'David Spiegelhalter is probably the greatest living statistical communicator … Read *The Art of Statistics* and learn. I did'
Tim Harford, author of *The Undercover Economist* and *Messy*

'If you want to develop the skills to see the world as it is, and to tell it how it is – honestly and seriously – this is the book'
Michael Blastland, co-author of *The Tiger That Isn't: Seeing Through a World of Numbers*

'H. G. Wells is frequently quoted as having said that "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." That day has certainly come. This wonderful book provides a non-technical and entertaining introduction to the basic tools of statistical thinking. Wells would have approved'
Professor Sir Adrian Smith FRS, Director, Alan Turing Institute

'Even those with expertise in statistics will find much within these pages to stimulate the mind, since Spiegelhalter combines clarity of thinking with superb communication skills and a wealth of experience. A real tour de force which deserves to be widely read'
Professor Dorothy Bishop, University of Oxford

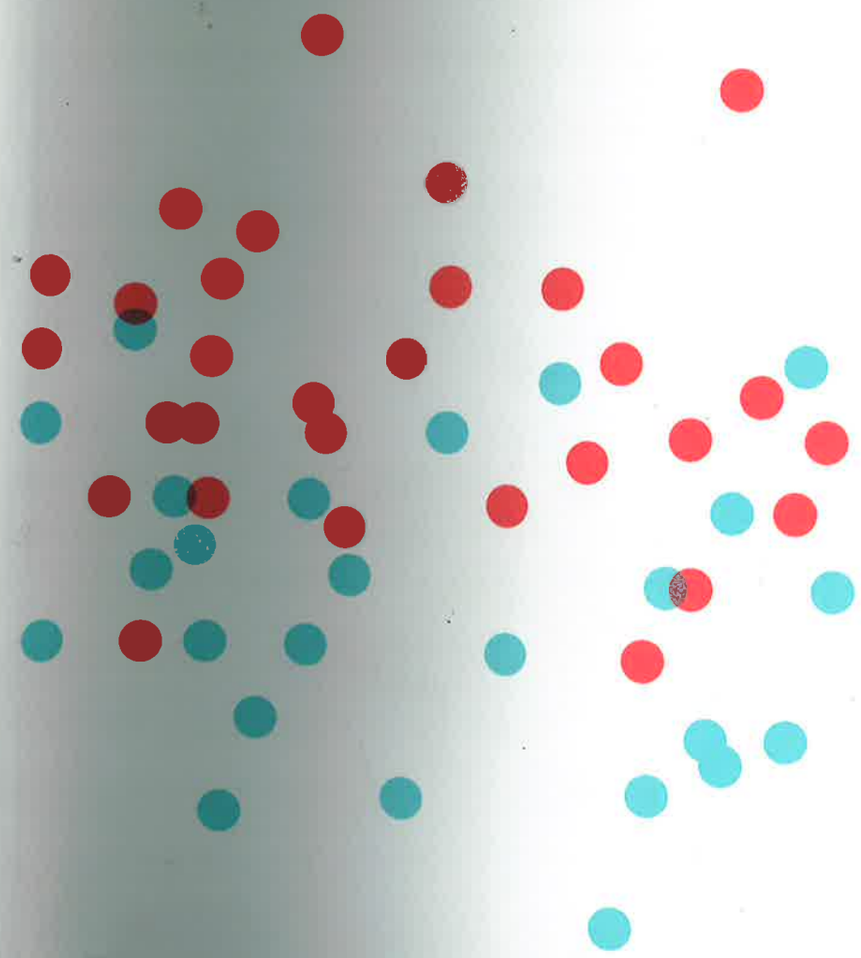
'A wonderfully accessible introduction to modern statistics'
Professor David J. Hand, author of *The Improbability Principle*

ISBN 978-0-241-39863-0
63599
FSC

The Art of Statistics David Spiegelhalter

A PELICAN BOOK

# The Art of Statistics
## Learning from Data
## David Spiegelhalter



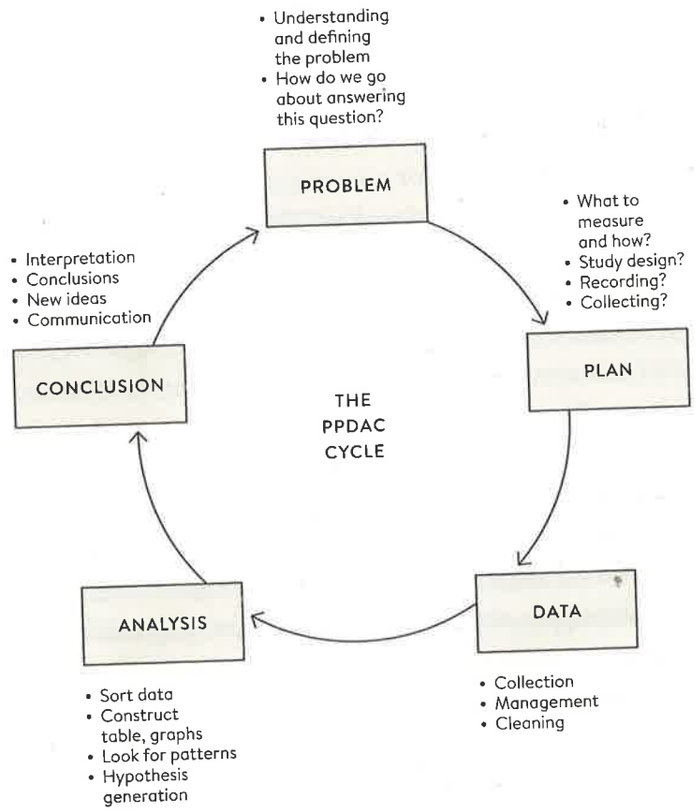



**Figure 0.3**
The PPDAC problem-solving cycle, going from Problem, Plan, Data, Analysis to Conclusion and communication, and starting again on another cycle.





designed. Unfortunately, in the rush to get data and start analysis, attention to design is often glossed over.

Collecting good Data requires the kind of organizational and coding skills that are being seen as increasingly important in data science, particularly as data from routine sources may need a lot of cleaning in order to get it ready to be analysed. Data collection systems may have changed over time, there may be obvious errors, and so on – the phrase 'found data' neatly communicates that it may be rather messy, like something picked up in the street.

The Analysis stage has traditionally been the main emphasis of statistics courses, and we shall cover a range of analytic techniques in this book; but sometimes all that is required is a useful visualization, as in Figure 0.1. Finally, the key to good statistical science is drawing appropriate Conclusions that fully acknowledge the limitations in the evidence, and communicating them clearly, as in the graphical illustrations of the Shipman data. Any conclusions generally raise more questions, and so the cycle starts over again, as when we started looking at th time of day when Shipman's patients died.

Although in practice the PPDAC cycle laid out in Figur 0.3 may not be followed precisely, it underscores that forma techniques for statistical analysis play only one part in th work of a statistician or data scientist. Statistical science i a lot more than a branch of mathematics involving esoter ic formulae with which generations of students have (ofte reluctantly) struggled.

## Summary

- A variety of statistics can be used to summarize the empirical distribution of data-points, including measures of location and spread.
- Skewed data distributions are common, and some summary statistics are very sensitive to outlying values.
- Data summaries always hide some detail, and care is required so that important information is not lost.
- Single sets of numbers can be visualized in strip-charts, box-and-whisker plots and histograms.
- Consider transformations to better reveal patterns, and use the eye to detect patterns, outliers, similarities and clusters.
- Look at pairs of numbers as scatter-plots, and time-series as line-graphs.
- When exploring data, a primary aim is to find factors that explain the overall variation.
- Graphics can be both interactive and animated.
- Infographics highlight interesting features and can guide the viewer through a story, but should be used with awareness of their purpose and their impact.





# Why Are We Looking at Data Anyway? Populations and Measurement

How many sexual partners have people in Britain *really* had?

The last chapter showed some remarkable results from a recent UK survey in which people reported the number of sexual partners they had had in their lifetime. Plotting these responses revealed various features, including a (very) long tail, a tendency to use round numbers such as 10 and 20, and more partners reported by men than women. But the researchers who spent millions of pounds collecting this data were not really interested in what these particular respondents said – after all, they were guaranteed complete anonymity. Their responses were a means to an end, which was to say something about the overall pattern of sexual partnerships in Britain – those of the millions of people who were *not* questioned about their sexual behaviour.

It is no trivial matter to go from the actual responses collected in a survey to conclusions about the whole of Britain. Actually, this is incorrect – it is incredibly easy to just claim that what these respondents say accurately represents what is really going on in the country. Media surveys about sex, where

## Summary

- Inductive inference requires working from our data, through study sample and study population, to a target population.
- Problems and biases can crop up at each stage of this path.
- The best way to proceed from sample to study population is to have drawn a random sample.
- A population can be thought of as a group of individuals, but also as providing the probability distribution for a random observation drawn from that population.
- Populations can be summarized using parameters that mirror the summary statistics of sample data.
- Often data does not arise as a sample from a literal population. When we have all the data there is, then we can imagine it drawn from a metaphorical population of events that could have occurred, but didn't.





# What Causes What?

Does going to university increase the risk of getting a brain tumour?

**Epidemiology** is the study of how and why diseases occur in the population, and Scandinavian countries are an epidemiologist's dream. This is because everyone in those countries has a personal identity number which is used when registering for health care, education, tax, and so on, and this allows researchers to link all these different aspects of people's lives together in a way that would be impossible (and perhaps politically controversial) in other countries.

A typically ambitious study was conducted on over 4 million Swedish men and women whose tax and health records were linked over eighteen years, which enabled the researchers to report that men with a higher socioeconomic position had a slightly increased rate of being diagnosed with a brain tumour. This was one of those worthy but rather unexciting studies that would typically not attract much attention, so a university communications officer thought it would be more interesting to say in a press release that 'High levels of education are linked to heightened brain tumour risk', even though the study was

## Summary

- Uncertainty intervals are an important part of communicating statistics.
- Bootstrapping a sample consists of creating new data sets of the same size by resampling the original data, with replacement.
- Sample statistics calculated from bootstrap re-samples tend towards a normal distribution for larger data sets, regardless of the shape of the original data distribution.
- Uncertainty intervals based on bootstrapping take advantage of modern computer power, do not require assumptions about the mathematical form of the population and do not require complex probability theory.





# Probability – the Language of Uncertainty and Variability

In 1650s France, the self-styled Chevalier de Méré had a gambling problem. It was not that he gambled too much (although he did), but he wanted to know which of two games he stood the greatest chance of winning -

Game 1: Throw a fair die at most four times, and win if you get a six.

Game 2: Throw two fair dice at most twenty-four times, and win if you get a double-six.

Which was his better bet?

Following good empirical statistical principles, the Chevalier de Méré decided to play both games numerous times and see how often he won. This took a great deal of time and effort, but in a bizarre parallel universe in which there were computers but no probability theory, the good Chevalier (real name Antoine Gombaud) would not have wasted his time collecting data on his successes – he would simply have simulated thousands of games.

Figure 8.1 displays the results of such a simulation, showing how the overall proportion of times that he wins each game changes as he 'plays' more and more. Although Game 2 looks the better bet for a while, after around 400 games of

## Summary

- The theory of probability provides a formal language and mathematics for dealing with chance phenomena.
- The implications of probability are not intuitive, but insights can be improved by using the idea of expected frequencies.
- The ideas of probability are useful even when there is no explicit use of a randomizing mechanism.
- Many social phenomena show a remarkable regularity in their overall pattern, while individual events are entirely unpredictable.





# Putting Probability and Statistics Together

*Warning. This is perhaps the most challenging chapter in this book, but persevering with this important topic will give you valuable insights into statistical inference.*

In a random sample of 100 people, we find that 20 are left-handed. What can we say about the proportion of the population who are left-handed?

In the last chapter we discussed the idea of a random variable – a single data-point drawn from a probability distribution described by parameters. But we are seldom interested in just one data-point – we generally have a mass of data which we summarize by determining means, medians and other statistics. The fundamental step we will take in this chapter is to consider those statistics as themselves being random variables, drawn from their own distributions.

This is a big advance, and one that has not only challenged generations of students of statistics, but also generations of statisticians who have tried to work out what distributions we should assume these statistics are drawn from. And given the discussion of the bootstrap in Chapter 7, it

## Summary

- Probability theory can be used to derive the sampling distribution of summary statistics, from which formulae for confidence intervals can be derived.
- A 95% confidence interval is the result of a procedure that, in 95% of cases in which its assumptions are correct, will contain the true parameter value. It cannot be claimed that a specific interval has 95% probability of containing the true value.
- The Central Limit Theorem implies that sample means and other summary statistics can be assumed to have a normal distribution for large samples.
- Margins of error usually do not incorporate systematic error due to non-random causes – external knowledge and judgement is required to assess these.
- Confidence intervals can be calculated even when we observe all the data, which then represent uncertainty about the parameters of an underlying metaphorical population.





# Answering Questions and Claiming Discoveries

Are more boys born than girls?

John Arbuthnot, a doctor who became physician to Queen Anne in 1705, set out to determine the answer to this question. He examined data on London baptisms for the 82 years between 1629 and 1710, and his results are shown in Figure 10.1 in terms of what is now known as the sex ratio, which is the number of boys born per 100 girls.

He found there had been more males than females baptized in every year, with an overall sex ratio of 107, varying between 101 and 116 over the period. But Arbuthnot wanted to claim a more general law, and so argued that if there were really no difference in the underlying rates of boys and girls being born, then each year there would be a 50:50 chance that more boys than girls were born, or more girls than boys, just like flipping a coin.

But to get an excess of boys in every year would then be like flipping a fair coin 82 times in a row, and getting heads every time. The probability of this happening is $1/2^{82}$, which is a very small number indeed, with 24 zeros after the decimal

## Summary

- Tests of null hypotheses – default assumptions about statistical models – form a major part of statistical practice.
- A P-value is a measure of the incompatibility between the observed data and a null hypothesis: formally it is the probability of observing such an extreme result, were the null hypothesis true.
- Traditionally, P-value thresholds of 0.05 and 0.01 have been set to declare 'statistical significance'.
- These thresholds need to be adjusted if multiple tests are conducted, for example on different subsets of the data or multiple outcome measures.
- There is a precise correspondence between confidence intervals and P-values: if, say, the 95% interval excludes 0, we can reject the null hypothesis of 0 at $P < 0.05$.
- Neyman–Pearson theory specifies an alternative hypothesis, and fixes Type I and Type II error rates for the two possible kinds of errors in a hypothesis test.
- Separate forms of hypothesis tests have been developed for sequential testing.
- P-values are often misinterpreted: in particular they do not convey the probability that the null hypothesis is true, nor does a non-significant result imply that the null hypothesis is true.



CHAPTER 11

# Learning from Experience the Bayesian Way

> I am not at all sure that the 'confidence' is not a 'confidence trick'.
>
> — Arthur Bowley, 1934

I must now make an admission on behalf of the statistical community. The formal basis for learning from data is a bit of a mess. Although there have been numerous attempts to produce a single unifying theory of statistical inference, none has been fully accepted. It is no wonder mathematicians tend to dislike teaching statistics.

We have already met the competing ideas of Fisher and Neyman–Pearson, and it is time to explore a third, Bayesian, approach to inference. This has only come to prominence in the last fifty years, but its basic principles go back somewhat further, in fact to the Reverend Thomas Bayes, a Nonconformist minister turned probability theorist and philosopher from Tunbridge Wells, who died in 1761.*

* He died with no knowledge whatsoever of his enduring legacy, and not only was his seminal paper published posthumously in 1763, but his name did not become associated with this approach until the twentieth century.

3. *Plan ahead, really ahead*: This includes the idea of pre-specification in confirmatory experiments – avoiding researcher degrees of freedom
4. *Worry about data quality*: Everything rests on the data.
5. *Statistical analysis is more than a set of computations*: Do not just plug into formulae or run procedures in software, without knowing why you are doing so.
6. *Keep it simple*: The main communication should be as basic as possible – do not show off skills in complex modelling unless they are really necessary.
7. *Provide assessments of variability*: With the warning that margins of error are generally bigger than claimed.
8. *Check your assumptions*: And make clear when this has not been possible.
9. *When possible, replicate!*: Or encourage others to do so.
10. *Make your analysis reproducible*: Others should be able to access your data and code.

Statistical science plays an important role in all our lives, and is constantly changing in response to the increasing quantity and depth of data becoming available. But the study of statistics does not just have an impact on society in general but on individuals in particular. From a purely personal perspective, putting this book together has made me realize how much my life has been enriched by engaging with statistics. I hope that you might feel the same – if not now, then in the future.



# Glossary

**absolute risk**: the proportion of people in a defined group who experience an event of interest within a specified period of time.

**adjustment/stratification**: inclusion into a regression model of known confounders which are not of direct interest, but are intended to allow a more balanced comparison between groups. The hope is that estimated effects associated with explanatory variables of interest should then be closer to causal effects.

**aleatory uncertainty**: unavoidable unpredictability about the future, also known as chance, randomness, luck and so on.

**algorithm**: a rule or formula that takes input variables and produces an output, such as a prediction, a classification, or a probability.

**artificial intelligence** (**AI**): computer programs intended to perform a task normally associated with human abilities.

**ascertainment bias**: when the chance of a person being sampled, or a feature being observed, depends on some background factor, for example when people in the treated arm of a randomized trial get closer supervision than the control group.

**average**: a generic term for a single representative value for a set of numbers, for example the mean, median or mode.

**Bayes factor**: the relative support given by a set of data for two alternative hypotheses. For hypotheses $H_0$ and $H_1$, and data $x$, the ratio is $p(x|H_0)/p(x|H_1)$.

**Bayesian**: the approach to statistical inference in which probability is used not only for aleatory uncertainty, but also epistemic uncertainty about unknown facts. Bayes' theorem is then used to revise these beliefs in the light of new evidence.

**Bayes' theorem**: a rule of probability that shows how evidence $A$ updates prior beliefs of a proposition $B$ to produce posterior beliefs $p(B|A)$, through the formula $p(B|A) = \frac{p(A|B)p(B)}{p(A)}$. This is easily proved: since $p(B \text{ AND } A) = p(A \text{ AND } B)$, the multiplication rule of probability means that $p(B|A)p(A) = p(A|B)p(B)$, and dividing each side by $p(A)$ gives the theorem.

**Bernoulli distribution**: if $X$ is a random variable which takes on the value 1 with probability $p$, and 0 with probability $1-p$, it is known as a Bernoulli trial with a Bernoulli distribution. $X$ has mean $p$ and variance $p(1-p)$.

**bias/variance trade-off**: when fitting a model to be used for prediction, increasing complexity will eventually lead to a model that has less bias, in the sense that it has greater potential to adapt to details of the underlying process, but more variance, since there is not enough data to be confident about the parameters in the model. These elements need to be traded off in order to avoid over-fitting.

**big data**: an increasingly anachronistic phrase sometimes characterized by four Vs: a huge Volume of data, a Variety of sources such as images, social media accounts or transactions, a high Velocity of acquisition, and possible lack of Veracity due to its routine collection.





**binary data**: variables that can only take on two values, often yes/no responses to a question. Can be mathematically represented by a Bernoulli distribution.

**binomial distribution**: when there are $n$ independent possibilities for an event to occur, each with the same probability, the observed number of events has a binomial distribution. Technically for $n$ independent Bernoulli trials $X_1, X_2 \ldots X_n$, each with probability $p$ of success, their sum $R = X_1 + X_2 + \ldots + X_n$, has a binomial distribution with mean $np$ and variance $np(1-p)$, where $P(R=r) = \binom{n}{r} p^r (1-p)^{n-r}$. The observed proportion $R/n$ has mean $p$ and variance $p(1-p)/n$: $R/n$ can therefore be considered as an estimator of $p$, with standard error $\sqrt{p(1-p)/n}$.

**blinding**: when those engaged in a clinical trial do not know what treatment a patient has been given, in order to avoid bias in outcome assessments. Single blinding is when patients do not know what treatment they have been given, double blinding means the people monitoring the patients do not know their treatment, triple blinding is when treatments are labelled say $A$ and $B$, and the statisticians analysing the data and the committee monitoring the results do not know which corresponds to the new treatment.

**Bonferroni correction**: a method for adjusting size (Type I error) or confidence intervals to allow for simultaneous testing of multiple hypotheses. Specifically, when testing $n$ hypotheses, for an overall size (Type I error) of $\alpha$, each hypothesis is tested with size $\alpha/n$. Equivalently, $100(1-\alpha/n)\%$ confidence intervals are quoted for each estimated quantity. For example, when testing 10 hypotheses with an overall $\alpha$ of 5%, then P-values would be compared to 0.05/10 = 0.005, and 99.5% confidence intervals used.

**bootstrapping**: a way of generating confidence intervals and the distribution of test statistics through resampling the observed data rather than through assuming a probability model for the underlying random variable. A basic bootstrap sample of a data set $x_1, x_2 \ldots x_n$ is a sample of size $n$ with replacement, so that the bootstrap sample will be drawn from the original set of distinct values, but not generally in the same proportions as the original data set.

**Brier score**: a measure for the accuracy of probabilistic predictions, based on the mean squared prediction error. If $p_1 \ldots p_n$ are the probabilities given to a set of $n$ binary observations $x_1 \ldots x_n$ taking on values 0 and 1, then the Brier score is $\frac{1}{n}\sum_i^n (x_i - p_i)^2$. Essentially a mean-squared-error criterion applied to binary data.

**calibration**: the requirement for the observed frequencies of events to match those expected by probabilistic predictions. For example, of the occasions when events are given a probability of 0.7, then the events should actually occur roughly 70% of the time.

**case-control study**: a retrospective study design in which people with a disease or outcome of interest (the cases) are matched with one or more people who do not have the disease (the controls), and the histories of the two groups are compared to see whether there are exposures which systematically differ between the two groups. This design can only estimate relative risks associated with exposures.

**categorical variable**: a variable that can take on two or more discrete values, which may or may not be ordered.

**Central Limit Theorem**: the tendency for the sample mean of a set of random variables to have a normal sampling distribution, regardless (with certain exceptions) of the shape of the





underlying sampling distribution of the random variable. If $n$ independent observations each have mean $\mu$ and variance $\sigma^2$, then under broad assumptions their sample mean is an estimator of $\mu$, and has an approximately normal distribution with mean $\mu$, variance $\sigma^2/n$, and standard deviation $\sigma/\sqrt{n}$ (also known as the standard error of the estimator).

**chi-squared test of association / goodness-of-fit test**: a statistical test that indicates the degree of incompatibility of data with an assumed statistical model comprising the null hypothesis, which may be one of lack of association, or some other specfied mathematical form. Specifically, the test compares a set of $m$ observed counts $o_1, o_2 \ldots o_m$ with a set of expected values $e_1, e_2 \ldots e_m$ which have been calculated under the null hypothesis. The simplest version of the test statistic is given as

$$X^2 = \sum_{j=1}^{m} \frac{(o_j - e_j)^2}{e_j}.$$

Under the null hypothesis $X^2$ will have an approximate chi-squared sampling distribution, enabling an associated P-value to be calculated.

**classification tree**: a form of classification algorithm in which features are examined in sequence, with the response indicating the next feature to examine, until a classification is made.

**confidence interval**: an estimated interval within which an unknown parameter may plausibly lie. Based on an observed set of data $x$, a 95% confidence interval for $\mu$ is an interval whose lower limit $L(x)$ and upper limit $U(x)$ has the property that, before observing the data, there is a 95% probability that the random interval $(L(X), U(X))$ contains $\mu$. The Central Limit Theorem, combined with the knowledge that close

to 95% of a normal distribution lies between the mean ± 2 standard deviations, means that a common approximation for a 95% confidence interval is the estimate ± 2 standard errors. Suppose we want to find a confidence interval for the difference $\mu_2 - \mu_1$ between two parameters $\mu_2$ and $\mu_1$. If $T_1$ is an estimator of $\mu_1$ with standard error $SE_1$, and $T_2$ is an estimator of $\mu_2$ with standard error $SE_2$, then $T_2 - T_1$ is an estimator of $\mu_2 - \mu_1$. The variance of the difference between two estimators is the sum of their variances, and so the standard error of $T_2 - T_1$ is given by $\sqrt{SE_1^2 + SE_2^2}$. From this a 95% confidence interval for the difference $\mu_2 - \mu_1$ can be constructed.

**confirmatory studies and analyses**: rigorous studies ideally done to a pre-specified protocol to confirm or negate hypotheses suggested by exploratory studies and analyses.

**confounder**: a variable which is associated with both a response and a predictor, and which may explain some of their apparent relationship. For example, the height and weight of children are strongly correlated, but much of this association is explained by the age of the child.

**continuous variable**: a random variable $X$ that can, at least in principle, take on any value within a specific range. It has a probability density function $f$ such that $P(X \leq x) = \int_{-\infty}^{x} f(t)dt$, and expectation given by $E(X) = \int_{-\infty}^{\infty} xf(x)dx$. The probability of $X$ lying in the interval $(A, B)$ can be calculated using $\int_A^B f(x)\,dx$.

**control group**: a set of individuals who have not been subject to the exposure of interest, say by randomization.

**control limits**: pre-specified limits for a random variable which are used in quality control to monitor deviation from an intended standard, say displayed on a funnel plot.





**count variables**: variables that can take on integer values 0, 1, 2 and so on.

**counter-factual**: a 'what-if' scenario in which an alternative history of events is considered.

**cross-sectional study:** when analysis is based solely on the current state of individuals, without any follow-up over time.

**cross-validation**: a way of assessing the quality of an algorithm for prediction or classification by systematically removing some cases to act as a test set.

**cox regression:** See **hazard ratio**.

**data literacy**: the ability to understand the principles behind learning from data, carry out basic data analyses, and critique the quality of claims made on the basis of data.

**data science**: the study and application of techniques for deriving insights from data, including constructing algorithms for prediction. Traditional statistical science forms part of data science, which also includes a strong element of coding and data management.

**deep learning**: a machine-learning technique that extends standard artificial neural network models to many layers representing different levels of abstraction, say going from individual pixels of an image through to recognition of objects.

**dependent events**: when the probability of one event depends on the outcome of another event.

**dependent, response or outcome variable**: the variable of primary interest that we wish to predict or explain.

**epidemiology**: the study of the rates of, and reasons for, the occurrence of disease.

**epistemic uncertainty**: lack of knowledge about facts, numbers or scientific hypotheses.

**error matrix**: a cross-tabulation of correct and incorrect classifications by an algorithm.

**expectation** (**mean**): the mean-average of a random variable. It is defined as $\sum xp(x)$ for a discrete random variable $X$ and $\int xp(x)\,dx$ for a continuous random variable. For example, if $X$ is the result of throwing a fair die, then $P(X = x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5, 6$, so that $E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$.

**expected frequencies**: the numbers of events expected to occur in the future, according to an assumed probability model.

**exploratory studies and analyses**: initial flexible studies which allow adaptive changes to design and analyses in order to pursue promising leads, and are intended to generate hypotheses to be tested in confirmatory studies.

**exposure**: a factor whose impact on a disease, death or other medical outcome is of interest, such as an aspect of the environment or behaviour.

**external validity**: when the conclusions of a study can be generalized to a target group, wider than the immediate population that has been studied. This addresses the relevance of a study.

**false discovery rate**: when testing multiple hypotheses, the proportion of positive claims that turn out to be false-positives.

**false-positive**: an incorrect classification of a 'negative' case as a 'positive' case.

**feature engineering**: in machine learning, the process of reducing the dimensionality of input variables, creating summary measures intended to encapsulate the information in the whole data.

**forensic epidemiology**: using knowledge about the causes of disease in populations when making judgements about the causes of disease in individuals.





**framing**: the choice of how to express numbers, which in turn can influence the impression given to audiences.

**funnel plot**: a plot of a set of observations from different units against a measure of their precision, where units might be institutions, areas or studies. Often two 'funnels' indicate where we would expect 95% and 99.8% of observations to lie, were there really no underlying differences between the units. When the distribution of the observations is approximately normal, the 95% and 99.8% control limits are essentially the mean ± two and three standard errors.

**hazard ratio**: when analysing survival times, the relative risk, associated with an exposure, of suffering an event in a fixed period of time. A Cox regression is a form of multiple regression when the response variable is a survival time, and the coefficients correspond to log(hazard ratios).

**hierarchical modelling**: in Bayesian analysis, when the parameters underlying a number of units, say areas or schools, are themselves assumed to be drawn from a common prior distribution. This results in shrinkage of the parameter estimates for individual units towards an overall mean.

**hypergeometric distribution**: the probability of $k$ successes in $n$ draws, without replacement, from a finite population of size $N$ that contains exactly $K$ objects with that feature, formally given by

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

**hypothesis testing**: a formal procedure for evaluating the support for hypotheses provided by data, generally an amalgam of classic Fisherian tests of a null hypothesis

using a P-value, and the Neyman–Pearson structure of null and alternative hypotheses and Type I and Type II errors.

**icon arrays**: a graphic display of frequencies using a set of small images, say of people.

**independent events**: $A$ and $B$ are independent if the occurrence of $A$ does not influence the probability of $B$, so that $p(B|A) = p(B)$, or equivalently $p(B, A) = p(B)p(A)$.

**independent variable / predictor**: a variable that is fixed by design or observation, and whose association with an outcome variable may be of interest.

**induction / inductive inference**: the process of learning about general principles from specific examples.

**inductive behaviour**: a proposal by Jerzy Neyman and Egon Pearson in the 1930s to frame hypothesis testing in terms of decision-making. The ideas of size, power and Type I and Type II errors are remnants.

**intention to treat**: the principle by which participants in randomized trials are analysed according to whatever intervention they were supposed to get, whether or not they actually received it.

**interactions**: when multiple explanatory variables combine to produce an effect different from that expected from their individual contributions.

**internal validity**: when the conclusions of a study truly apply to the population of a study. This addresses the rigour with which a study has been conducted.

**inter-quartile range**: a measure of the spread of a sample or a population distribution, specifically the distance between the 25th and 75th percentiles. Equivalent to the difference between the 1st and 3rd quartiles.





**Law of Large Numbers**: the process by which the sample mean of a set of random variables tends towards the population mean.

**least-squares**: suppose we have a set of $n$ paired numbers, $(x_1, y_1)$, $(x_2, y_2)$ . . . $(x_n, y_n)$, and $\overline{x}$, $s_x$ are the sample mean and standard deviation of the $x$s, and $\overline{y}$, $s_y$ are the sample mean and standard deviation of the $y$s. Then the least-squares regression line is given by

$$\hat{y} = b_0 + b_1\,(x - \overline{x}),$$

where

- $\hat{y}$ is the predicted value for the dependent variable $y$ for a specified value of the independent variable $x$.
- The gradient is $b_1 = \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.
- The intercept is $b_0 = \overline{y}$. The least-squares line goes through the centre of gravity $\overline{x}$, $\overline{y}$.
- The $i$th residual is the difference between the $i$th observation and its predicted value, $y_i - \hat{y}_i$.
- The adjusted value of the $i$th observation is the residual added to the intercept, i.e., $y_i - \hat{y}_i + \overline{y}$. It is intended to be the value we would have observed were this an 'average' case, that is with $x = \overline{x}$ rather than $x = x_i$.
- The residual sum of squares (RSS) is the sum of the squares of the residuals, so that $\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. The least-squares line is defined as the line that minimizes the residual sum of squares.
- The gradient $b_1$ and Pearson's correlation coefficient $r$ are related through the formula $b_1 = rs_y/s_x$. So if the standard deviations of the $x$s and $y$s are the same, then the gradient is exactly equal to the correlation coefficient.

**likelihood**: a measure of the evidential support provided by data for particular parameter values. When a probability distribution for a random variable depends on a parameter, say $\theta$, then after observing data $x$ the likelihood for $\theta$ is proportional to $p(x|\theta)$.

**likelihood ratio**: a measure of the relative support that some data provides for two competing hypotheses. For hypotheses $H_0$ and $H_1$, the likelihood ratio provided by data $x$ is given by $p(x|H_0)/p(x|H_1)$.

**logarithmic scale**: The logarithm to base 10 of a positive number $x$ is denoted by $y = \log_{10} x$, or equivalently $x = 10^y$. In statistical analysis, $\log x$ generally denotes the natural logarithm $y = \log_e x$, or equivalently $x = e^y$ where $e$ is the exponential constant 2.718.

**logistic regression**: a form of multiple regression when the response variable is a proportion, and the coefficients correspond to log(odds ratios). Suppose we observe a series of proportions $y_i = r_i/n_i$, assumed to arise from a binomial variable with underlying probability $p_i$ with a corresponding set of predictor variables $(x_{i1}, x_{i2} \ldots x_{ip})$. The logarithm of the odds of the estimated probability $\hat{p}_i$ is assumed to be a linear regression:

$$\log \frac{\hat{p}_i}{(1-\hat{p}_i)} = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}.$$

Suppose one of the predictor variables, say $x_1$, is binary with $x_1 = 0$ corresponding to not being exposed to a potential hazard, and $x_1 = 1$ corresponding to being exposed. Then the coefficient $b_1$ is a log (odds ratio).

**lurking factor**: in epidemiology, an exposure that has not been measured but may be a confounder responsible for some of the observed association: for example, when socioeconomic





status has not been measured in a study relating diet with disease.

**machine learning**: procedures for extracting algorithms, say for classification, prediction or clustering, from complex data.

**margin of error**: after a survey, a plausible range in which a true characteristic of a population may lie. These are generally 95% confidence intervals, which are approximately ± 2 standard errors, but sometimes error-bars are used to represent ± 1 standard error.

**mean (of a population)**: *see* **expectation**

**mean (of a sample)**: suppose we have a set of $n$ data-points, which we label as $x_1, x_2, \ldots, x_n$. Then their sample mean is given by $m = (x_1 + x_2 + \ldots + x_n)/n$, which can be written as $m = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$. For example, if 3, 2, 1, 0, 1 are the numbers of children reported by 5 people in a sample, then the sample mean is $(3 + 2 + 1 + 0 + 1)/5 = 7/5 = 1.4$.

**mean-squared-error (MSE)**: a measure of performance when predictions $t_1 \ldots t_n$ are made of observations $x_1 \ldots x_n$, given by $\frac{1}{n}\sum_{i=1}^{n}(x_i - t_i)^2$

**median (of a sample)**: the value mid-way along the ordered set of data-points. If the data-points are put in order, we denote the lowest by $x_{(1)}$, the second lowest by $x_{(2)}$, and so on until the maximum value $x_{(n)}$. If $n$ is odd, then the sample median is the middle value $x_{\left(\frac{n+1}{2}\right)}$; if $n$ is even, then the average of the two 'middle' points is taken as the median.

**meta-analysis**: a formal statistical method for combining the results from multiple studies.

**mode (of a population distribution)**: the response with the maximum probability of occurring.

**mode (of a sample)**: the most common value in a set of data.

**multi-level regression and post-stratification** (**MRP**): a modern development in survey sampling in which fairly small numbers of responders are obtained from many areas. A regression model is then built relating responses to demographic factors, allowing for additional between-area variability using hierarchical modelling. Knowing the demographics of all areas then allows both local and national predictions to be made, with appropriate uncertainty.

**multiple linear regression**: suppose that for every response $y_i$ there are a set of $p$ predictor variables $(x_{i1}, x_{i2} \ldots x_{ip})$. Then a least-squares multiple linear regression is given by

$$\hat{y}_i = b_0 + b_1(x_{i1} - \overline{x}_1) + b_2(x_{i2} - \overline{x}_2) + \ldots + b_p(x_{ip} - \overline{x}_p),$$

where the coefficients $b_0, b_1 \ldots b_p$ are chosen to minimize the residual sum of squares RSS $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. The intercept $b_0$ is simply the mean $\overline{y}$, and the formula for the remaining coefficients is complex but easily computed. Note that $b_0 = \overline{y}$ is the predicted value of an observation $y$ whose predictor variables were the averages $(\overline{x}_1, \overline{x}_2 \ldots \overline{x}_p)$, and, just as for a linear regression, an adjusted $y_i$ is given by the residual plus the intercept, or $y_i - \hat{y}_i + \overline{y}$.

**multiple testing**: when a series of hypothesis tests are carried out, so increasing the chance of at least one false-positive claim (Type 1 error).

**normal distribution**: $X$ has a normal (Gaussian) distribution with mean $\mu$ and variance $\sigma^2$ if it has a probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty \le x \le \infty.$$





Then $E(X) = \mu$, $V(X) = \sigma^2$, $SD(X) = \sigma$. The standardized variable $Z = \frac{X-\mu}{\sigma}$ has mean 0 and variance 1, and is said to have a standard normal distribution. We write $\Phi$ for the cumulative probability of a standard normal variable $Z$.

For example, $\Phi(-1) = 0.16$ is the probability of a standard normal variable being less than $-1$, or equivalently, the probability of a general normal variable being less then one standard deviation below the mean. The $100p\%$ percentile of the standard normal distribution is $z_p$ where $P(Z \le z_p) = p$. Values of $\Phi$ are available in standard software or tables, as are percentage points $z_p$: for example, the 75th percentile of the standard normal distribution is $z_{0.75} = 0.67$.

**null hypothesis**: a default scientific theory, generally representing the absence of an effect or a finding of interest, which is tested using a P-value. Generally denoted $H_0$.

**objective priors**: an attempt to remove the subjective element in Bayesian analysis, by pre-specifying prior distributions that are intended to represent ignorance about parameters, and so let the data speak for itself. No overall procedure for setting such priors has been established.

**odds, odds ratios**: if the probability of an event is $p$, the odds of the event is defined by $\frac{p}{(1-p)}$. If the odds of an event in the exposed group is $\frac{p}{(1-p)}$, and the odds in the non-exposed group is $\frac{q}{(1-q)}$: the odds ratio is then given by $\frac{p}{(1-p)} / \frac{q}{(1-q)}$. If $p$ and $q$ are small, then the odds ratio will be close to the relative risk $p/q$, but odds ratios and relative risks start to differ when the absolute risks are much more than 20%.

**one-sided and two-sided tests**: a one-sided hypothesis test is used when a null hypothesis specifies that, say, the effect of

a medical treatment is negative. This would only be rejected by large positive values of a test statistic representing an estimated treatment effect. A two-sided test would be appropriate for a null hypothesis that a treatment effect, say, is exactly zero, and so both positive and negative estimates would lead to the null being rejected.

**one-tailed and two-tailed P-values**: those corresponding to one-sided and two-sided tests.

**over-fitting**: building a statistical model that is over-adapted to training data, so that its predictive ability starts to decline.

**parameters**: the unknown quantities in a statistical model, generally denoted with Greek letters.

**Pearson correlation coefficient**: for a set of $n$ paired numbers, $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, when $\overline{x}$, $s_x$ are the sample mean and standard deviation of the $x$s, and $\overline{y}$, $s_y$ are the sample mean and standard deviation of the $y$s, the Pearson correlation coefficient is given by

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

Suppose $x$s and $y$s have both been standardized to Z-scores given by $u$s and $v$s respectively, so that $u_i = (x_i - \overline{x})/s_x$, and $v_i = (y_i - \overline{y})/s_y$. Then the Pearson correlation coefficient can be expressed as $\sum_{i=1}^{n} u_i v_i$, that is the 'cross-product' of the Z-scores.

**percentile (of a population)**: there is, for example, a 70% chance of drawing a random observation below the 70th percentile. For a literal population, it is the value below which 70% of the population lie.

**percentile (of a sample)**: the 70th percentile of a sample, for example, is the value that is 70% along the ordered data set:





the median is therefore the 50th percentile. Interpolation between points may be necessary.

**permutation/randomization test**: a form of hypothesis test in which the distribution of the test statistic under the null hypothesis is obtained by permuting the labels of the data, rather than through a detailed statistical model for the random variables. Suppose the null hypothesis is that a 'label', say being male or female, is not associated with an outcome. Randomization tests examine all possible ways in which labels for individual data-points can be rearranged, each of which are equally likely under the null hypothesis. The test statistic for each of these permutations is calculated, and the P-value is given by the proportion that lead to more extreme test statistics than that actually observed.

**placebo**: a dummy treatment given to the control arm of a randomized clinical trial, such as a sugar pill disguised to look like the treatment being tested.

**Poisson distribution**: a distribution for a count random variable $X$ for which $P(X = x|\mu) = e^{-\mu}\frac{\mu^x}{x!}$ for $x = 0, 1, 2 \ldots$ Then $E(X) = \mu$ and $V(X) = \mu$.

**population**: a group from which it is assumed your sample data are drawn, and which provides the probability distribution for a single observation. In a survey this may be a literal population, but when making measurements, or when having all possible data, the population becomes a mathematical idealization.

**population distribution**: when the population literally exists, the pattern of potential observations in the entire population. It also refers to the probability distribution of a generic random variable.

**posterior distribution**: in Bayesian analysis, the probability distribution of unknown parameters after taking into account observed data through Bayes' theorem.

**power of a test**: the probability of correctly rejecting the null hypothesis, given the alternative hypothesis is true. It is one minus the Type II error rate of a statistical test, and is generally denoted by $1 - \beta$.

**PPDAC**: a proposed structure for the 'data cycle', comprising Problem, Plan, Data collection, Analysis (exploratory or confirmatory) and Conclusions and communication.

**practical significance**: when a finding is of genuine importance. Large studies may give rise to results that are statistically but not practically significant.

**predictive analytics**: using data to create algorithms for making predictions.

**prior distribution**: in Bayesian analysis, the initial probability distribution for the unknown parameters. After observing data, it is revised to the posterior distribution using Bayes' theorem.

**probabilistic forecast**: a prediction in the form of a probability distribution for a future event, rather than a categorical judgement of what will happen.

**probability**: the formal mathematical expression of uncertainty. Let $P(A)$ be the probability for an event $A$. Then the rules of probability are:

1. Bounds: $0 \leq P(A) \leq 1$, with $P(A) = 0$ if $A$ is impossible and $P(A) = 1$ if $A$ is certain.
2. Complement: $P(A) = 1 - P(\text{NOT } A)$.
3. Addition rule: If $A$ and $B$ are mutually exclusive (i.e., one at most can occur), $P(A \text{ OR } B) = P(A) + P(B)$.





4. Multiplication rule: For any events $A$ and $B$, $P(A \text{ AND } B) = P(A|B)P(B)$, where $P(A|B)$ represents the probability for $A$ given $B$ has occurred. $A$ and $B$ are independent if and only if $P(A|B) = P(A)$, i.e., the occurrence of $B$ does not affect the probability for $A$. In this case we have $P(A \text{ AND } B) = P(A)P(B)$, the multiplication rule for independent events.

**probability distribution**: a generic term for a mathematical expression of the chance of a random variable taking on particular values. A random variable $X$ has a probability distribution function defined by $F(x) = P(X \leq x)$, for all $-\infty < x < \infty$, i.e., the probability that $X$ is at most $x$.

**prosecutor's fallacy**: when a small probability of the evidence, given innocence, is mistakenly interpreted as the probability of innocence, given the evidence.

**prospective cohort study**: when a set of individuals are identified, background factors measured, and then they are followed up and relevant outcomes observed. Such studies are lengthy and expensive, and may not identify many rare events.

**P-value**: a measure of discrepancy between data and a null hypothesis. For a null hypothesis $H_0$, let $T$ be a statistic for which large values indicate inconsistency with $H_0$. Suppose we observe a value $t$. Then a (one-sided) P-value is the probability of observing such an extreme value, were $H_0$ true, that is $P(T \geq t|H_0)$. If both small and large values of $T$ indicate inconsistency with $H_0$, then the two-sided P-value is the probability of observing such a large value in either direction. Often the two-sided P-value is simply taken as double the one-sided P-value, while the R software uses the total probability of events which have a lower probability of occurring than that actually observed.

**quartiles (of a population)**: the 25th, 50th and 75th percentiles.

**randomized controlled trial (RCT)**: an experimental design in which people or other units being tested are randomly allocated to different interventions, thus ensuring, up to the play of chance, that the groups are balanced in both known and unknown background factors. If the groups show subsequent differences in outcome, then either the effect must be due to the intervention or a surprising event has occurred, whose probability can be expressed as a P-value.

**random match probability**: in forensic DNA testing, the probability that a person randomly drawn from a relevant population would match the observed DNA profile that connects a suspect with a crime.

**random variable**: a quantity assumed to have a probability distribution. Before they are observed, random variables are generally given a capital letter such as $X$, while observed values are denoted $x$.

**range (of a sample)**: the maximum minus the minimum, denoted $x_{(n)} - x_{(1)}$.

**rate ratio**: the relative increase in the expected number of events in a fixed period of time associated with an exposure. A Poisson regression is a form of multiple regression when the response variable is the observed rate, and the coefficients correspond to log(rate ratios).

**Receiver Operating Characteristic (ROC) curve**: for an algorithm that generates a score, we can choose a particular threshold for the score above which a unit is classified as 'positive'. As this threshold varies, the ROC curve is formed by plotting the resulting sensitivity (true-positive rate) on the $y$-axis versus one minus specificity (false-positive rate) on the $x$-axis.





**regression coefficient**: an estimated parameter in a statistical model, that expresses the strength of relationship between an explanatory variable and an outcome in multiple regression analysis. The coefficient will have a different interpretation depending on whether the outcome variable is a continuous variable (multiple linear regression), a proportion (logistic regression), a count (Poisson regression) or a survival time (Cox regression).

**regression to the mean**: when a high or low observation is followed by one that is less extreme, through the process of natural variation. It occurs because part of the reason for the initial extreme case was chance, and this is unlikely to repeat to the same extent.

**relative risk**: if the absolute risk among people who are exposed to something of interest is $p$, and the absolute risk among people who are not exposed is $q$, then the relative risk is $p/q$.

**reproducibility crisis**: the claim that many published scientific findings are based on work of insufficient quality, so that the results fail to be reproduced by other researchers.

**residual**: the difference between an observed value and that predicted by a statistical model.

**residual error**: the generic term for the component of the data that cannot be explained by a statistical model, and so is said to be due to chance variation.

**retrospective cohort study**: when a set of individuals are identified at a point in the past, and their subsequent outcomes traced up to the present day. Such a study does not require an extended period of follow-up, but is dependent on the appropriate explanatory variables having been measured in the past.

**reverse causation**: when an association between two variables initially appears to be causal, but could in fact be acting in the opposite direction. For example, people who do not drink alcohol tend to have poorer health outcomes than moderate drinkers, but this is at least partly due to some non-drinkers having given up alcohol due to poor health.

**sample distribution**: the pattern made by a set of numerical or categorical observations. Also known as the empirical or data distribution.

**sample mean**: *see* **mean (of a sample)**

**sampling distribution**: the probability distribution of a statistic.

**sensitivity**: the proportion of 'positive' cases that are correctly identified by a classifier or test, often termed the true-positive rate. One minus sensitivity is also known as the observed Type II error or false-negative rate.

**sequential testing**: when a statistical test is repeatedly carried out on accumulating data, thus inflating the chance of a Type I error occurring at some point. A 'significant result' is guaranteed if the process is continued for long enough.

**shrinkage**: the influence of a prior distribution in Bayesian analysis, in which an estimate tends to be pulled towards either an assumed or an estimated prior mean. This is also known as 'borrowing strength', since, say, estimated rates of disease in a specific geographical area are influenced by rates in other areas.

**signal and the noise**: the idea that observed data arises from two components: a deterministic signal which we are really interested in, and random noise that comprises the residual error. The challenge of statistical inference is to appropriately identify the two, and not be misled into thinking that noise is actually a signal.





**Simpson's paradox**: when an apparent relationship reverses its sign when a confounding variable is taken into account.

**size of a test**: the Type I error rate of a statistical test, generally denoted by $\alpha$.

**skewed distribution**: when a sample or population distribution is highly asymmetric, and has a long left- or right-hand tail. This might typically occur for variables such as income and sales of books, when there is extreme inequality. Standard measures (such as means) and standard deviations can be very misleading for such distributions.

**Spearman's rank correlation**: the rank of an observation is its position in the ordered set, where 'ties' are considered to have the same rank. For example, for the data (3, 2, 1, 0, 1) the ranks are (5, 4, 2.5, 1, 2.5). Spearman's rank correlation is simply the Pearson's correlation when the $x$s and $y$s are replaced by their respective ranks.

**specificity**: the proportion of 'negative' cases that are correctly identified by a classifier or test. One minus specificity is also known as the observed Type I error, or false-positive rate.

**standard deviation**: the square root of the variance of a sample or distribution. For well-behaved, reasonably symmetric data distributions without long tails, we would expect most of the observations to lie within two sample standard deviations from the sample mean.

**standard error**: the standard deviation of a sample mean, when considered as a random variable. Suppose $X_1, X_2 \ldots X_n$ are independent and identically distributed random variables drawn from a population distribution with mean $\mu$ and standard deviation $\sigma$. Then their average $Y = (X_1 + X_2 + \ldots + X_n)/n$ has mean $\mu$ and variance $\sigma^2/n$. The standard deviation of $Y$ is

$\sigma/\sqrt{n}$, known as the standard error, and estimated by $s/\sqrt{n}$, where $s$ is the sample standard deviation of the observed $X$'s.

**statistic**: a meaningful number derived from a set of data.

**statistical inference**: the process of using sample data to learn about unknown parameters underlying a statistical model.

**statistical model**: a mathematical representation, containing unknown parameters, of the probability distribution of a set of random variables.

**statistical science**: the discipline of learning about the world from data, typically involving a problem-solving cycle such as PPDAC.

**statistical significance**: an observed effect is judged to be statistically significant when its P-value corresponding to a null hypothesis is less than some pre-specified level, say 0.05 or 0.001, meaning such an extreme result was unlikely to occur were the null hypothesis, and all other modelling assumptions, to hold.

**supervised learning**: construction of a classification algorithm based on cases with confirmed membership of classes.

***t*-statistic**: a test statistic used to test a null hypothesis of a parameter being zero, formed by the ratio of an estimate to its standard error. For large samples, values of above 2 or below −2 correspond to a two-sided P-value of 0.05; exact P-values can be obtained from statistical software.

**Type I error**: when a true null hypothesis is incorrectly rejected in favour of an alternative, so a false-positive claim is made.





**Type II error**: when an alternative hypothesis is true, but a hypothesis test does not reject the null hypothesis, so the conclusion is a false-negative.

**unsupervised learning**: identification of classes based on cases with no identified membership, using some form of clustering procedure.

**variability**: the inevitable differences that occur between measurements or observations, some of which may be explained by known factors, and the remainder attributed to random noise.

**variance**: for a sample $x_1 \ldots x_n$ with mean $\bar{x}$, this is generally defined as $s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$ (although the denominator can also be $n$ rather than $n - 1$). For a random variable $X$ with mean $\mu$, the variance is $V(X) = E(X - \mu)^2$. The standard deviation is the square root of the variance, so $SD(X) = \sqrt{V(X)}$.

**wisdom of crowds**: the idea that a summary derived from a group opinion is closer to the truth than the majority of the individuals.

**Z-score**: a means of standardizing an observation $x_i$ in terms of its distance from the sample mean $m$ expressed in terms of sample standard deviations $s$, so that $z_i = (x_i - m)/s$. An observation with a Z-score of 3 corresponds to being 3 standard deviations above the mean, which is a fairly extreme outlier. A Z-score can also be defined in terms of a population mean $\mu$ and standard deviation $\sigma$, in which case $z_i = (x_i - \mu)/\sigma$.