8.5232

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.8467     0.1441   82.22   <2e-16 ***
x1            6.2508     0.2038   30.68   <2e-16 ***
x2           -3.3177     0.2038  -16.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.756 on 363 degrees of freedom
Multiple R-Squared: 0.7687,     Adjusted R-squared: 0.7674
F-statistic: 603.1 on 2 and 363 DF,  p-value: < 2.2e-16
```

Autocorrelation

- **relationship between today's and yesterday's temperature?**
- **relationship between today's and yesterday's residual?**

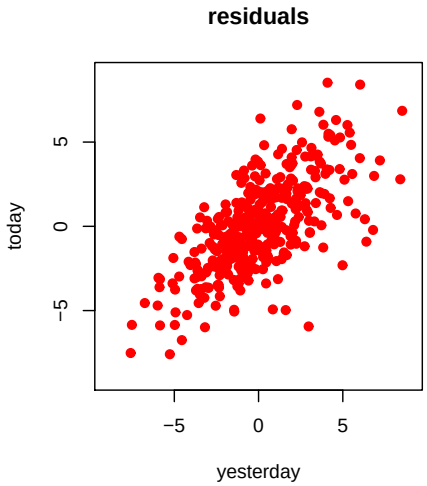


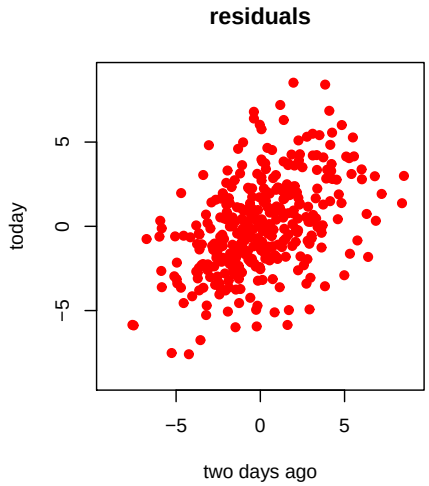


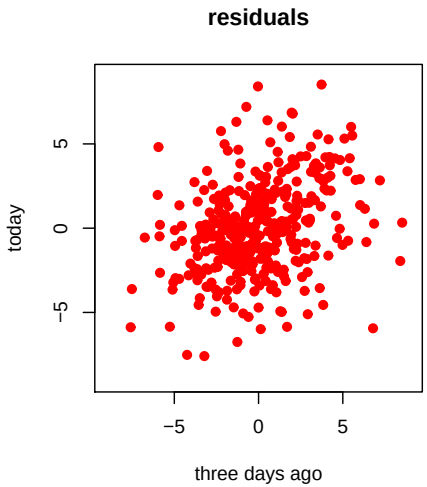
- **how and why are the two relationships different?**

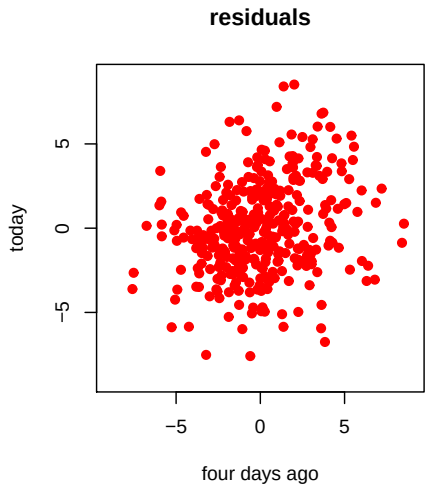
Autocorrelation (2)

How does the relationship between residuals today and *k* days ago change as *k* increases?



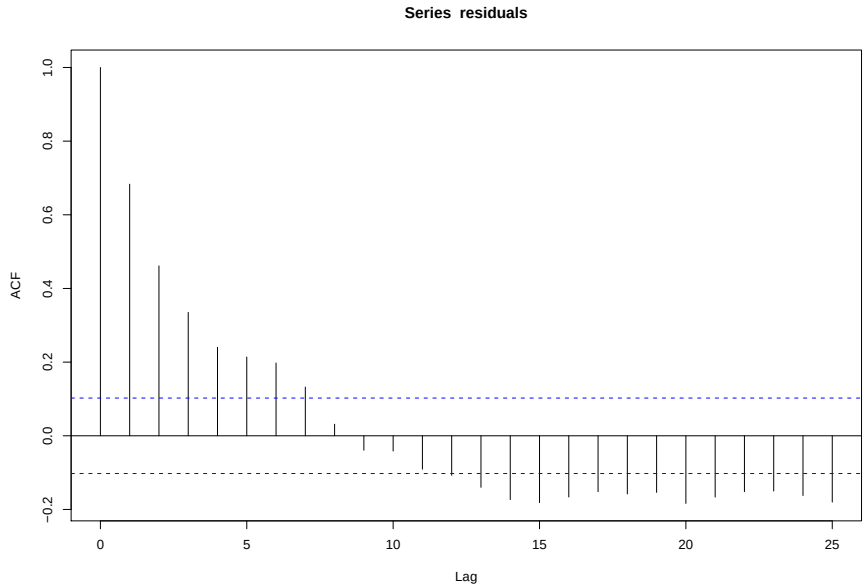






Autocorrelation (3)

- **lag-k autocorrelation is the correlation between pairs of values from the same time series k time-units apart**
- **correlogram is a plot of lag-k autocorrelation against k**



- **dashed lines at $\pm 2/\sqrt{n}$ are pointwise 95% limits for uncorrelated residuals**
- **but overall pattern is more important than individual numerical values**

Exercise

Imagine that you have data on daily maximum temperatures for several years, up to today.

1. **How would make a forecast of:**
   - **tomorrow's temperature?**
   - **the temperature one month from now?**
2. **In what ways are your two answers different, and why?**

Time series models and random effects

$$Y(t) = \alpha + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p) + \text{residual}(t)$$

1. Model consists of a time-varying mean, also called the trend,

$$\mu(t) = \alpha + \beta_1 \cos(2\pi t/p) + \beta_2 \sin(2\pi t/p)$$

and stochastic variation, residual(t), about the trend.

2. Decompose the residual into two terms:

$$\text{residual}(t) = S(t) + Z_t$$

- $\text{Cov}\{S(t), S(t-u)\} = \sigma^2 \exp(-u/\phi)$ (random effect)
- Z_t uncorrelated $\text{N}(0, \tau^2)$ (noise/measurement error)

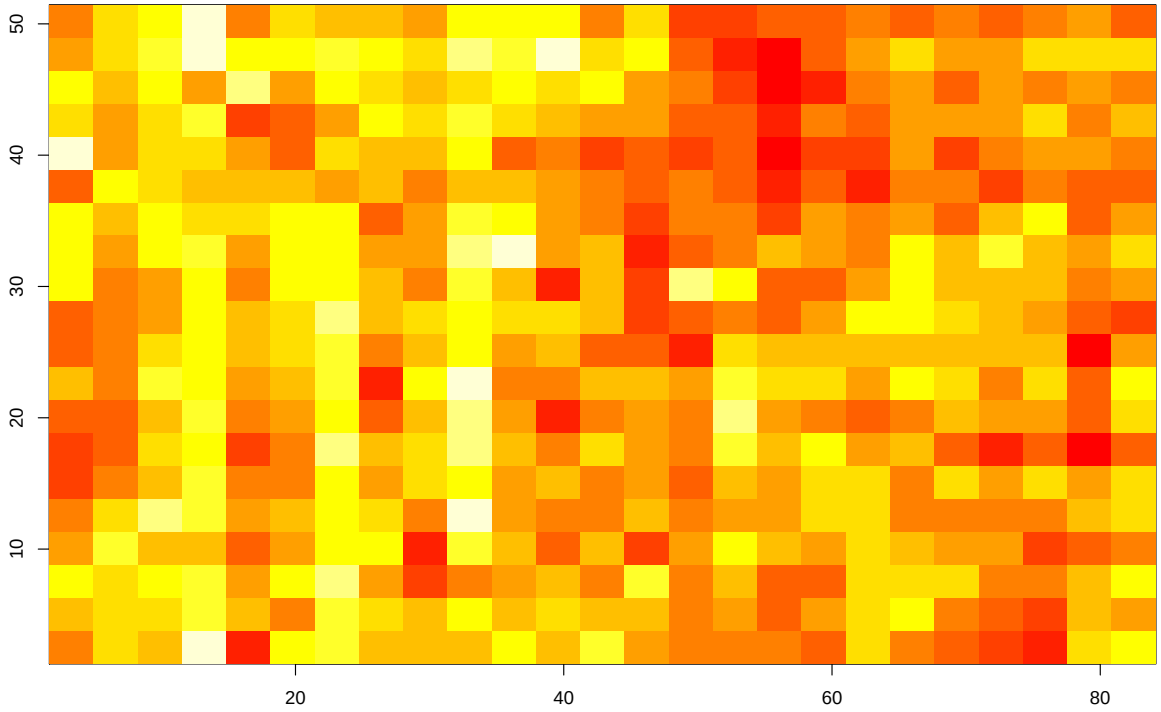
Diggle (1990), Priestley (1981)

Space

Spatial Data Formats

spatially aggregated; spatially sampled; point process

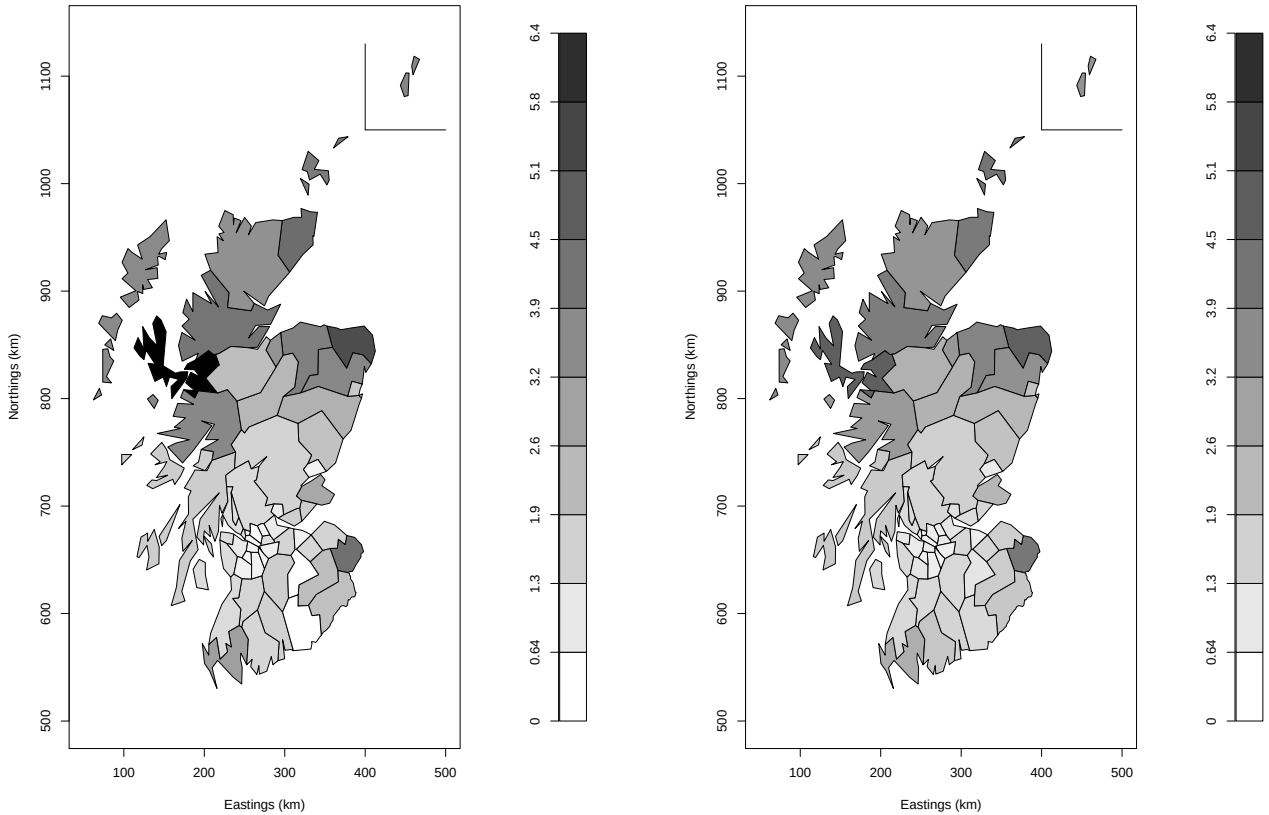
Example 1a. Wheat uniformity trial



Mercer and Hall (1911)

Example 1b. Cancer atlases

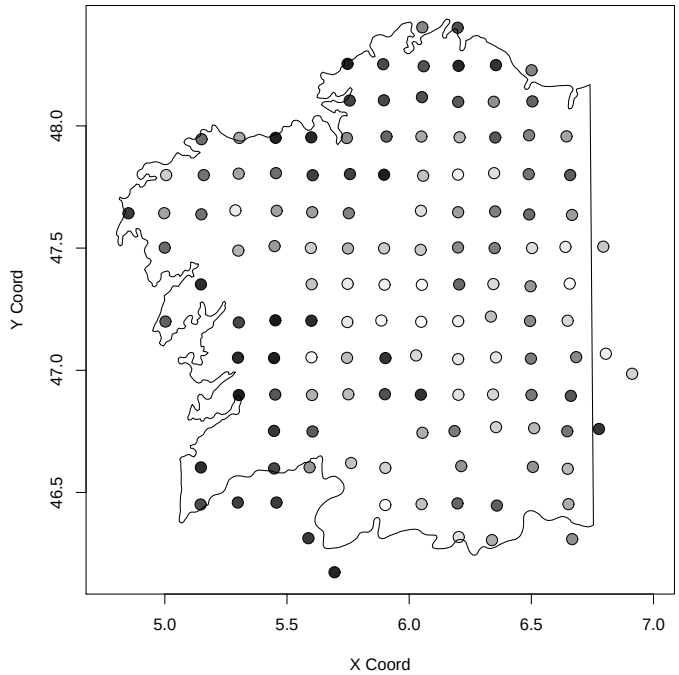
Raw (left panel) and spatially smoothed (right panel) relative risk estimates for
lip cancer in 56 Scottish counties



Wakefield (2007)

Example 2. Galicia biomonitoring study

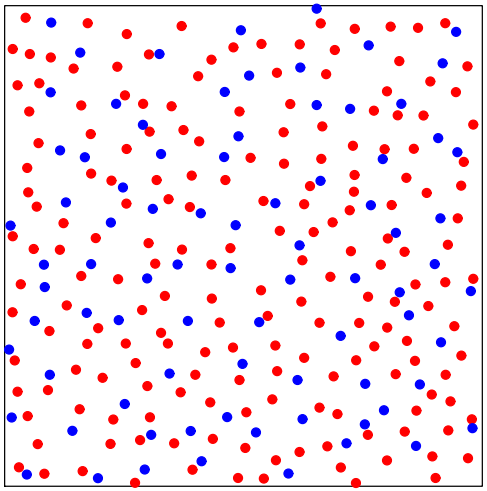
Lead concentrations measured in samples of moss, map shows locations and log-concentrations

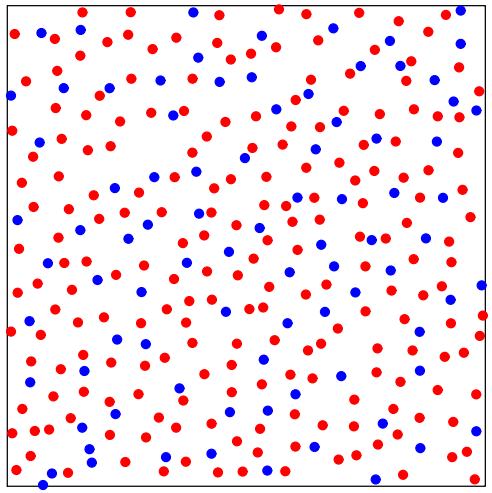


Diggle, Menezes and Su (2010)

Example 3. Retinal mosaics

Locations of two types of light-responsive cells in macaque retina (2 animals)





Eglen and Wong (2008)

Spatial dependence

First law of geography: all things are related but close things are more strongly related than distant things

Spatial dependence may be:

- **functional (deterministic)**
- **statistical (stochastic)**

Origin of statistical dependence may be mechanistic (causal) or empirical (descriptive/predictive)

Geostatistical regression modelling

A common scenario in tropical disease epidemiology:

- **Objective:** understand geographical variation in risk
- **Study-design:** Identify n communities in study-region A, by a suitable sampling scheme (not necessarily completely random)
- **Data:** Measure empirical prevalence and hypothesised risk-factors on each sampled community, also risk-factors on all unsampled communities for which estimates of risk are required
- **Model:** either a linear or a generalized linear model for prevalence (or transformation thereof)
- **Diagnostic checking:** all the usual plus check for spatial independence of residuals

Diggle and Giorgi (2019)

A model-based approach: random effects again

$$Y_i = \mu(x_i) + S(x_i) + Z_i : i = 1, ..., n$$

- x_i = location
- Y_i = (transformed) prevalence
- $\mu(x_i) = u_i'\beta$ (fixed effects)
- $S(x)$ = spatially correlated process, variance σ^2
- Z_i = uncorrelated residuals, variance τ^2

Estimating correlation structure

- **Regression residuals,** $r_i = Y_i - u_i\hat{\beta}$
- **Each** r_i **estimates** $S(x_i) + Z_i$
- **Do we need the random effect process** $S(x)$**?**

The variogram

- $d_{ij} = ||x_i - x_j||$**(distance)** $\quad v_{ij} = \frac{1}{2}(r_i - r_j)^2$
- $V(d_{ij}) = \mathrm{E}[v_{ij}^2]$
- $V(d) = \tau^2 + \sigma^2\{1 - \rho(d)\}$

Onchocerciasis: a vector-borne disease

"A small black fly, with thick shoulders and bullet-head, infests the place, and torments the naked arms and legs of the people with its sharp stings to an extent that must make life miserable to them"

John Hanning Speke, Uganda, 1864

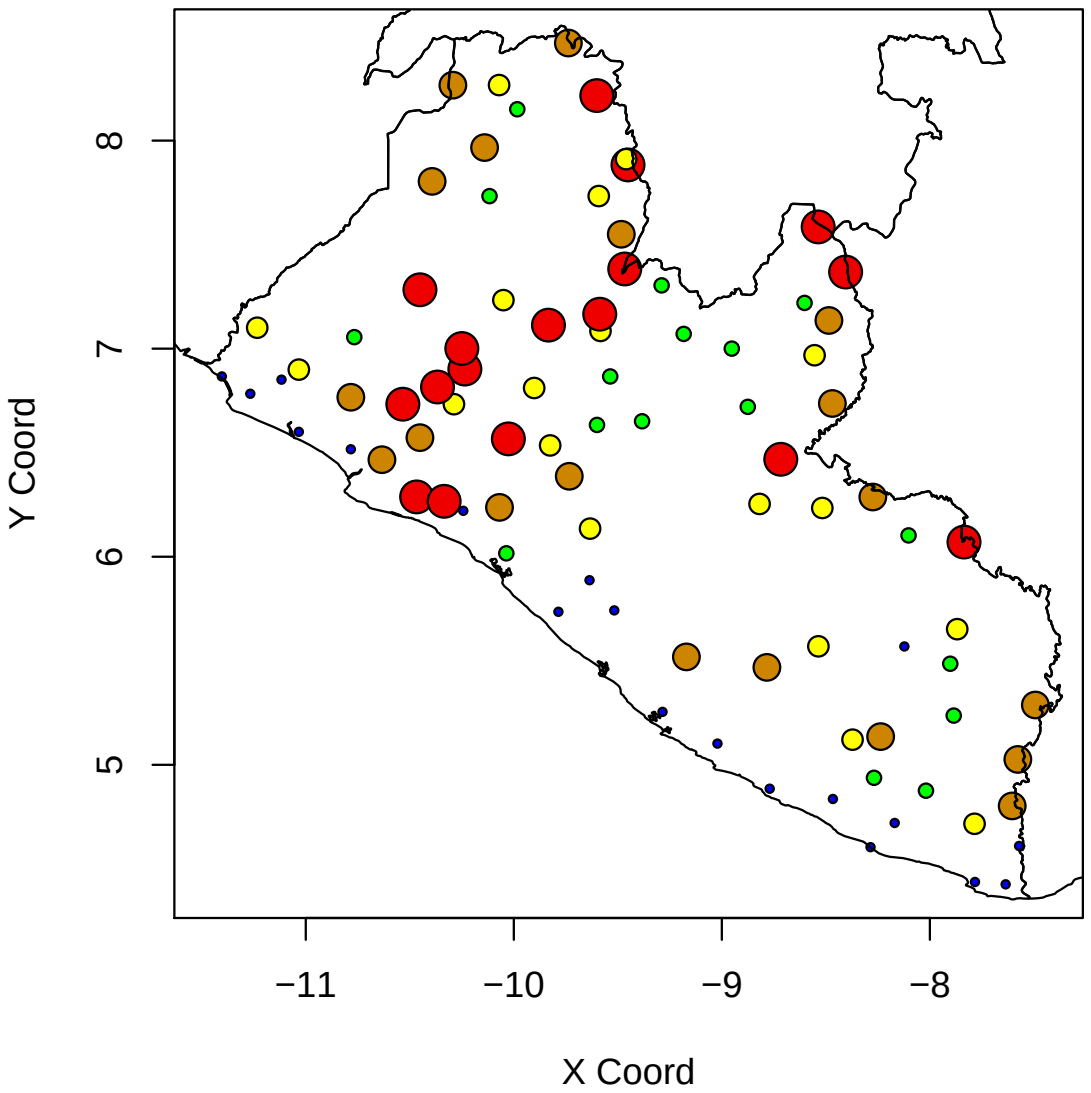
The first written description of the onchocerciasis vector, *Simulium damnosum* **(blackfly)**

Onchocerciasis: aka River Blindness

MDA: a tool for control of vector-borne filarial disease

- **Ivermectin (Mectizan): annual dose clears microfilarial infections of the blood**
- **generally considered safe, with no serious side-effects**
- **mass distribution made possible by donation programme (Merck)**
- **used in multi-national programmes to combat onchocerciasis and lymphatic flariasis**

A prevalence survey data-set: onchocerciasis in Liberia



- **prevalence data from 90 villages in Liberia**
- **sample sizes 40 to 50**
- **empirical prevalences 0% to 35%**

Analysing prevalence data: a geostatistical model

Data: $(x_i, d(x_i), n_i, Y_i) : i = 1, ..., n$

Model:

- $\log[p(x_i)/\{1 - p(x)_i\}] = d(x)_i'\beta + U_i + S(x_i)$
- $U_i \sim \mathrm{N}(0, \nu^2)$, mutually independent
- $S(x) \sim$ Gaussian process, spatially correlated
- $Y_i | U_i, S(x_i) \sim \mathrm{Binomial}\{n_i, p(x_i)\}$

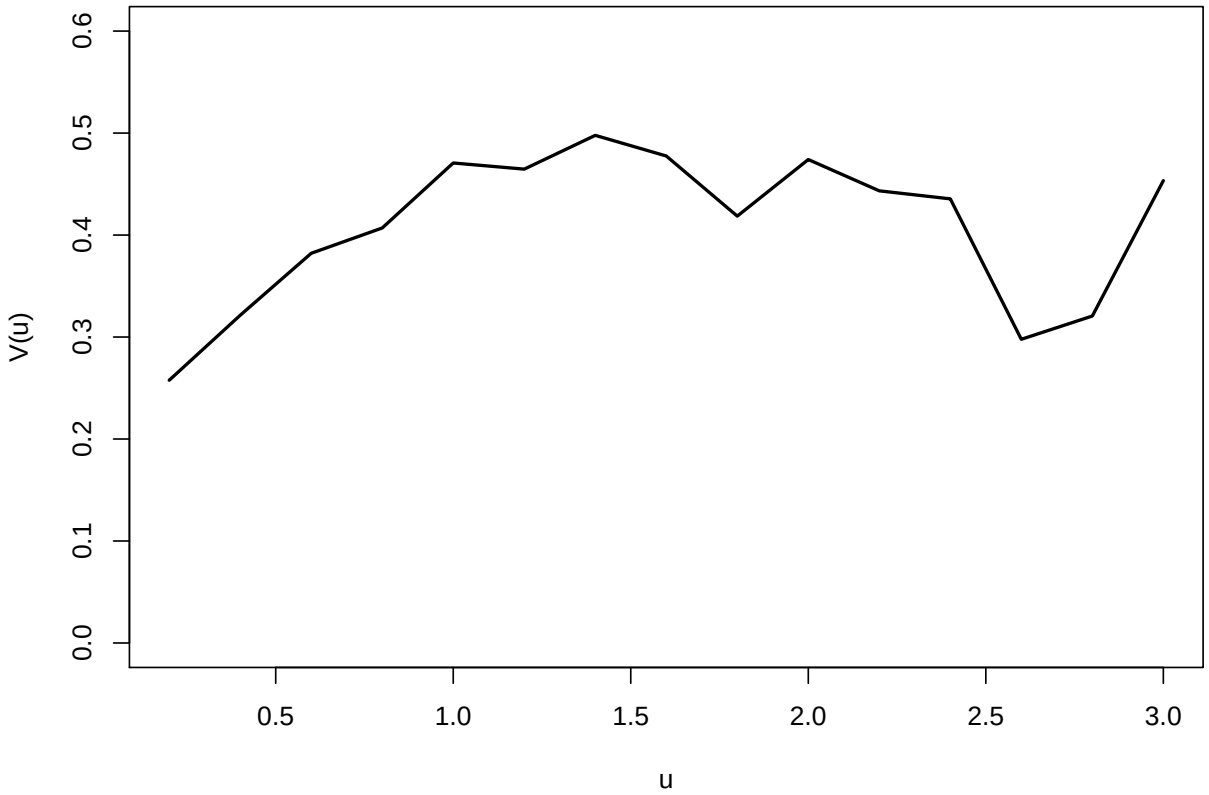
Exploratory analysis: empirical logits

- **fitting the binomial logistic model is computationally demanding, and requires judgement:**
  - **convergence of iterative algorithms**
  - **judicious choice of approximations**
- **empirical logit transform:**

$$Z_i = \log\{(Y_i + 0.5)/(n_i - Y_i + 0.5)\}$$

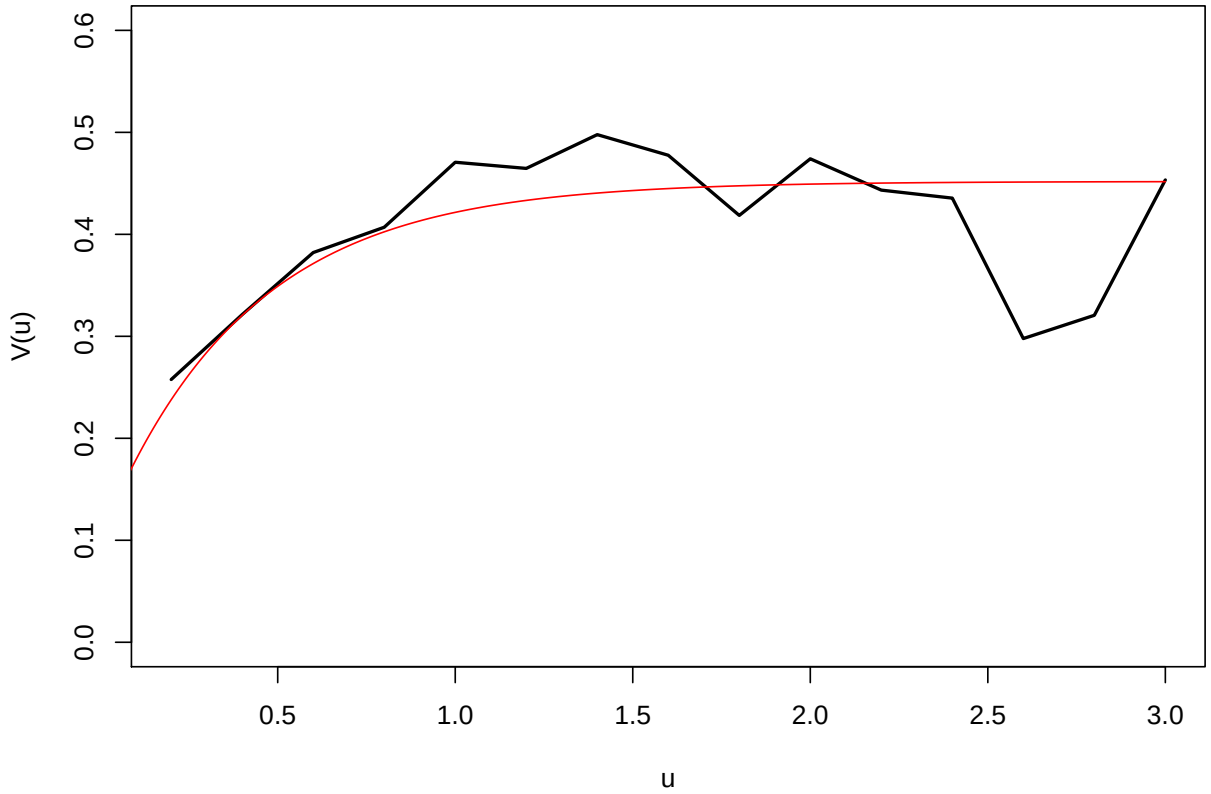
- **fit linear model with Z_i as response**

Exploratory analysis of onchocerciasis data



- **residual variogram after fitting linear trend surface**

Exploratory analysis of onchocerciasis data



- **fitted Matérn model with $\kappa = 0.5$**

Fitting the binomial logistic model

- **likelihood function involves intractable high-dimensional integral**
- **need to use Monte Carlo methods**
- **Monte Carlo maximum likelihood or Bayesian estimation according to choice**
- **for large data-sets, algorithms need careful tuning to preserve accuracy while remaining computationally feasible**

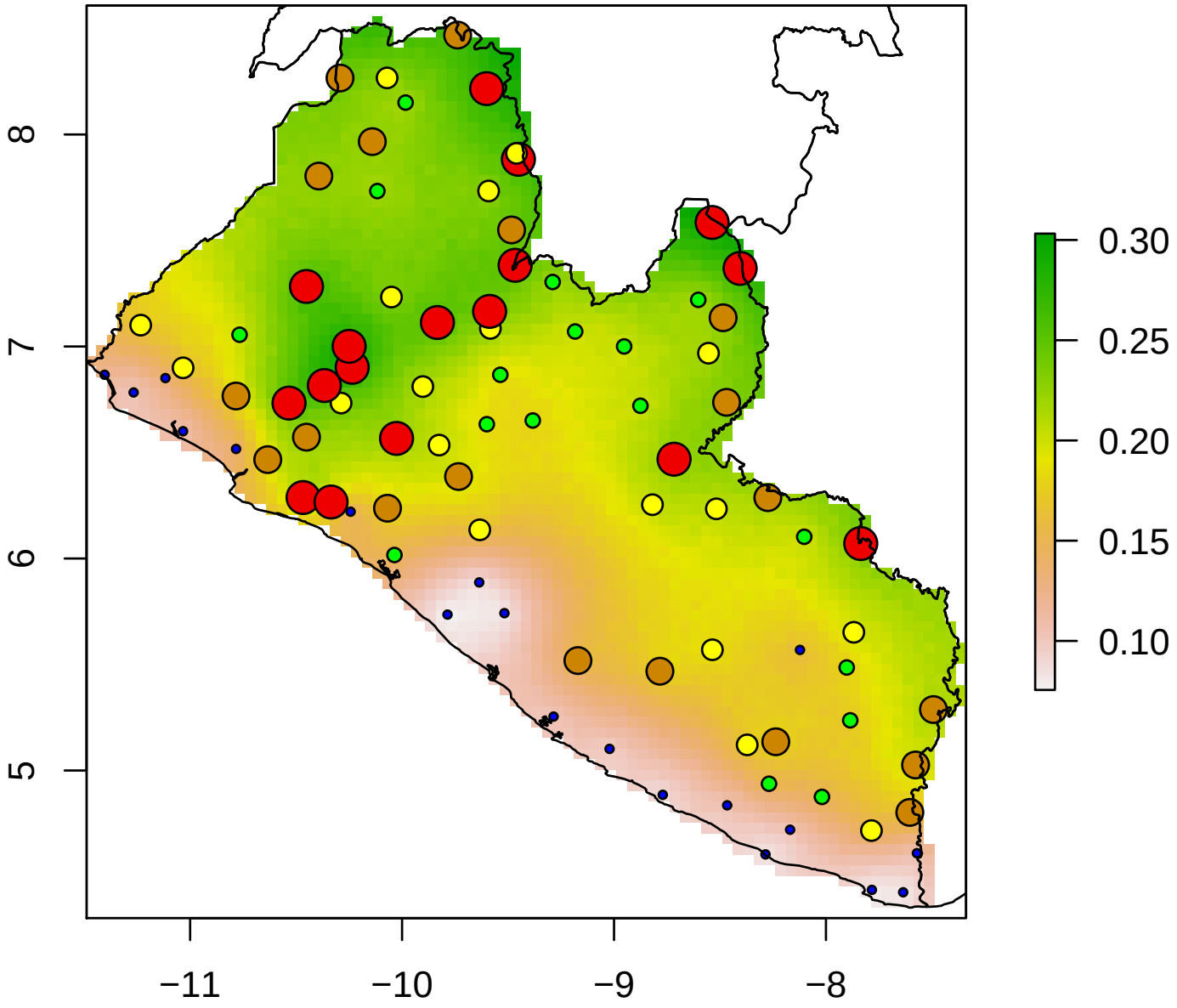
PrevMap: fitting the binomial logistic model

```
library(PrevMap)
response<-cbind(npositive,ntested-npositive)
fit.glm<-glm(response~longitude+latitude,
       data=data,family=binomial)
beta<-fit.glm$coef; par0<-c(beta,fit$cov.pars,fit$nugget)
mcmc<-control.mcmc.MCML(n.sim=10000,burnin=2000,thin=8,
       h=1.65/(nrow(data)**(1/6)))
fit.bl<-binomial.logistic.MCML(npositive~longitude+latitude,
       units.m= ntested,coords=~longitude+latitude,
       data=data,par0=par0,control.mcmc=mcmc,kappa=0.5,
       start.cov.pars=c(par0[5],par0[6]/par0[4]))
```

PrevMap: mapping the results

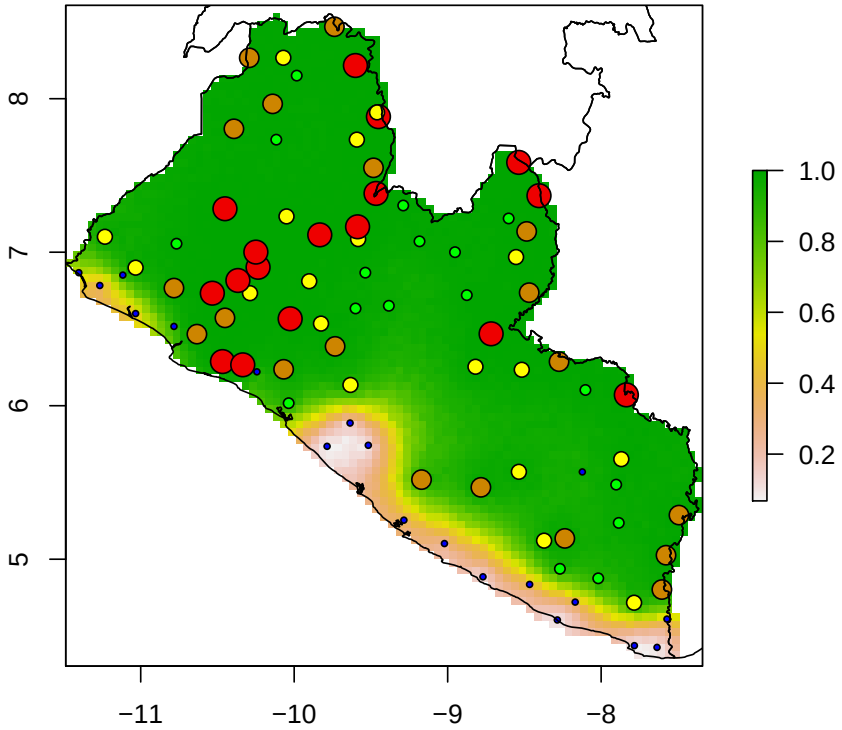
```
library(splancs)
xy<-data.frame(longitude=data$longitude,latitude=data$latitude)
par(pty="s"); pointmap(xy)
poly<-getpoly()
grid.predict<-as.data.frame(gridpts(poly,xs=0.1,ys=0.1))
names(grid.predict)<-c("longitude","latitude")
predict.MCML<spatial.pred.binomial.MCML(fit.bl,
       grid.pred=grid.predict,predictors=grid.predict,
       control.mcmc=mcmc,scale.predictions="prevalence",
       standard.errors=TRUE,thresholds=0.2,
       scale.thresholds="prevalence")
plot(predict.MCML,type="prevalence")
points(gd,add=TRUE)
polymap(poly,add=TRUE)
```

Liberia: onchocerciasis prevalence map

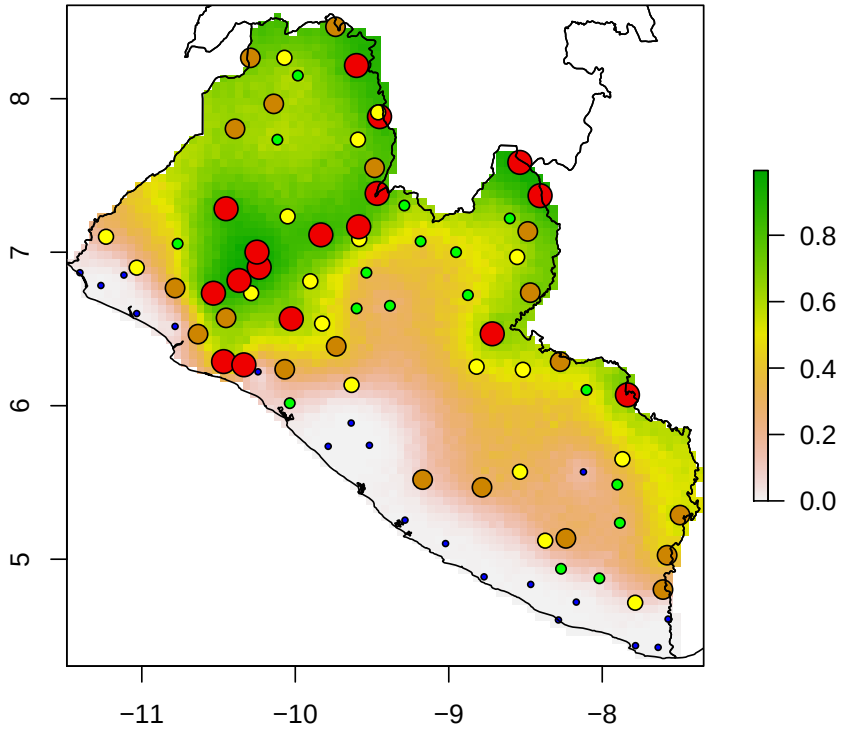


Liberia: onchocerciasis exceedance maps

P(prevalence>10%)



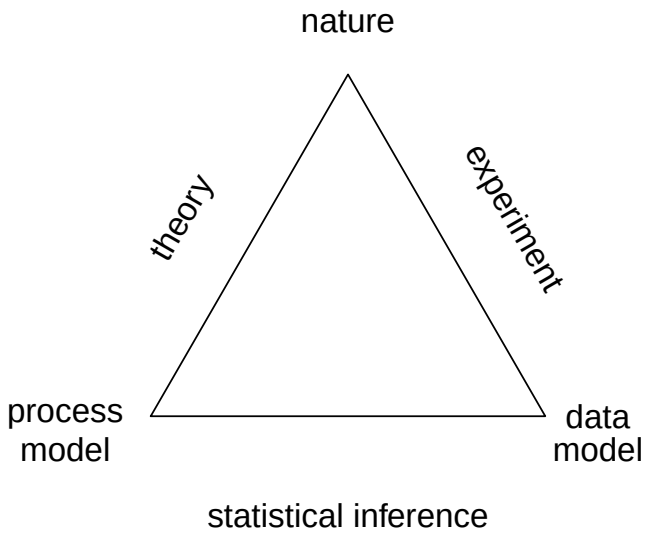
P(prevalence>20%)



Summary

- Statistics is fundamentally about understanding variation
- Design: eliminate extraneous sources of variation
- Model: acknowledge remaining sources of variation
- Inference: estimate model parameters efficiently so that interpretations are:
  - valid (honest)
  - efficient (as precise as possible)
- Design and modelling both involve scientific judgement
- Inference should follow automatically (likelihood-based)
- And the inference should answer the original question

Statistics and scientific method



Statistical method :

- **a device to answer a question**
- **a bridge between theoretical and applied science**
- **a framework to enable principled inference in the presence of uncertainty**

- **Scientific purpose is more important than data-format**
- **Analyse problems, not data**

Diggle, P.J. and Chetwynd, A.G. (2011). *Statistics and Scientific Method: an Introduction for Students and Researchers*. **Oxford: Oxford University Press.**