Chapter 1

Data Analysis and Medical Statistics

Andrew Biffen and David Ashton-Cleary

Learning Objectives

- To acquire an understanding of how statistical considerations shape design and implementation of clinical studies
- To be able to define a specific research question with due consideration to potential bias, ethical considerations and appropriate randomization
- To recognize different types of data and select appropriate measures of central tendency and spread which describe the data
- To understand probability theory and select and use appropriate tests of statistical significance
- To gain an overall understanding of the clinical application of statistics, its strengths and its limitations, and to critique statistical analyses of academic papers

Chapter Content

Scenario

You are anaesthetizing for a list of total knee replacements and wonder whether there is evidence as to whether femoral nerve block makes any difference to post-operative pain. You conduct a literature review and turn up some studies.

How would you approach evaluation of the evidence? What makes a good trial? How would you design a trial yourself if you found the evidence to be insufficient to answer your question?

Introduction

The first medical statistics book published in Britain appeared in 1931. Since that time, scores of anaesthetists have relegated their knowledge of statistics to a brief window immediately prior to an examination. Engagement with statistics is important in our ongoing appraisal of available evidence; particularly in a time when medical reversal (new research which contradicts established practice) and research fraud means we should be well equipped to interpret published studies.

Study Design and Data Collection

A bad study design begets bad results. Bad results derive from bad data or the wrong data. Therefore, before embarking on a study, it is important to pay heed to some key questions:

- What is the research question?
- What variables need to be measured?
- What is the main outcome variable?
- Are we aiming to deliver an intervention (experimental) or purely perform observation?
- Is a comparison group needed?
- What will be the duration of the study?
- How much will it cost?

Defining the clinical research question is the first step in devising a study (or reviewing the literature). One method to consider is PICO. This model raises key aspects to consider in formulating the question:

- P – patient/population
- I – intervention
- C – comparison
- O – outcome

For example, 'in major trauma patients over the age of 18 years (P) does the administration of tranexamic acid (I) rather than placebo (C) reduce mortality (O)?'

Studies can be classified as either observational or experimental.

Observational Studies

As suggested by the name, observational studies involve observing what happens (without interfering). There are three key types of observational study.

1. Cross-sectional Study

This is a snapshot, e.g. a survey. As such, it cannot determine the direction of causation but such studies are cheap and easy to perform. They may be:

- Descriptive – looking at the prevalence or incidence of a condition
- Analytic – studying variables among groups

2. Cohort Study

This may be prospective or retrospective. The aim is to identify the factors which influence the chance of a particular outcome occurring. To do so prospectively, we would take a group of people and follow them up over a period of time; we are looking to see whether exposure to a particular factor affects an outcome of interest.

Pros

- Can study several outcomes from the same risk factor
- Clear time-order of events
- Less potential for bias than case–control studies
- Suitable for studying rare risk factors

Cons

- Lengthy follow-up required
- Study participants may drop out or change relevant habits

3. Case–Control Study

This starts with a case group that have a condition of interest, for example hypercholesterolaemia. A group of matched individuals without the condition is then recruited to act as the controls. The control individuals are matched to the case individuals on a range of factors, for example gender and age. Prior exposure to risk factors is determined, to look for associations with a chosen outcome. The potential risk factors should be chosen in advance of the study and be clearly defined. By definition, these studies are retrospective.

Pros

- Relatively cheap, easy and quick to perform
- Suitable for studying rare conditions

Cons

- Recall bias – case-patients may remember certain aspects better than control-individuals
- Not suitable for studying rare risk factors

Experimental Studies

Unlike observational studies, experimental studies depend on an intervention being carried out. They are a prospective comparison of two or more treatments, one of which forms the control (this may be the current, standard treatment or a placebo). Having chosen the treatments to compare, we must decide on the outcome measures. Primary and secondary endpoints should be determined in advance.

Outcome Measures and Their Uncertainty

Outcome measures, sometimes referred to as endpoints, are the specific factors quantified to demonstrate effect within a study. The primary outcome relates to the main objective of the study. Secondary endpoints may also be studied. They should be clearly defined in advance (sometimes referred to as, *a priori*; from the Latin, 'from the earlier'; in other words, determined before the experiment).

The CONSORT (Consolidated Standards of Reporting Trials) statement provides recommendations for best-practice reporting of randomized trials. Item number six from their checklist of information to include when reporting a trial relates to outcomes [1]. Namely, that outcome measures (both primary and secondary) should be 'completely defined pre-specified . . . , including how and when they were assessed'. This is important to avoid data being indiscriminately analyzed after the study and yielding chance findings. As an example, the article 'Gone fishing in a fluid trial' by Hjortrup *et al.* intentionally demonstrates the flaw associated with post-hoc analysis of data [2]. They proved a mortality benefit with being born under the Zodiac sign, Pisces, when taking part in a trial of intravenous fluids!

Uncertainty around outcome measures could arise from combined endpoints, e.g. a composite of mortality, cardiovascular events and stroke.

Sampling

In order to determine what size of study sample is required, a power calculation will be performed by the statistician at the study design phase. This topic is dealt with in more detail in the section on hypothesis testing later in the chapter. In essence, it accounts for the predicted magnitude of the effect of the intervention or treatment (e.g. does a new treatment reduce mortality by 2% or 20%) along with a few other factors to determine the sample size required for the study.

Generally speaking, it is not possible to collect data from the entire target population. Therefore, a sample is used, and statistical inference applied. The aim is to have a sample that is representative of the population of interest. This can be achieved by various sampling methods (Table 1.1).

Table 1.1 Various methods of sampling the population

Simple random sampling	Random selection from the population
Systematic random sample	Random first recruit, e.g. 19th admission to ICU this year, then interval sampling thereafter. The interval is determined as population size divided by sample size, e.g. sample of 200 from 1,000 ICU admissions means sampling every fifth patient.
Stratified random sample	Pooled sample from a number of random samples from several strata within the population, e.g. random sample of women, of men, of smokers, etc. Size of samples from each stratum determined by relative size within population: if smoking prevalence is 10%, the final pool will contain smokers and non-smokers in a 1:9 ratio as well.
Cluster sample	Simple random sample from groups within the population which are otherwise expected to be homogeneous, e.g. random sample of post-cardiac arrest patients, admitted to different ICUs.
Consecutive sample	Every eligible individual is recruited from the start of the study period until the required sample size is achieved.

Having obtained our sample, for a prospective study it is then necessary to allocate patients to the intervention and control groups. Allocation is a separate process to sampling but is also undertaken in a random manner. Randomization is used to ensure that any differences between the groups within the study are due to chance, i.e. to reduce the chance of statistical error. This can be done using a random-number table. Due to the nature of randomization, it may be necessary to undertake block randomization in order to maintain equal numbers of subjects in each group.

Blinding refers to obscuring to which group each patient has been assigned. It is not always possible, for example in the case of a study comparing having surgery or not. However, it is a powerful method to avoid bias (see below). Blinding can refer to study participants (patients) or investigators. Where neither party is aware which arm the participant is in, the study is double-blinded. A study which is not blinded is 'open' or 'open-label'.

Bias

Bias refers to a situation that results in a difference between study results and reality. There are numerous ways by which bias may be introduced. Bias may lead to overestimation or underestimation of an effect.

Selection Bias

The study population is not representative of the actual population. This should be reduced through appropriate randomization. Sub-types include:

- Ascertainment: when the sample is not randomly selected
- Attrition: differences in those lost to follow-up to those not
- Response: differences between volunteers to a study and non-volunteers
- Survivorship: study participants have to survive long enough to receive the intervention of note, but survival is measured from an earlier time

Information Bias

The incorrect recording of measurements, to include:

- Central tendency: responses on a Likert scale (1–5 Strongly disagree/Strongly agree) tend to merge to the centre (neither agree nor disagree)

- Lead-time: the advent of new diagnostic tests results in patients entered into a study later being diagnosed earlier in the disease process, leading to an apparent increase in survival – not due to the intervention
- Measurement: from an inaccurate or badly calibrated measuring device
- Misclassification: wrongly classifying an outcome variable
- Observer: also known as assessment bias – when an observer over- or under-reports a variable
- Reporting: study participants may give answers they think the investigator wants to hear or withhold information they (wrongly) believe to be irrelevant

Publication Bias

A tendency of journals only to publish papers with positive results. There is more detail on this in respect of meta-analysis and systematic review towards the end of the chapter.

Confounding

A confounding variable affects both the independent and the dependent variables but is not part of the exposure-outcome causal pathway. Female patients have a higher incidence of PONV and only female patients undergo gynaecological procedures but such procedures do not directly predispose to PONV: gender is a confounding variable here (Figure 1.1).

Confounding effects are controlled for at the design stage of a study by restriction, matching or randomization and at the data analysis stage by stratification or adjustment.

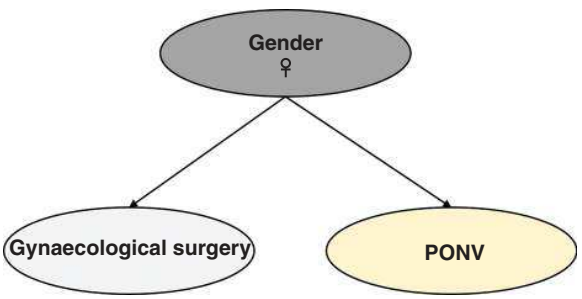

Figure 1.1 Confounding variables; causality indicated by arrows

Other Considerations

Crossover occurs when each group acts as both the control and treatment group within the same study. This can lead to a range of inaccuracies in the final data. Care must be taken to ensure any drug effects are not carried over to the control period. This is ensured by including a suitable wash-out period between the two arms of the study.

Intention-to-treat analysis involves analyzing all data as if any subjects that were lost are still in the study. Subjects may be lost for a number of reasons, e.g. refusal, moving house, dying. The analysis of all subjects aims to achieve a more true-to-life estimate of any treatment effect. In addition, maintenance of the sample size ensures that the power of the study is not diminished by the loss of subjects.

All clinical trials must undergo consideration by an ethics committee. Trials must abide by the Declaration of Helsinki. In addition, training in Good Clinical Practice (GCP) standards is important for individuals involved in research, to ensure that they have a full understanding of their responsibilities as part of the research team.

Descriptive Statistics

Having undertaken the observation phase of a study, you will hopefully have a large number of values: your data. The next step is to characterize and describe the data. This is important in two distinct ways. Firstly, it is helpful to gain a very crude understanding of any patterns in the values. If you have measured the heights of a class of school children, is there a typical range of heights for that age group or is this group unusually tall for their age? Secondly, what type of data have you collected? This directly informs what types of statistical analyses are appropriate to use in gaining a greater understanding of the meaning behind your data. In reality, you ought to understand this aspect at the planning stage of your study but determining data types and matching them to statistical tests is a crucial skill in critical appraisal of medical evidence: the literature is scattered with studies which misidentify the data type or use the incorrect statistical test or often both.

Data Types

Variables are classified as follows:

Qualitative or Categorical

- Dichotomous: data which can only have two categories, e.g. gender, malignant vs. benign, weight >40 kg vs. weight <40 kg
- Nominal: the data have more than two categories but these have no intrinsic order, e.g. different types of lung cancers observed in a study of mine workers
- Ordinal (or ranked): the data fall into several categories which have an intrinsic order but, importantly, cannot have a value assigned to them, e.g. ASA score, GCS.

Categorical data can cause some confusion. The categories should be thought of as having a 'label', not a value, but this can be easily forgotten, particularly with ordinal variables which, by definition, use numbers as the labels; you cannot perform quantitative analysis on observations of GCS – a mean GCS of 8.7 is meaningless.

Quantitative or Metric

- Discrete: the data reflect whole-number counts, e.g. number of patients responding to a treatment, platelet count
- Continuous: the possible number of measured variables is only limited by the resolution of the measuring equipment, e.g. height, temperature, BMI

Quantitative data are easy to spot; they have units of measurement. Sometimes these are not always reported or used in everyday practice, which can lead to misidentifying the variable type. BMI is a good example of this – the units are kg m^{-2} but these are colloquially omitted.

Continuous variables can be subclassified to interval and ratio variables. By definition, a ratio variable is one where zero denotes none of that quantity. Consider temperature as an example. Kelvin is the ratio variable for temperature; 20 K is twice as hot as 10 K. By contrast, an interval variable has a relative zero point, e.g. temperature measured in degrees Celsius. Whilst the interval between 10 °C and 20 °C equates to that between 30 °C and 40 °C, 20 °C is not twice as hot as 10 °C; 10 °C and 20 °C actually equate to 283.15 K and 293.15 K so clearly one is not twice as hot as the other.

Data Representation

In simple terms, data can be depicted in rows and columns – a table or, more formally, a contingency table. This form of presentation is the basis for undertaking chi-squared analysis (see later). Graphical depictions of the data can be a more intuitive representation of the interdependence between two variables. The simplest types are those which demonstrate the frequency distribution for the values. There are two distinct types. For qualitative data, a bar chart is used (Figure 1.2). The height of the bars corresponds to the number in each category. The order of the bars along the *x*-axis is arbitrary (except in the case of ordinal/ranked data). By convention, the bars do not 'touch', which emphasizes the fact that the groups represented by the bars are not numerically contiguous.

By contrast, quantitative data are represented as a histogram (Figure 1.3). For continuous values, the data are first split into groups, sometimes called 'bins'. For example, age data from a population may be grouped

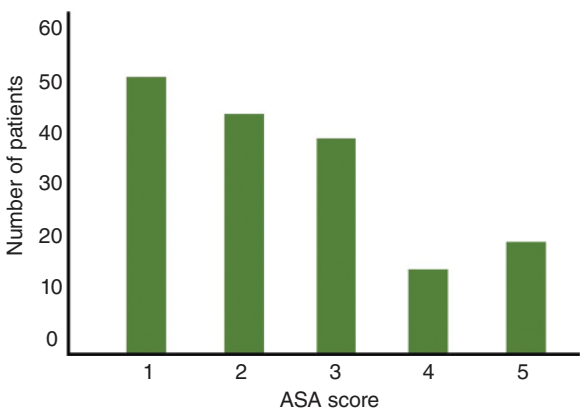

Figure 1.2 A bar chart of patients by ASA score

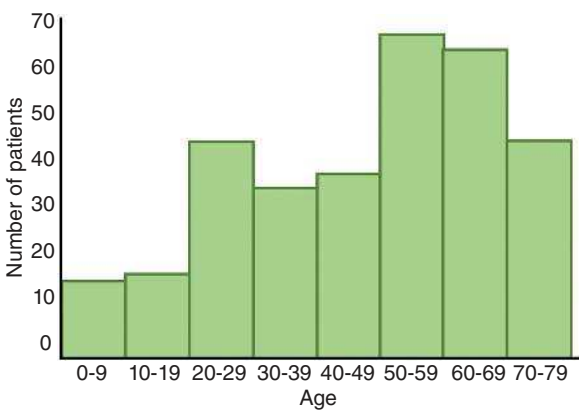

Figure 1.3 A histogram of patients' ages

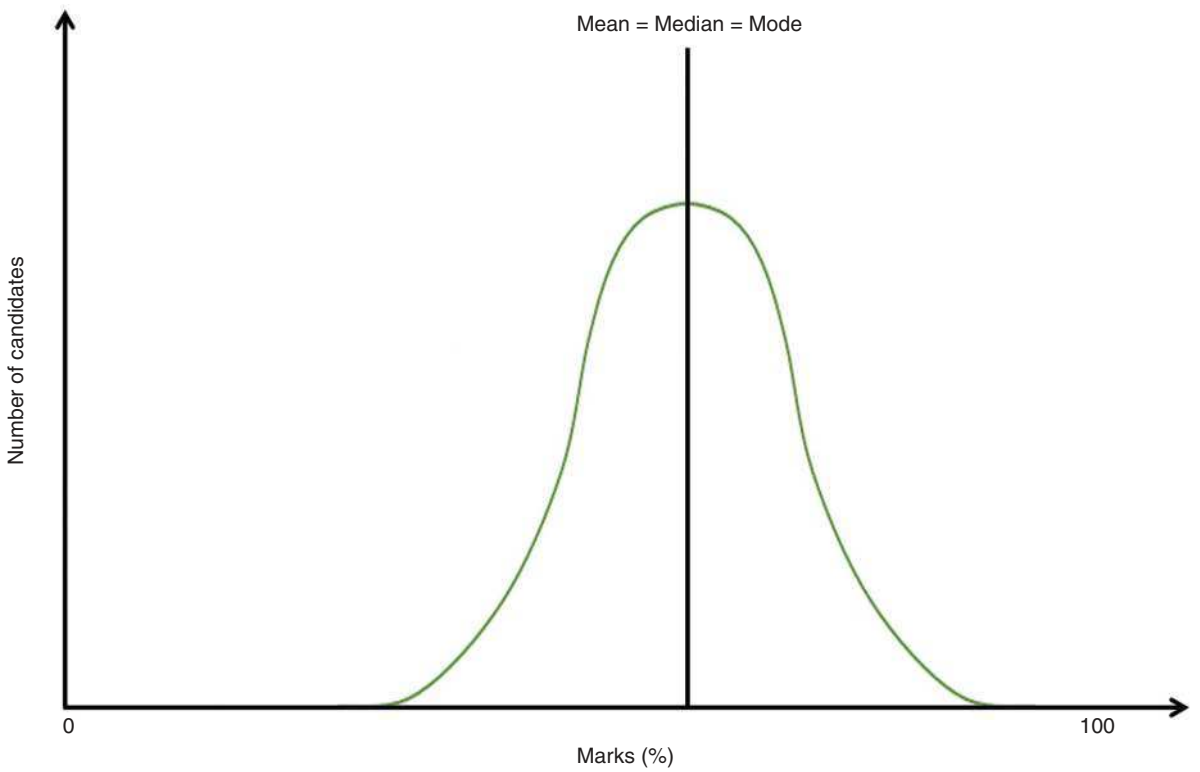

Figure 1.4 The normal distribution of a set of FRCA scores for a group of candidates

into bins of 40–49.9 years, 50–59.9 years and so on. The bins represent contiguous sections of the *x*-axis and so the bars touch on a histogram and clearly have an order of sequence. Because each patient in each bin in our example could have an age anywhere within the range of the bin, it is the bar area rather than just the height which is proportional to the frequency of patients falling within the bin.

The most familiar distribution for natural phenomena, e.g. height, weight or IQ, is the *normal distribution*, sometimes referred to as a 'bell curve' on account of the characteristic shape (Figure 1.4). It is a symmetrical distribution around the most common value which, in the case of a truly normal distribution, is simultaneously the median, mode and mean (see later).

Other data sets are asymmetrical. This can be the case for variables that would be expected to demonstrate a normal distribution and this is often due to the effects of a small sample size where outliers in the population can distort the distribution. This is referred to as *skew*. Skew can be negative or positive and this simply refers to where the outliers lie on the *x*-axis. In a positive skew, the outliers lie towards the right of the *x*-axis (and the bulk of the population more towards the left) and vice versa for a negative skew. Some find it counterintuitive that the description of skew seems to refer to where the outliers are, rather than the bulk of the sample. A useful aide memoire is that the 'skew' produces a 'skewer-like' projection off the side of the bell curve. Thus, in a positive skew, the 'skewer' points to the right of the *x*-axis and vice versa. Figure 1.5 demonstrates the distribution of marks from two further sittings of the FRCA – the solid line shows positive skew, the dashed line shows negative skew. For a positive skew, mode $<$ median $<$ mean whereas for a negative skew, mode $>$ median $>$ mean (as a rule of thumb).

If skew results from sample-size issues, a larger sample will tend to reduce the skew. This is simply due to the fact that a larger sample is more likely to also include a few outliers at the opposite end of the *x*-axis. For example, if we recorded the blood pressure of all patients attending a cardiology outpatient clinic, we might find a high proportion of patients with a high blood pressure and only a few with normal or low values; a negative skew. If we take a larger population, for example everyone walking into the hospital (outpatients, staff and visitors), we would expect to mostly encounter normotensive individuals and a

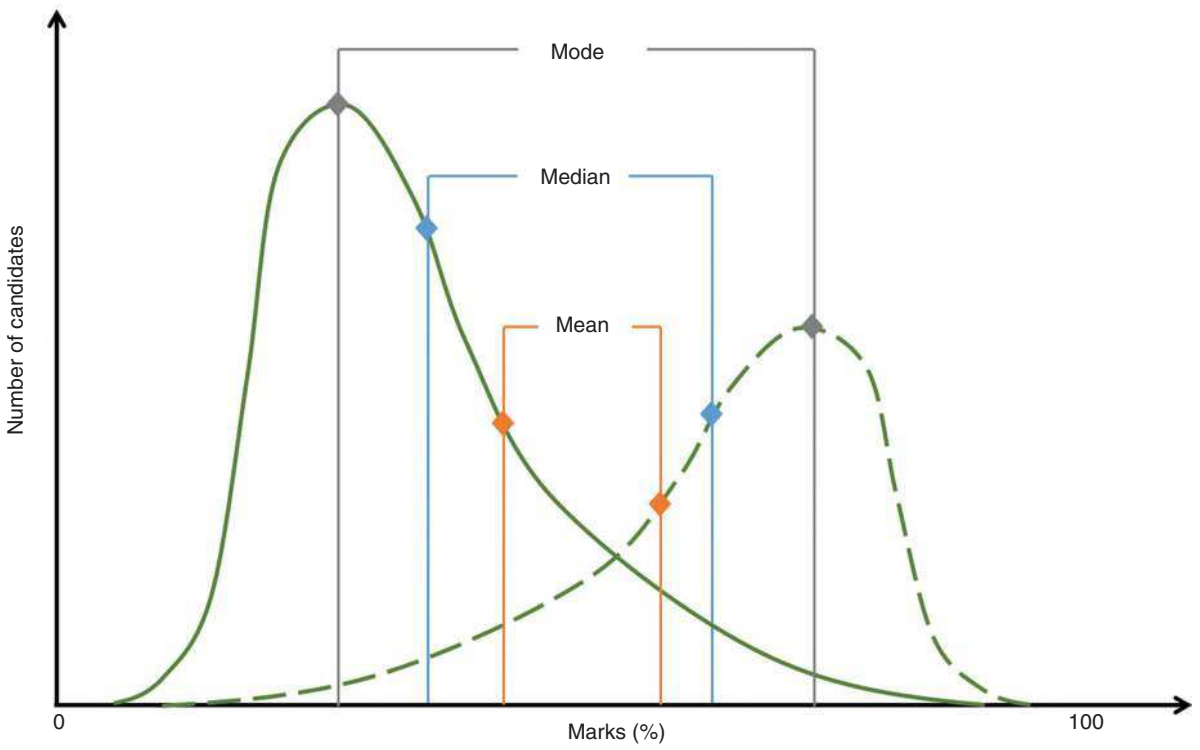

Figure 1.5 Skewed distributions – solid line is positively skewed, dotted line is negatively skewed

group of healthy, cycling enthusiasts/anaesthetists with lower blood pressures. The cyclists would balance the clinic patients and a normal distribution would be expected.

The characteristics of a normal distribution are mathematically a little more complex. Firstly, the median must equal the mean. The 'tails' of a normal distribution are summarized by the standard deviation (see later) and kurtosis. Determining the kurtosis is a complex calculation, yielding a unitless pure number. A value of 3 suggests a normal distribution (mesokurtic); higher values (leptokurtic) suggest extreme outliers, long fat tails and a narrow, thin central peak; whereas lower values (platykurtic) suggest no outliers, i.e. short thin tails and a broad, flat central peak. A calculated skew of zero is a characteristic of a normal distribution but it does not guarantee it; it could result from a balance of a short, fat tail on one end of the distribution and a long, thin tail on the other. Specific tests such as the Kolmogorov–Smirnov test can test the normality of a distribution directly.

The Exponential Distribution

One particular form of skewed distribution is the exponential distribution (Figure 1.6). In this distribution, all outliers are in one direction (an extreme positive skew in the case of a negative exponential). The exponential distribution is also known as the negative exponential distribution because of the mathematical relationship which describes it; it includes a negative exponent. To explain what that means, it's worth looking at a common example from the FRCA syllabus. After a bolus dose of an intravenous medication, such as propofol, the reduction from the peak blood concentration follows a negative exponential relationship. The concentration of propofol at any specified time after the bolus, (C_t), can be found as a function of the initial concentration (C_0) and the time between C_0 and C_t. Here is the maths: the initial concentration is multiplied by the mathematical constant, 'e', raised by an 'exponent' (hence the name of the relationship) of $-kt$, which is the rate constant (k) times the time elapsed (t):

$$C_t = C_0 \times e^{-kt}$$

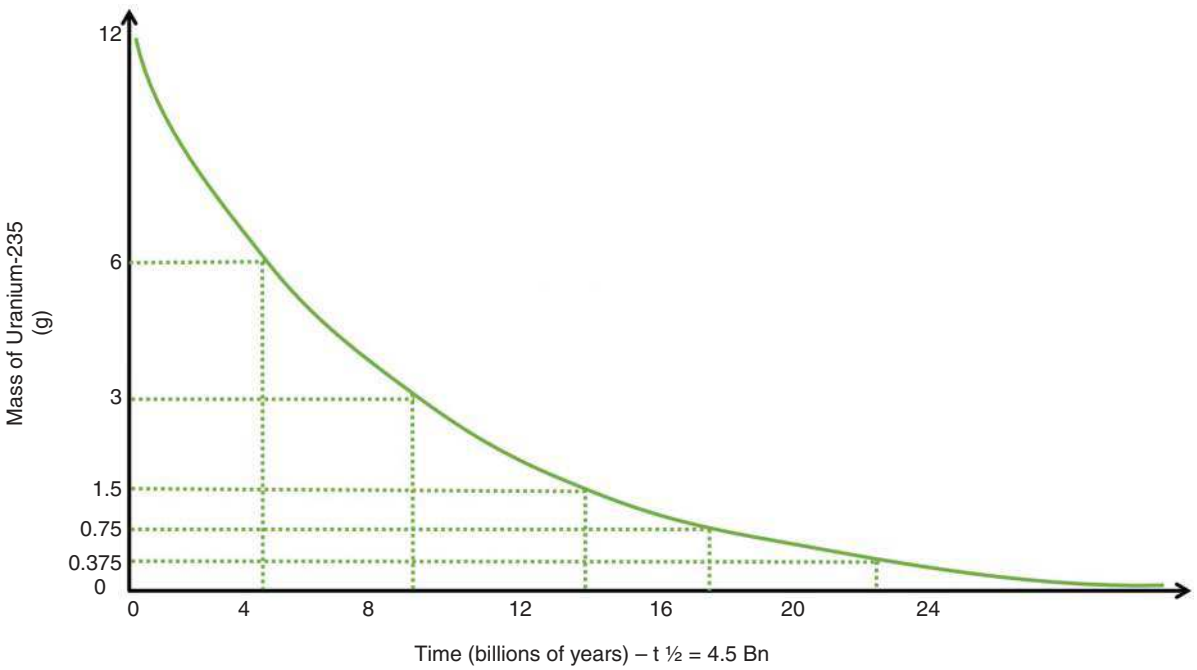

Figure 1.6 The exponential distribution; decay of a 12-g sample of uranium-235

Of course, as any good FRCA candidate knows, the pharmacological reality is a little more complex and the actual mathematical relationship describes the interplay of exponential decay between other pharmacokinetic compartments, resulting in a biexponential relationship.

Euler's number, *e*, is a mathematic constant named in honour of Swiss mathematician Leonhard Euler. It is the number whose natural logarithm = 1. Like 0, 1, i and π, it is one of the five irrational numbers in mathematics (ones which are not ratios of integers). It was actually discovered by fellow Swiss mathematician, Joseph Bernoulli, when he was studying the financial concept of compound interest. It is also known as Napier's number (who first described natural logarithms). Euler's number is different from Euler's constant (also known as the Euler–Mascheroni constant).

All statistical distributions can be described by a mathematical relationship but the exponential relationship is worthy of the particular focus we have just paid it, due to its clinical and medical physics applications, such as pharmacokinetic phenomena and radioactive decay (see Chapter 15). An everyday example is the computer-controlled pumps used for total intravenous anaesthesia. These perform these exponential calculations in real time.

Summary Statistics

To mathematically describe the data, we need to use *summary statistics*. Probably the first area of interest is the part of the data set which contains the most-frequently-occurring value, for example, the commonest weight within a group of surgical patients. Measures of central tendency, such as averages (see below), allow us to do this, and measures of spread, such as standard deviation, then allow a mathematical description of how the rest of the sample relates to these central values. These statistics are useful for ordinal (qualitative) and all quantitative data (Table 1.2).

The measures of central tendency most commonly used are the averages. The term 'average' is used a little loosely. Technically, it includes mean, median and mode although it is often simply used as a synonym for the mean.

The mode is very straightforward to work out; it is simply the most frequently occurring value in the data set. Consequently, it is not affected by outliers but,

Table 1.2 Appropriate measures of central tendency for different data types

	Mode	Median	Mean
Categorical nominal	Yes	No	No
Categorical ordinal	Yes	Yes	No
Quantitative discrete	Yes	Yes	Yes
Quantitative continuous	No	Yes	Yes

equally, it barely describes any of the data. The only common application of the mode is in describing multimodal distributions. Age of onset of type-1 diabetes is an example of a bimodal distribution; there is a peak between 5 and 9 years and between 10 and 14 years of age. Continuous data are difficult to describe with the mode. Imagine a truly continuous sample of patient weights, accurate to grams. It is almost inconceivable that even two patients would weigh the same to this level of accuracy so the sample would have no mode.

The median is the middle value in the range when the data are sorted into ascending order. It is not particularly affected by skewed data sets, which is why it is the average of choice for non-normally distributed data. However, as with the mode, it does ignore most of the data apart from those at the centre.

To obtain the mean, all values are added together and divided by the number of values. This has the advantage that all data are incorporated in the measure. This is also its disadvantage as skewed data are poorly described by the mean; you should only use it with normally distributed data. It cannot be used with ordinal data.

A note on discrete data and calculating the mean. This might seem counterintuitive as you may well end up with a non-integer result, e.g. 0.7 of a patient. It is, however, perfectly correct to say that, for example, an average of 28.7 patients are admitted to an ICU per month. That's not quite the same as saying that, in an average month, 28.7 patients are admitted. We don't know what 'an average month' is and that phrase would imply the month is the average 'thing' and we therefore would have to account for what we mean by 0.7 of a patient.

Measure of Spread

Measures of spread help to describe how the remainder of the data relate to the central portion and, in turn, reflect the validity of the measure of central tendency. A widely spread data set suggests the presence of outliers, the potential for skew and hence an unreliable measure of central tendency.

The range simply describes the difference in value between the smallest and largest value in the sample. It is not affected by where the rest of the data lie between those limits. As such, it tells us nothing about skew (asymmetry of the bell curve) or kurtosis (the narrowness of the bell); both of these are measures of central spread. That said, outlier data do dramatically affect the validity of the range as a description of the rest of the data.

Quartiles, as the name suggests, divide the data set into four portions, in exactly the same way that the median divides the data in halves. Quartile 1 includes the first 25% of the data from the smallest value. Quartile 2 includes the next 25% up to the median, and so on. The interquartile range (IQR) is the difference between the lower limit of quartile 2 and the upper limit of quartile 3 – the middle 50% of the data. This therefore excludes any outliers but can still be affected by a skew. By definition, of course, the IQR fails to reflect 50% of the data – the upper 25% and lower 25%. The IQR and median are the basis for a box-and-whisker plot (Figure 1.7). The box contains the IQR and is bisected by a line representing the median. The whiskers extend out to the maximum and minimum values. The figure also demonstrates the effect of skew on the appearance of the plot (Group B in Figure 1.7 has a negative skew).

To describe the distribution of data within these wider bands of ranges and quartiles, we can describe their position mathematically in relation to the mean. Simple measures are the absolute variation and variance but these are somewhat simplistic and do not provide a particularly useful number (so much so, you may not even have heard of them before). The standard deviation (SD) overcomes these technical issues. In turn, it can also be used to quantitatively describe the shape of a distribution (see earlier). Mathematically it is calculated as:

$$SD = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Where x = each value, $\bar{x}$ = mean, n = number of values. Standard deviation is only suited to continuous data and should only be used for normally distributed data. The mean ± 2 SD defines a range which will contain 95.4% of all data points if the data is normally distributed.

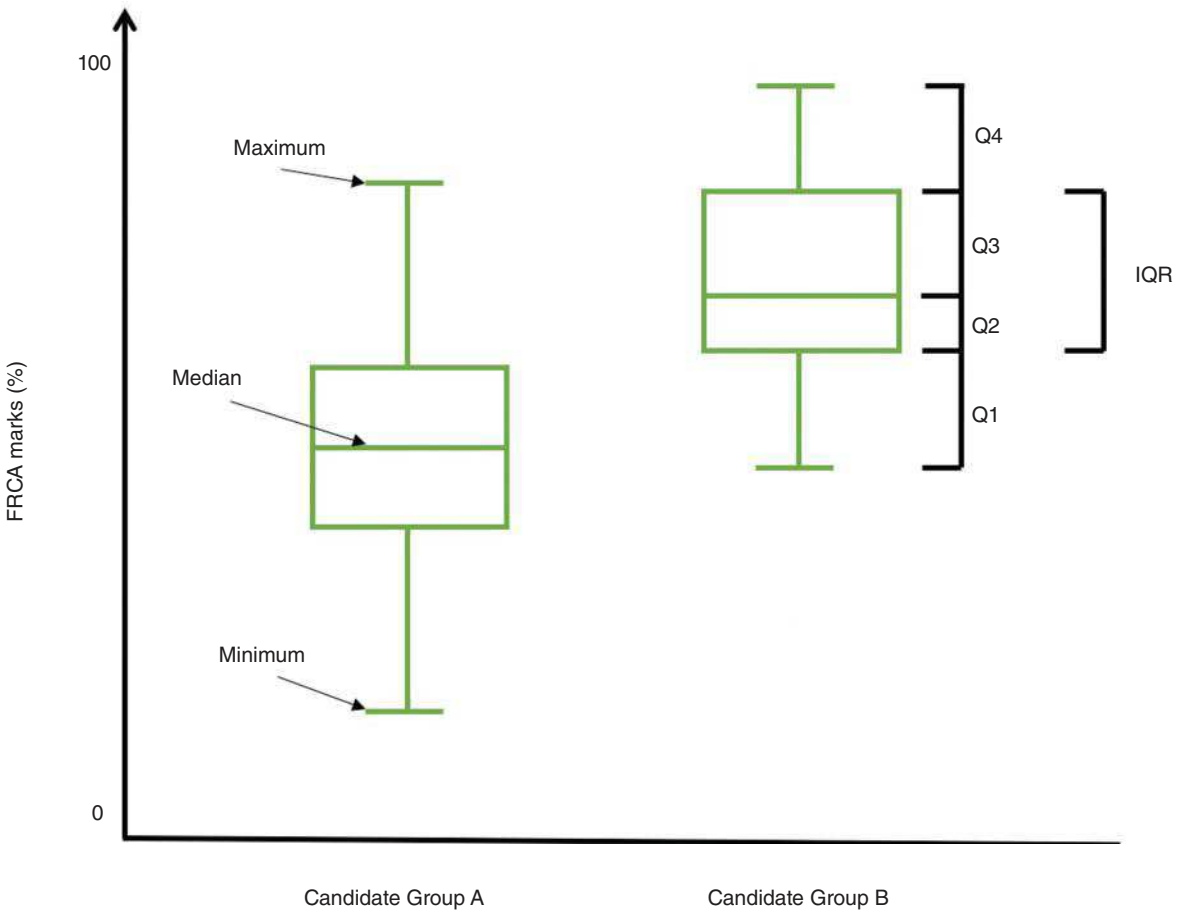

Figure 1.7 Box-and-whisker plot of FRCA scores for two different groups. The components of the box-and-whisker plot are demonstrated for Group B's plot

Statistics vs. Parameters

It is worth mentioning that we have talked principally about statistical terms so far. Statistics are descriptors of a sample. They are often then extrapolated to a population. For example, if we measure all the heights of a class of medical students we could generate a mean height. That mean is a 'parameter' of the class – it is the population mean for that group because we have measured every individual in the population (the population being the class). If we want to use that result as a surrogate for all UK 18-year-old adults, we refer to it as a statistic because it is calculated from a sample, not the whole population of UK 18-year-olds. Just remember: **s**tatistics for **s**amples, **p**arameters for whole **p**opulations. We can calculate the difference between any given statistic and the corresponding parameter by calculating the standard error. The most common example is the standard error of the mean (SEM). The larger the sample, the smaller the SEM – the closer you come to sampling the entire population, the closer the sample mean approximates to the population mean and the SEM tends towards zero:

$$SEM = \frac{\sigma}{\sqrt{n}}$$

The standard deviation of the population (σ) is obviously not often known and so the sample SD is usually substituted but this renders the SEM an approximation:

$$SEM \approx \frac{SD}{\sqrt{n}}$$

Table 1.3 summarizes the measures of spread and for which types of variables they can be applied.