AN INTRODUCTION TO STATISTICS

An introduction to everyday statistics—2

P Driscoll, F Lecky, M Crosby

Objectives
- Describe central tendency and variability
- Summarising datasets containing two variables

In covering these objectives we will deal with the following terms:
- Mean, median and mode
- Percentiles
- Interquartile range
- Standard deviation
- Standard error of the mean

In the first article of this series, we discussed graphical and tabular summaries of single datasets. This is a useful end point in its own right but often in clinical practice we also wish to compare datasets. Carrying this out by simply visually identifying the differences between two graphs or data columns lacks precision. Often therefore the central tendency and variability is also calculated so that more accurate comparisons can be made.

Central tendency and variability

It is usually possible to add to the tabular or graphical summary, additional information showing where most of the values are and their spread. The former is known as the central tendency and the latter the variability of the distribution. Generally these summary statistics should not be given to more than one extra decimal place over the raw data.

Key point
Central tendency and variability are common methods of summarising ordinal and quantitative data

CENTRAL TENDENCY

There are a variety of methods for describing where most of the data are collecting. The choice depends upon the type of data being analysed (table 1).

Mean

This commonly used term refers to the sum of all the values divided by the number of data points. To demonstrate this consider the following example. Dr Egbert Everard received much praise for his study on paediatric admissions on one day to the A&E Department of Deathstar General (article 1). Suitably encouraged, he reviews the waiting time for the 48 paediatric cases involved in the study (table 2).

Table 2 Waiting time for paediatric A&E admissions in one day to Deathstar General

Case	Waiting time (min)	Case	Waiting time (min)	Case	Waiting time (min)	Case	Waiting time (min)
1	46	13	19	25	10	37	9
2	45	14	20	26	17	38	31
3	44	15	19	27	14	39	18
4	26	16	14	28	15	40	7
5	52	17	27	29	22	41	19
6	14	18	26	30	27	42	22
7	19	19	24	31	45	43	14
8	18	20	17	32	18	44	11
9	14	21	19	33	22	45	17
10	34	22	28	34	10	46	26
11	18	23	37	35	19	47	37
12	22	24	12	36	15	48	41

Considering cases 1 to 12, the mean waiting time is:

46+45+44+26+52+14+19+18+14+34+18+22/12

Which is:

352/12 = 29.3 minutes

Key point
Mean = Σx/n. Where Σx = the sum of all the measurements and n = the number of measurements

Consequently the mean takes into account the exact value of every data point in the distribution. However, the mean value itself may not be possible in practice. An example of this is the often quoted mean of "2.4 children" per family. Another major advantage of using the mean is that combining values from several small groups enables a larger group mean to be estimated. This means Egbert could determine the mean of all 48 cases by repeating this procedure for cases 13 to 24; 25 to 36 and 37 to 48. The four separate means could then be added together and divided by 4 to give the overall mean waiting time.

The problem in using the mean is that it is markedly affected by distribution of the data and outliers. The latter are the extreme values of the data distribution but are not true measures of the central tendency. The mean is therefore ideally reserved for normally distributed data with no outliers.

Accident and Emergency Department, Hope Hospital, Salford M6 8HD

Correspondence to:
Mr Driscoll, Consultant in A&E Medicine (pdriscoll@hope.srht.nwest.nhs.uk)

Table 1 Applicability of measure of central tendency

	Mean	Median	Mode
Useful with quantitative data	✓	✓	✓
Useful with ordinal data	×	✓	✓
Useful with nominal data	×	×	✓
Effected by outliers	✓	×	×
Effected by lack of normal distribution	✓	×✓	×

Key points
- The mean reflects all the data points
- Small group means can be combined to estimate an overall mean
- The mean is commonly used for quantitative data

Median

The mean cannot be used for ordinal data because it relies on the assumption that the gaps between the categories are the same. It is possible however to rank the data and determine the midpoint. When there is an even number of cases the mean of the two middle values is taken as the median. In Egbert's study, the median waiting time for cases 1 to 12 would therefore be calculated by:

Listing the data points in rank order:

14 14 18 18 19 22 26 34 44 45 46 52

Determining the midpoint that lies half way between the 6th and 7th data point:

(22 + 26)/2 = 24 minutes

Key point

Median = value of the (n+1)/2 observation. Where n is the number of observations in ascending order

In contrast with the mean, the median is less affected by outliers, being responsive only to the number of scores above and below it rather than their actual value. It is therefore better at describing the central tendency in non-normally distributed data and when there are outliers. However, in bimodal distribution, neither median or mean will provide a good summary of the central tendency.

A common example of the use of medians is seen in the longevity of grafts, prostheses and patients. It is used because the loss to follow up, and the unknown end point of the trial, prevents a mean being used to determine the average survival duration.

Key points
- Medians only reflect the middle data points
- You cannot combine medians to get an overall value
- Median is commonly used for ordinal data, and for quantitative data when the data are skewed or contains outliers

Mode

This equals the commonest obtained value on a data scale. The mode is also used in nominal data to describe the most prevalent characteristic particularly when there is a bimodal distribution. Consequently in Egbert's study, there would be two mode values for cases 1 to 12—that is, 14 and 18 minutes.

The mode can be demonstrated graphically. Another advantage is that it is relatively immune to outliers. However, the peaks can be obscured by random fluctuations of the data when the sample size is small. Furthermore, unlike the mean and median, it can change markedly with small alterations in the sample values.

Key points
- Mode only reflects the most common data points
- Modes cannot be combined to get an overall value
- Mode is commonly used for nominal data

Key points
- The mean, median and mode will all be the same in normally distributed data but differ when the distribution is skewed (fig 1)
- When dealing with skewed data, the mean is positioned towards the long tail, the mode towards the short one and the median is between the two. The median value divides the curve into two equally sized areas (fig1(B) and (C))
- The mode is a good descriptive statistic but of little other use
- The mean and median provide only one value for a given dataset. In contrast there can be more than one mode

VARIABILITY

This is a measure of the distribution of the data. It can be described visually by plotting the frequency distribution curve. However, it is best to quantify this spread using methods dependent upon the type of data you are dealing with (table 3). Though there are measures of spread of nominal data, they are rarely used other than to provide a list of the categories present.[1] We will therefore concentrate on methods used to describe the variability of ordinal and quantitative data.

Range

This is the interval between the maximum and minimum value in the dataset. It can be profoundly effected by the presence of outliers and does not give any information on where the majority of the data lies. As a consequence other measures of variability tend to be used.

Percentiles

It is possible to divide up a distribution into 100 parts such that each is equal to 1 per cent of the area under the curve. These parts are known as percentiles. This process is used to identify data points that separate the whole distribution into areas representing particular percentages. For example, as described previously, the median divides the distribution into two equal areas. It therefore represents the 50th centile. Similarly the 10th centile would be the data point dividing the distribution into two areas one representing 10% of the area and the other 90% (fig 2).

To determine the value of a particular percentile we use the same process required to work out the median. To demonstrate this

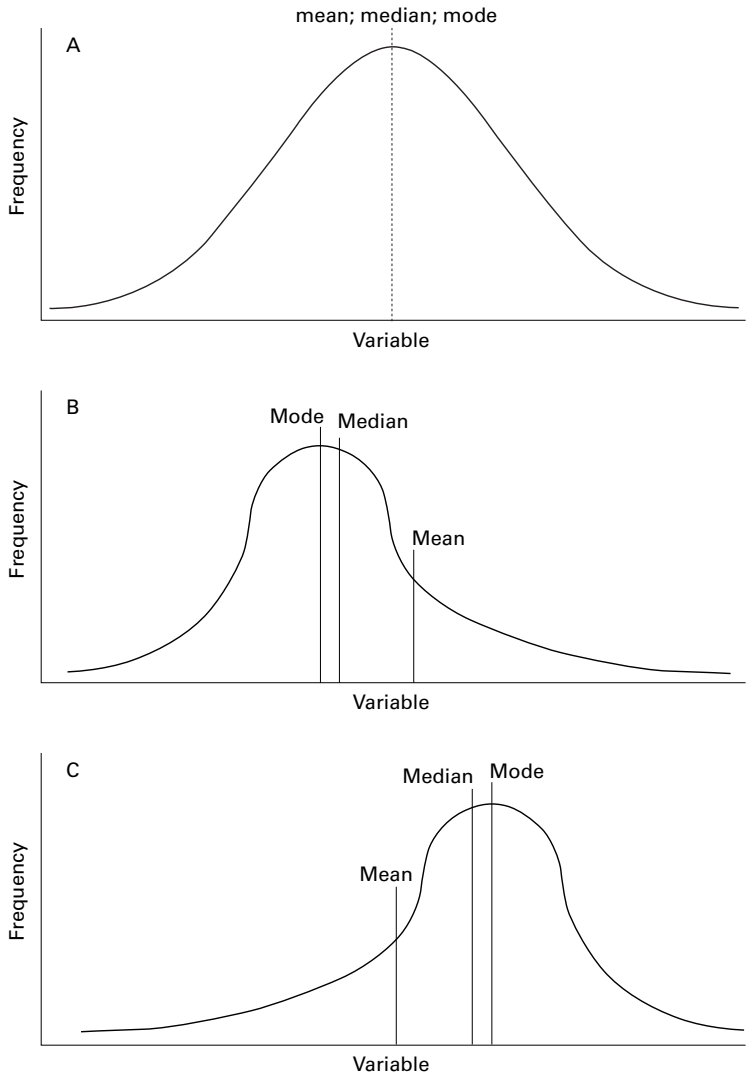

Figure 1 Location of mean, median and mode with different distributions. (A) Normal distribution. (B) Right (positive) skewed distribution. (C) Left (negative) skewed distribution.

Table 3 Methods of quantifying distribution

	Range	*IQ range*	*SD*	*SEM*
Descriptive of sample variability	✓	✓	✓	✓
Describes quantitative data	✓	✓	✓	✓
Describes ordinal data	✓	✓	×	×

Key point
Percentiles are values that identify a particular percentage of the whole distribution

consider how Egbert would determine the 25th centile in his waiting time data.

Firstly, he needs to sort them into rank order (table 4). He then locates the position of the 10th percentile by:

$10/100 \times (n + 1)$ where n is the number of data points in the distribution.

The position of the 10th percentile is therefore $0.1 \times 49 = 4.9$

This means the 10th centile lies nine tenths of the way from the 4th (that is, 10 minutes) to the 5th (that is, 11 minutes) data points. Consequently the 10th percentile is 10.9 minutes. In other words, 10 per cent of all the waiting times in Egbert's study have a value of 10.9 minutes or less. An equally correct way of expressing this result would be say that 90% of all the waiting times in Egbert's study have a value of 10.9 minutes or greater (fig 2).

Key points
- Value of the X percentile = X (n+1)th observation/100. Where n is the number of observations in ascending order
- Percentiles can only be calculated for quantitative and numeric ordinal data

The two most commonly used percentiles are the 25th and 75th ones. As they are marking out the lower and upper quarters of the distribution they are known as the lower and upper quartile values. The data points between them represent the interquartile range.

Interquartile (IQ) range

As the lower and upper ends of the distribution are eliminated, the interquartile range shows the location of most of the data and is less affected by outliers. It represents a good method of summarising the variability when dealing with ordinal data or distributions that are not "normal".

The IQ range can also be used to provide a good graphical representation of the data. As the lower and upper quartile values represent the 25th and 75th percentile, and the median the 50th, these can be used to draw a "box and whisker" plot (fig 3). The box represents the central 50% of the data and the line in the box marks the median. The whiskers extend to the maximum and minimum values. In the case of outliers the whiskers are restricted to a length one and a half the interquartile range with other outliers being shown separately (fig 3). Box and whisker plots allow the reader to quickly judge the centre, spread and symmetry of the distribution. In so doing it enables two distributions to be compared (see later). Nevertheless, from a mathematical point of view, the IQ range does not provide the versatility required for detailed comparisons of normally distributed quantitative data. For this we need the variance and standard deviation.

Key points
- The IQ range is a good method of describing data but not the most efficient way of describing data variability when comparing groups
- The IQ range is commonly used to describe the variability of ordinal data and quantitative data that are skewed or have outliers.

Variance

An obvious way of looking at variability is to measure the difference of each value from the sample mean—that is, the calculated mean of

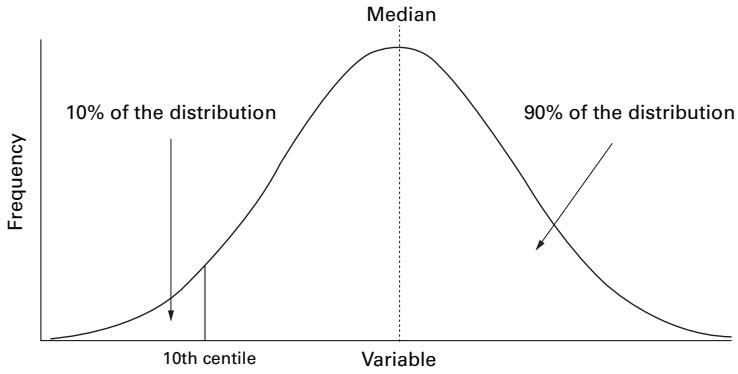

Figure 2 Frequency distribution showing the 10th centile.

Table 4 Waiting time for paediatric A&E admissions in one day to Deathstar General—sorted by rank order

Rank order	Waiting time (min)	Rank order	Waiting time (min)	Rank order	Waiting time (min)	Rank order	Waiting time (min)
1	7	13	15	25	19	37	27
2	9	14	17	26	19	38	28
3	10	15	17	27	20	39	31
4	10	16	17	28	22	40	34
5	11	17	18	29	22	41	37
6	12	18	18	30	22	42	37
7	14	19	18	31	22	43	41
8	14	20	18	32	24	44	44
9	14	21	19	33	26	45	45
10	14	22	19	34	26	46	45
11	14	23	19	35	26	47	46
12	15	24	19	36	27	48	52

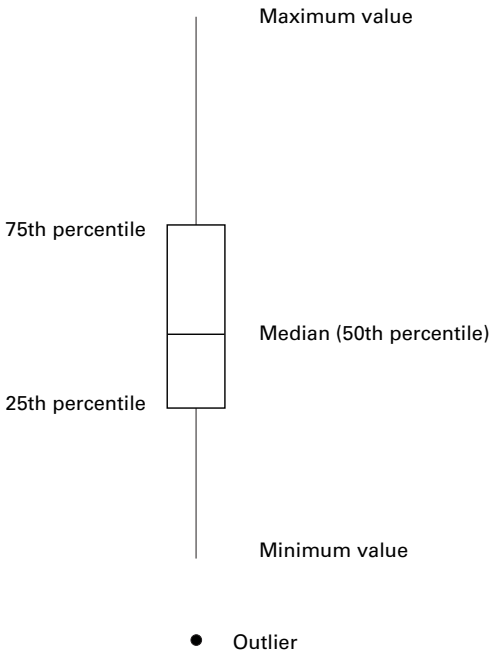

Figure 3 A box and whisker plot.

the study group. However, in simply adding up these numbers you would always arrive at zero because the sum of the differences bigger and smaller than the mean are equal but opposite. However, by squaring all the differences before adding them together, a positive result is always achieved. The sum of these squares about the mean is usually known as the "sum of the squares". For example, in a study the following differences were found between the mean and each value:

−3, −2, −2, −1, 0, 0, 0, 1, 1, 1, 2, 3

Therefore the sum of the squares is:

+4 +4 + 1 + 0 +0 +0 + 1 +1 +1 + 4 + 9 = 34

The size of the sum of the squares will vary according to the number of observations as well as the spread of the data. To get an average, the sum of the squares is divided by the number of observations. The result is called the "variance".

Key point

Variance = Σ (Mean − x)^2/(n). Where Σ (Mean − x)^2= the sum of the squares of all the differences from the mean

Standard deviation

Variance has limited use in descriptive statistics because it does not have the same units as the original observations. To overcome this the square root of the variance can be calculated. This is known as the standard deviation (SD) and it has the same units as the variables measured originally. If the SD is small then the data points are close to one another, whereas a large SD indicates there is a lot of variation between individual values. In fact the mean for positive values is no longer an adequate measure of the central tendency when it is smaller than the SD.

Key points

- The standard deviation is a measure of the average distance individual values are from the mean
- The bigger the variability the bigger the standard deviation
- The standard deviation is used to measure the variability of quantitative data that are not markedly skewed and do not have outliers

When the data distribution is normal, or nearly so, it is possible to calculate that 68.3% of the data lies +/− 1 SD from the mean; 95.4% between +/− 2 SD; 99.7% between +/− 3 SD (fig 4).

The "normal range" corresponds to the interval that includes 95% of the data. This is equal to 1.96 SD on either side of the mean but is often approximated to mean (2SD). If this criterion is used to diagnose health and disease, it follows that 5% of all results from normal individuals will be classified as being pathological (fig 4).

STANDARD ERROR OF THE MEAN

It is useful at this stage to consider the standard error of the mean (SEM). Usually in clinical practice we do not know what the true overall mean is. All we have are the results from our study and the standard deviation. Fortunately it can be shown mathematically that, provided our study group size is over 10 and randomly selected, the overall mean lies within a normal distribution whose centre is our study group's mean (fig 5). The standard deviation of this distribution is called the SEM. This can be

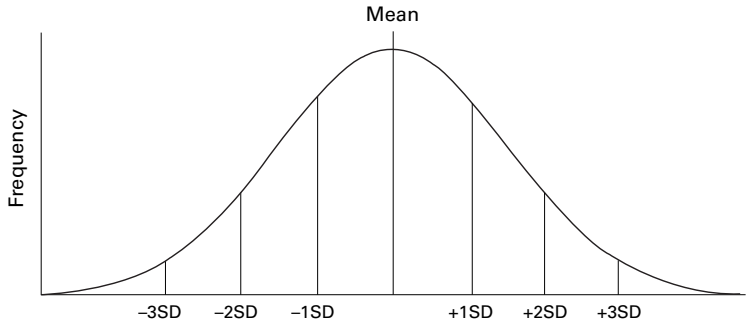

Figure 4 A normal distribution curve divided up into different multiples of the standard deviation (SD).

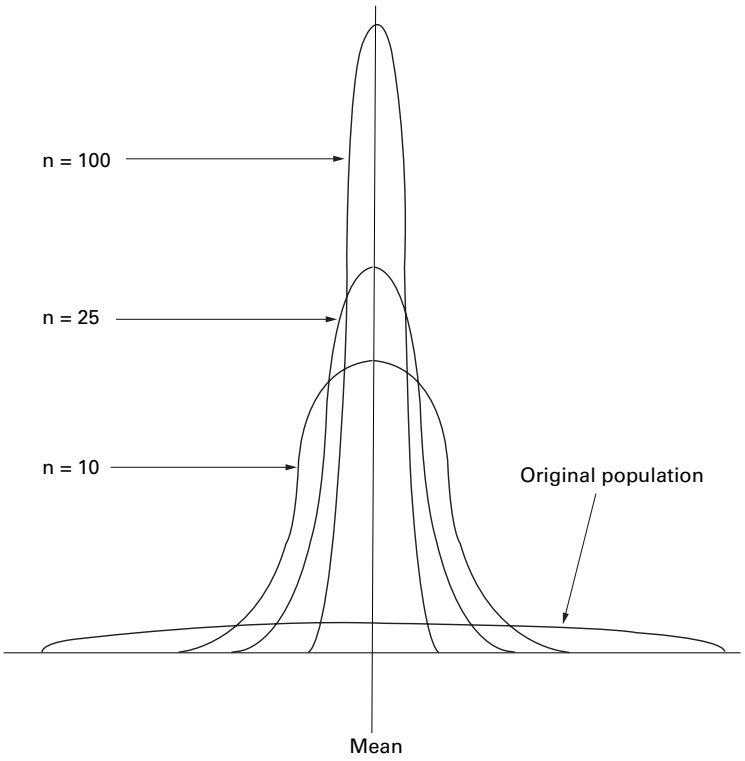

Figure 5 Distribution of the mean of samples of various sizes.

calculated by dividing the sample's standard deviation by the square root of the number of observations in the sample.

Key point
SEM = S√nD√n

As the distribution is normal, the same principles apply as found with the SD above. Consequently there is a 95% chance that the overall mean lies +/− 1.96 SEM from the sample mean.

To demonstrate these points, consider what Watenpaugh and Gaffney are saying when they write, "One hour after infusion, the capillary reabsorption of fluids was −236 +/− SEM 102 ml/h".[2]

This means we can be 95% sure that the overall mean reabsorption rate lies somewhere between:

−236 +/− (1.96 × 102) ml/h = −435.9 ml/h to −36.1 ml/h

Key points
- SEM is used when describing the overall mean, not the study group's mean
- SEM is a measure of the variance of the overall mean as estimated from the study group. It is *not* a measure of the distribution of the data.

The size of the SEM tends to decrease as the study group size gets larger or the variation within the sample becomes less marked. Consequently samples greater than 30 with little variation will have small SEM. As a result, the overall mean can be reliably pin pointed to lie within a narrow range.

So far in article 1 and 2, we have concentrated on summarising single variable datasets. In practice this is often a preliminary step to combining and comparing different collections of data.

Summarising data from two variables

Again the method used for this is dependent upon the type of data being dealt with (table 5). In subsequent articles we will discuss how each of these various types of comparisons are carried out. For now, we need to consider the important points that have to be borne in mind when presenting the summary of two variables.

CROSS TABULATION

When comparing nominal datasets, a frequency table can be drawn showing the number and proportion (percentage) of cases with a combination of variables. If there are large numbers, and marked variations, it is often better to present the data as a percentage of either the row or column totals. This makes interpretation by the reader easier. However, when using percentage, you need to indicate the total number from which they were derived. If one, or more, of the datasets are ordinal or quantitative then the groups must be arranged in the correct order (fig 6).

Table 5 Methods used to compare different types of data

	Nominal	*Ordinal*	*Quantitative—discrete*	*Quantitative—continuous*
Nominal	Cross tabulation Bar chart	Cross tabulation Bar chart	Cross tabulation Graphical	Graphical
Ordinal		Cross tabulation Bar chart	Cross tabulation Graphical	Graphical
Quantitative—discrete			Cross tabulation Graphical	Graphical
Quantitative—continuous				Scatter plot with regression and/or correlation

Triage category	Waiting time (min) 0–30	31–60	61–90	90–120	Row total (%)
Yellow	1	5			6 (12.5)
Green	11	12	10		33 (68.8)
Blue				1	1 (2.1)
Not recorded		3	5		8 (16.7)
Column total (%)	12 (25.0)	20 (41.7)	15 (31.3)	1 (2.1)	48 (100)

Number of missing observations = 1

Figure 6 Cross tabulation using an example of quantitative data (waiting times) and ordinal data (Triage category) from Egbert's study on paediatric A&E attendances.

BAR CHART

You can use bar charts when comparing nominal and ordinal datasets. In the latter case the groups must be arranged in the most appropriate order (fig 7). This comparison has to be done visually and is therefore prone to error.

GRAPHICAL REPRESENTATION

The dot plot provides a useful way of allowing ungrouped, discrete data to be visually compared. As with the bar chart, in the dot plot the values of the variable are listed on the vertical axis and the categories on the horizontal.

Figure 8 shows the values of an objective measure of muscle bulk in the hand according to a simple clinical classification of muscle wasting for 61 elderly subjects. With the dot plot each individual value is shown (fig 8 (A)). It is therefore easy to see that there is a large amount of scatter and that the distributions are slightly skewed. As described previously the box and whisker plots can be used to summarise the data and demonstrate the difference between the groups (fig 8 (B)). This shows those with most wasting tend to have smaller cross sectional areas (CSA). Further investigation reveals that when the CSA are log transformed the distribution is more symmetrical. We can therefore now use means and SD to summarise the transformed data (fig 8C). These are shown against a logarithmic scale in their original units as it makes them easier to interpret.

Graphical representation of central tendency and variance allow data to be summarised in a readily accessible format (figs 8 (B) and 9 (A)). This is commonly done by marking the central tendency and providing a measure of the variability. However, if it is important to represent each reading, a dot plot can be used. This enables the corresponding points to be joined when considering a series of readings from the same subject (fig 9(B)).

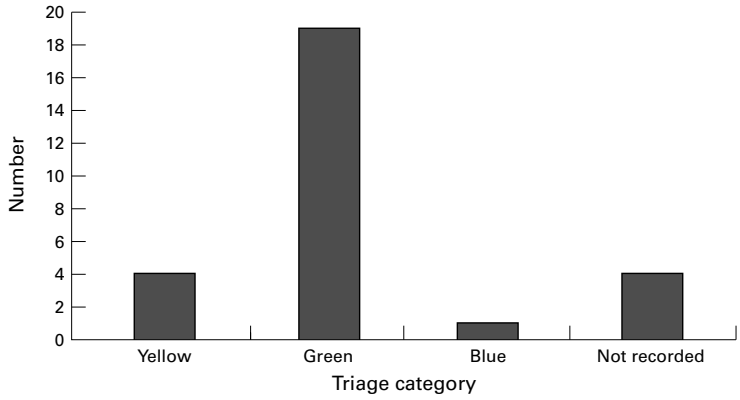

Figure 7 Frequency of male paediatric admissions in each triage category.

SCATTER PLOT, CORRELATION AND REGRESSION

A scatter plot enables the relation between quantitative-continuous variables to be demonstrated (fig 10). By convention the variable that is "controlling" the other is known as the independent variable and its value is recorded on the horizontal axis. These independent

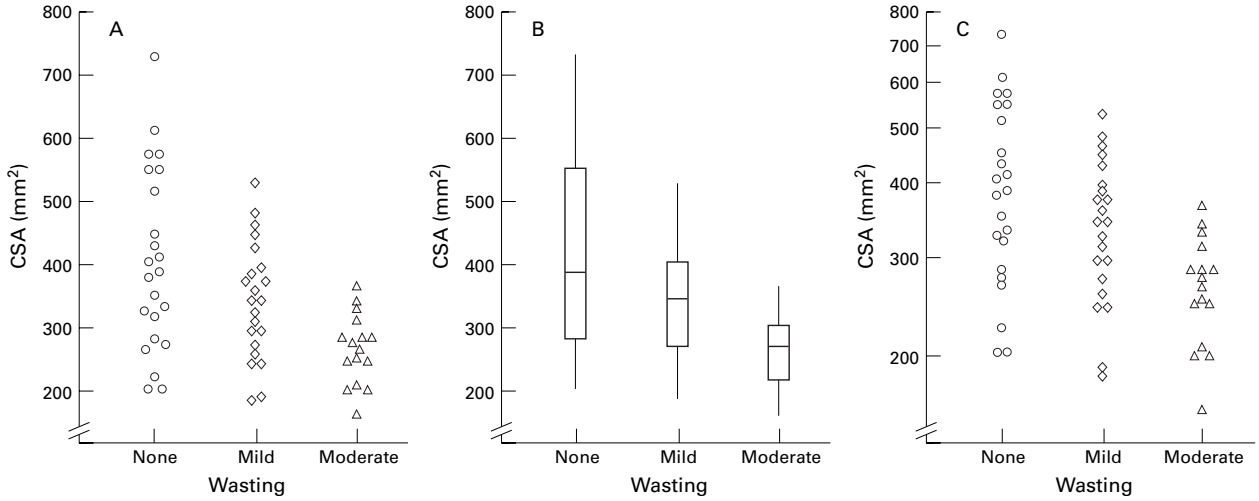

Figure 8 Three methods of showing the central tendency and variance using the same data. (A) Scatter plot. (B) Box and whisker plot. (C) Scatter plot using a logarithmic scale.

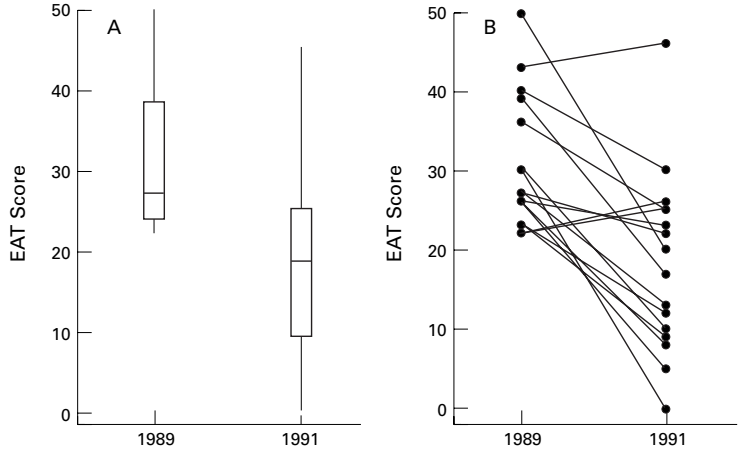

Figure 9 Two methods of presenting the same data. (A) provides a measure of the central tendency and variance using a box and whisker plot. However, if the change in each subject is needed, individual values should be plotted and joined (B).

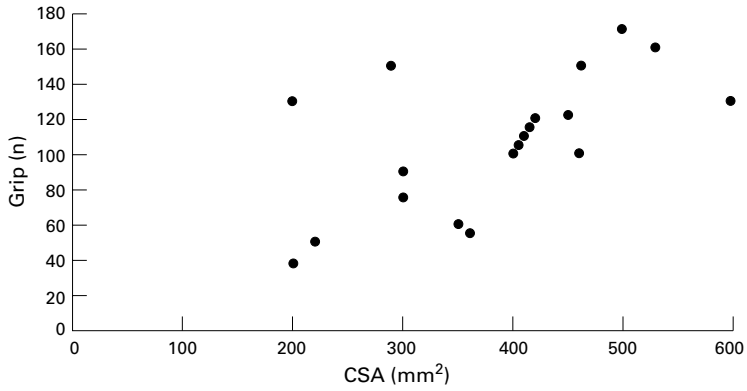

Figure 10 Grip strength and cross sectional area (CSA) of index finger-thumb web space.

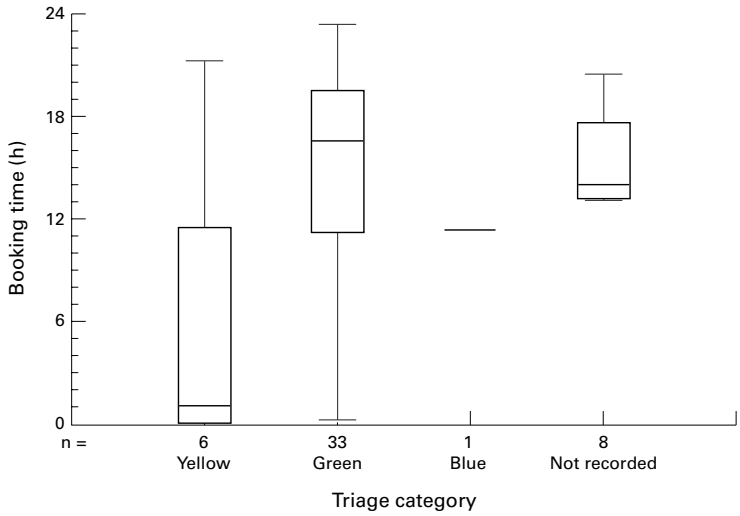

Figure 11 Box and whisker plot of booking time by triage category.

variables are those that are varied by the experimenter, either directly (for example, treatment given) or indirectly through the selection of subjects (for example, age). The dependent variables are the outcome of this experimental manipulation and they are plotted on the vertical axis.

Summary

When confronted with a vast amount of information, first consider the type of data present. Then summarise appropriately (see article 1), noting the central tendency and variability. Once this is complete it is possible to combine the information from two or more variables taking into account what types of data they are.

Quiz

(1) If data are skewed to the right, which is greater the median or the mean?
(2) Describe two main differences between the mean and the median.
(3) Consider Egbert's waiting time data listed in table 4.
  (a) What is the mean, median and mode?
  (b) What is the 75th centile?
  (c) What is the interquartile range?
(4) Figure 11 shows a box and whisker plot from Egbert's study.
  (a) What do the box's represent
  (b) What does the line in the box's represent
  (c) How would you describe the box and whisker plot of triage category (green) and booking time?
(5) One for you to do on your own.
Mullner *et al* carried out an experiment to look at myocardial function following successful resuscitation after a cardiac arrest.[3] Part of their results are shown in table 6:
  (a) Identify the different types of data
  (b) Identify the central tendency and variability for each variable
  (c) Present a tabular summary of the patient's ages
  (d) What graphical summary would you use to compare the frequency of VF with site of the myocardial infarction?
  (e) What graphical summary would you use to show the comparison between the site of the myocardial infarction and the duration of the arrest?
  (f) What method would you use to show the comparison between the systolic blood pressure and duration of arrest?

Answers

(1) The mean
(2) The mean reflects all the data points whereas the median uses only the middle ones. Secondly, unlike medians, small group means can be combined to give an overall mean.
(3) (a) Mean, median and mode:
Mean = 1100/48 = 22.9 minutes
Median = 49/2nd data point = half way between the 24th and 25th data point = 19 minutes
Mode = 19 minutes
  (b) The 75th centile equals: 75× 49/100th data point = 36.75th data point. This means it is 0.75 of the way from the 36th to the 37th data point. However, using table 4, both data points are 27 minutes.
The 75 th centile is therefore 27 minutes.
  (c) The interquartile range equals:
This is the range between the 25th and the 75th centile

Table 6

Patient	*Age (y)*	*ECG*	*Systolic BP (mm Hg)*	*Duration of arrest (min)*	*Epinephrine (mg)*	*MI location*
1	44	VF	98	15	2	Anterior
2	58	Asystole	112	2	0	None
3	57	VF	100	12	0	None
4	62	VF	90	27	4	Inferior
5	52	VF	110	25	10	Inferior
6	65	VF	117	48	6	Anterior
7	42	VF	113	13	0	Inferior
8	46	VF	105	9	0	None
9	59	VF	180	15	1	Inferior
10	56	VF	97	47	8	Anterior
11	73	VF	114	18	1	None
12	60	VF	165	25	2	Subendocardial
13	80	PEA	115	29	2	None
14	62	Asystole	85	26	7	None
15	58	VF	104	15	6	Inferior
16	50	VF	101	20	4	None
17	72	VF	100	32	3	Inferior
18	46	VF	120	23	2	Subendocardial
19	72	VF	138	22	3	Inferior
20	50	VF	102	42	3	None

MI = myocardial infarction; VF = ventricular fibrillation; PEA = pulseless electrical activity.

Using the process described in 3 (b), the 25th centile is 15 minutes.
The interquartile range is therefore 15–27 minutes.

(4) (a) The middle 50% of the data

(b) The median

(c) Negative skewed distribution (long tail on the side with the lowest numbers)

The authors would like to thank Sally Hollis and Alan Gibbs who helped with an earlier version of this article and the invaluable suggestions from Jim Wardrope, Paul Downs and Iram Butt.

1 Bowlers D. Measures of spread of nominal variables. In: *Statistics from scratch.* Chichester: John Wiley, 1996:118–23.
2 Watenpaugh D, Gaffney F. Measurement of net whole-body transcapillary fluid transport and effective vascular compliance in humans. *J Trauma* 1998;**45**:1062–8.
3 Mullner M, Domanovitis H, Sterz F, *et al*. Measurement of myocardial contractility following successful resuscitation: quantitated left ventricular systolic function utilising non-invasive wall tress analysis. *Resuscitation* 1998;**39**:51–9.

Further reading

Altman D, Bland J. Quartiles, quintiles, centiles and other quantiles. *BMJ* 1994;**309**:996.
Bowlers D. Measure of average. In: *Statistics from scratch.* Chichester: John Wiley, 1996:83–113.
Coogan D. *Statistics in clinical practice.* London: BMJ publication, 1995.
Gaddis G, Gaddis M. Introduction to biostatistics: part 2, descriptive statistics. *Ann Emerg Med* 1990;**19**:309–15.
Glaser A. Descriptive statistics. In: *High-yield biostatistics*. Philadelphia: Williams and Wilkins 1995:1–18.

Funding: none.

Conflicts of interest: none.

AN INTRODUCTION TO STATISTICS

An introduction to statistical inference—3

P Driscoll, F Lecky, M Crosby

Objectives
- Discuss the principles of statistical inference
- Quantifying the probability of a particular outcome
- Discuss clinical versus statistical significance

In covering these objectives we will introduce the following terms:
- Population and sample
- Parameter and statistic
- Null hypothesis and alternative hypothesis
- Type I and II errors

The previous two articles discussed summarising data so that useful comparisons can be made. Another common problem encountered is estimating a value in a larger group based upon information collected from a small number of subjects. To see how statistics can be used to achieve this, it is helpful to begin by reviewing the meaning of five commonly used terms:
- Population and sample
- Parameter and statistic
- Element

The word "population" describes a large group that includes every possible case. In contrast, a "sample" is a smaller group of subjects who are part of the population. Therefore the population of UK emergency departments would have every emergency department in the UK whereas those in the north west would represent a sample.

A value measured in a population is known as a "parameter". Consequently the trolley waiting time in UK emergency departments would be a parameter. The term "statistic" is used to denote the same variable when it is measured in a sample. Finally each separate observation in either a population or sample is called an "element" and it is often labelled with the letter X. The number of elements in a population is given the letter N and in a sample, n.

Key point
A population contains all the elements from X_1 to X_N and a sample has n of the N elements.

It is often not possible to record all the elements of a population. For example, a study investigating the peak flow in asthmatic patients attending UK emergency departments cannot review every patient. However, it is feasible to record the peak flow in a sample of asthmatic emergency department attendances. From this statistic an estimation of the population's peak flow can be inferred. The name given for manipulating data in this way is therefore called inferential statistics. It can also be used to make estimations about a sample based upon information from a population.

In carrying out inferential statistics it is important to ensure that samples are representative of the whole population and have been randomly selected. If this is not the case, bias will be introduced and a perverse answer could result. For example, inferential statistics could be used for making a national generalisation following a survey on the waiting times in 20 emergency departments. However, problems would arise if the sample did not represent the population. For example, if the investigation looked at district general hospital emergency departments in London then it is unlikely to be an accurate reflection of all the emergency departments in the UK.

It is also important that each subject has an equal chance of being included in the sample. Consequently, if the trolley waits for elderly patients was being studied, all such times need to be recorded not just those measured when there is an apparent delay. A possible way of achieving this goal is by random sampling. This is a method of selecting subjects such that each member of the population has an equal and independent chance of being picked. A variety of techniques can be used including flicking a coin, drawing numbers from a hat and reading from random number tables. However, by themselves, these techniques may not be adequate because populations can be made up of different types of groups of various sizes. This heterogeneity could have a bearing on the study. For example the stage of a disease, and the age or sex of a patient may change the response to a particular drug. As these are not necessarily evenly distributed in a population it is important they are adequately covered in the sampling process. This is achieved by stratified random sampling in which the population is initially divided into homogeneous groups from which random samples are taken. In this way representative samples of the whole population can be obtained.

Allowing for uncertainty

Measurements on people vary because we are not all the same. Therefore the peak flow in 20

Accident and Emergency Department, Hope Hospital, Salford M6 8HD
P Driscoll
F Lecky
M Crosby

Correspondence to:
Mr Driscoll, Consultant in Accident and Emergency Medicine (pdriscoll@hope.srht.nwest.nhs.uk)

Accepted 22 June 2000

Key points
- *Population* This is a complete group (that is, having every eligible person (or item) with a particular characteristic). Depending on the study the actual size of a population varies from modest (for example, all minor injury units in a particular region) to huge (for example, all emergency treatment centres in Europe).
- *Sample* This is a subset of the population that has been collected for a study. As with the population, the size of a sample can vary.
- *Parameter* This is the value of a variable in a population.
- *Statistic* This is the value of a variable in a sample.
- *Element* This is a single observation.
- *Statistical inference* This enables statements to be made about a sample based upon a population's parameter. It also allows the converse to occur but in this case it is dependent upon random, representative samples being taken.

randomly selected adults from your waiting room would vary over a range of values. This would be the case even if all the subjects were perfectly healthy. We could reduce this range by being very selective about who we measured, for example being male, 1.75 meters tall, good looking and from Yorkshire! However, even if you managed to find 20 such subjects, the peak flows would still vary.

A useful way of considering this variation is to plot it as a frequency distribution. When the values from a reasonably homogeneous group are shown in this way the curve takes upon the shape of a "normal distribution".[1] This is because many of the recordings are clustered around the mean value with a symmetrical fall off in the frequency of recordings as you move from the centre.

Inferential statistics enables us to take account of these variations when estimations about a parameter or statistic are been made. It does this by quantifying the probabilities of the possible outcomes.

Probability

In view of the variations discussed previously we cannot predict with absolute certainty the outcome of an event prior to it happening. This uncertainty is often referred to as "a matter of chance". It is however possible to quantify this uncertainty by calculating the probability of an event occurring.

Key point
Probability is the proportion of cases in a study that have a particular result.

Probability (p) is invariably expressed as a decimal value between 0 and 1, where zero means that an outcome will never happen and 1 means it will always occur. Therefore, if 30% of patients survive after a cardiac arrest in hospital, the probability of survival would be written as 0.3. As it cannot be larger than 1, it follows that the probability of an event not occurring is 1−p. Therefore the probability of not surviving a cardiac arrest is 1−0.3 = 0.7.

Key points
- Probability values are expressed as a decimal from 0 to 1
- The probability of an event occurring cannot be negative
- The probability of an event not occurring is 1−p.

Probabilities are often used to help guide management. For example after a head injury, a patient with no skull fracture and no neurological signs has a 0.00017 probability of developing an operable intracranial haematoma. This is such a low number that we tend to discharge these patients under the care of a sensible adult.

When using probabilities in determining clinical decision making, one tends to err on the side of caution so that no cases are missed. This is seen with the Ottawa knee protocol (box 1). A level is chosen where the probability of missing a fracture is zero. Consequently radiographs are reserved for those patients with particular presenting symptoms.

Box 1 Ottawa knee rule[2]
Order radiography of the knee if any of the following are present:
- Patient older than 55 years
- Tenderness at the head of the fibula
- Isolated tenderness of the patella
- Inability to flex to 90 degrees
- Inability to transfer weight for four steps both immediately after injury and in the A&E department

COMBINING PROBABILITIES

Though a single finding, or test, can help in clinical decision making, in practice we often rely on several results before making a diagnosis. This process entails combining the probabilities of each of the possible outcomes.

The chance of a particular outcome in any of the tests carried out is equal to the sum of all the probabilities. For example, if the probability of a patient in your waiting room being blood group O is 0.46, the chance of either of two unrelated patients being blood group O is 0.46 + 0.46 = 0.92. This is known as the additional rule of probability and it can be written as: Probability of outcome A or B = probability of outcome A + probability of outcome B

To calculate the probability of a specific combination of independent outcomes occurring (for example, the probability of outcome A and B), the separate outcome probabilities need to be multiplied together. Therefore, the probability of both patients being blood group

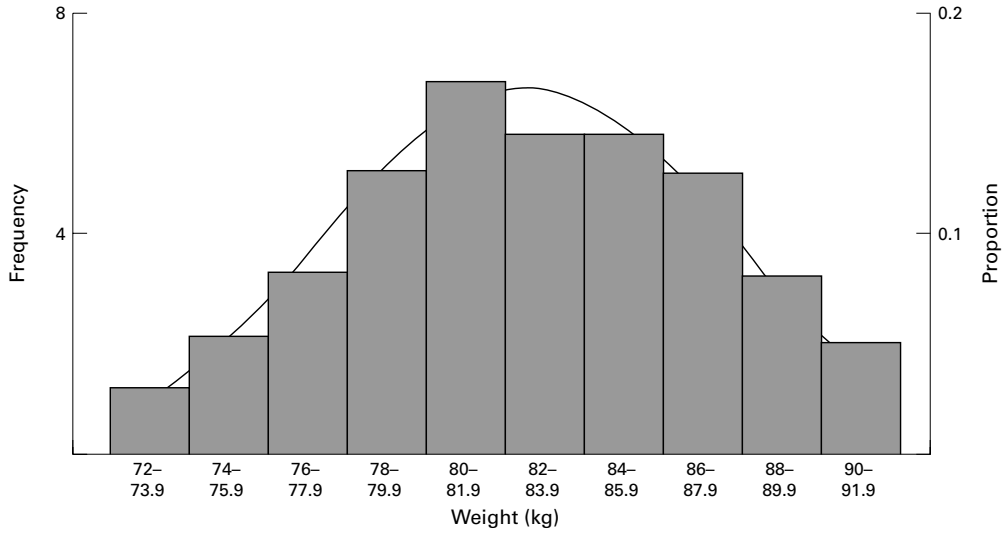

Figure 1 Weight of staff in the emergency department of Deathstar General.

O is $0.46 \times 0.46 = 0.21$. This is the multiplication rule of probability and it can be written as: Probability of outcome A and B = probability of outcome A × probability of outcome B

However, the method used to calculate the chance of a particular combination varies with the independence of the outcomes. Independence in this context means the chances of a particular result will not make another outcome more or less likely. In the example given above this obviously applies—a particular patient's blood group will not change the chances of an unrelated patient being of certain blood type.

It also follows that if the outcomes are not independent then the multiplication probability will not apply. This can be used to detect factors that are related.[3] To demonstrate this assume the probability of loosing a finger in your community is 0.01 and the probability of working at "Happy crusher" the local steal works is 0.1. If these were independent of one another the probability of working at Happy crusher and loosing a finger should be:

$$0.1 \times 0.01 = 0.001$$

(that is, one chance in a 1000)

If you find out that in practice the figure is 0.01 you would suspect there is a connection and the factors are not independent.

In clinical practice we are often dealing with outcomes that are not mutually exclusive.[4] Consequently you usually need to take into account the probability of a combination occurring. This can be calculated by modifying the equation above to: Probability of outcome A or B = probability of outcome A + probability of outcome B−probability of outcome A and B.

Table 1

	Telephone questionnaire		
	Yes	*No*	*Total*
Emergency admission			
Yes	18	42	60
No	12	28	40
Total	30	70	100

The following problem helps to demonstrate these points. Let us say the probability of a person having an emergency admission to hospital at some stage in their life is 0.6. They also have a 0.3 probability of being asked to complete a telephone questionnaire in the same time span. If these results were completely independent of one another, the probability of having an: Emergency admission and a telephone questionnaire = $0.6 \times 0.3 = 0.18$

Consequently the probability of having either an: Emergency admission or a telephone questionnaire = 0.6 + 0.3−0.18 = 0.72

To illustrate what these figures mean it is helpful to use a 2×2 table after converting the probabilities into actual numbers. This is done by assuming we are dealing with 100 people (table 1).

From table 1 you can see that 18 people will have both an emergency admission and a telephone questionnaire some time in their life. Seventy two will have one or both. This number is the total of those having a telephone questionnaire only (12) plus people having an emergency admission only (42) plus those having both an emergency admission and a telephone questionnaire (18).

There is another method of calculating probabilities when dealing with data that have only two possible outcomes. Examples of such binomial data include live or die; boy or girl, success or failure. Consequently the outcomes are mutually exclusive. The probability of a specific combination of these outcomes can be determined by use of the binomial probability tables.[5] These list the chances of obtaining each of the possible outcomes.

As binomial distributions deal specifically with combinations of independent, mutually exclusive events, they are often not applicable to emergency medicine. In contrast, they are used in genetic counselling where some inheritance disorders, such as Tay-Sachs disease, follow a binomial distribution. Koosis provides a comprehensive account for those who wish to learn more about this topic.[5]

DISPLAYING PROBABILITIES

In article 1 we discussed how a frequency distribution could be used to show graphically the number of cases at each particular value of a variable. This is demonstrated in the following example. The specialist registrar at Deathstar General, Dr Egbert Everard is concerned about the health of the 40 male staff in the emergency department. He therefore decides to weigh them and plot the results as a frequency distribution (fig 1). Probabilities can be demonstrated in a similar way. To do this Egbert divides the number of cases in each weight category by the total number of cases in the whole study (that is, 40). This gives him the proportion of cases at each value. These values can then be joined up to produce a distribution curve (fig 1).

Key point
You can express the data in a frequency distribution as a distribution of probability.

It is possible to use these distribution curves to calculate the probability of having a value equal to, or greater than, a particular number. For example the proportion of staff with a weight greater than, or equal to, 80 kg is represented by the area under the curve to the right of 80 kg mark (fig 2). Probability distributions have a further useful property in that the area under the whole curve is equal to one. This is because it represents the sum of all the possible probabilities. Consequently the proportion of staff with a weight less than 80 kg is represented by the area under the curve to the left of the 80 kg mark. This is equal to [1−shaded area].

Statistical inference in medical studies commonly use probabilities in this way to test the null hypothesis.

Testing the null hypothesis

Consider what you would do if asked to make recommendations for your emergency department on a new drug for asthma care following a successful trial. Firstly, you would need to be to sure the patients were representative and randomly chosen. Secondly, any difference in effect attributable to the new treatment would need to be judged in the light of the differences between patients simply attributable to chance variation.

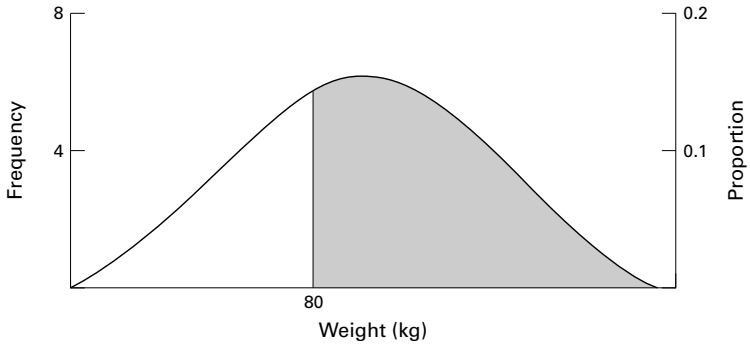

Figure 2 Distribution curve of the proportion of staff with a weight greater than or equal to 80 kg.

We have seen that the probabilities of various outcomes can be quantified using statistical inference. However, it is not practical to test all of the infinite number of possible differences between the population and sample. Consequently only the possibility of there being no difference between the population and sample is tested. It is then feasible to determine the probability that a difference equal, or greater, to that found in the study could be attributable to normal variation. This is known as testing the null hypothesis.

Key point
The null hypothesis states that the difference between the groups being tested is attributable to chance variation.

The probability of the null hypothesis being correct is called the p value, a frequently used term in medical journals. For example, in a study comparing the rehabilitation time after ankle sprains with new and standard treatment, it was found that the mean difference was four days ($p = 0.01$). Consequently the chance that a difference this big, or bigger, occurring when the null hypothesis is correct is 1 in a 100. This means that it is more likely that there is a difference between the two treatments. This is called the alternative hypothesis.

Key point
The actual p value should be provided to two decimal places.

Nowadays the p value is calculated by computer but the statistical tests used to work it out depend upon the data in question and the type of study. The choice of test is therefore important so that meaningful results are obtained.

Key point
The p value is derived from the raw data using statistical calculations and tables appropriate to the test carried out.

Later in this series the tests commonly used in emergency medicine will be described so that you will be able to choose the correct one. At this point however it is useful to test our understanding of the role of the null hypothesis and p values by considering the results of a recent publication. Sunde *et al* compared the time from turning on the monitor to starting chest compression in different types of cardiac arrest.[6] In cases of asystole, the median time delay was 29 seconds. This was significantly shorter than the time found in patients with electromechanical dissociation (EMD) (109 seconds, $p < 0.001$). What does this mean?

The null hypothesis in this study is that the time delays before cardiopulmonary resuscitation in patients with asystole and EMD are the same. However, the p values indicate that the

probability of a difference of 80 seconds being attributable to chance is less than one in a thousand. It is therefore more likely that the alternative hypothesis is correct and there is a difference between these two groups.

STATISTICAL SIGNIFICANCE

Rejecting the null hypothesis means that a "significant difference" exists between the populations studied that cannot be explained by chance alone.

Statistical methods used to test the null hypothesis are termed "tests of significance"

A 2×2 table can be constructed for the four possible outcomes of the null hypothesis (NH) tested (table 2).

TYPE I ERROR

These mistakes occur when statistical tests indicate that the null hypothesis is unlikely (that is, the p value is low) but in actual fact there is no difference between the study groups. By convention, "statistical significance" is accepted if the chance of making such an error is less than 0.05. When these arbitrary levels are given for a study they are often referred to as α. Consequently when $\alpha = 0.05$ it is considered tolerable to make a "type I" mistake 1 in 20 times. However, the level considered significant is determined by you the investigator. It may be that a lower value of α is required when testing the use of an expensive or potentially toxic treatment. In this way the chances of falsely rejecting the null hypothesis can be kept small. In these cases you may wish to use a value of 0.01 (that is, a 1 in a 100 chance of falsely rejecting the null hypothesis) rather than the usual 0.05.

In considering the arbitrary level demarcating type I errors, it is also important to be aware that the value for p is markedly affected by both the sample sizes and the magnitude of any difference (that is, the point estimate). This is demonstrated in table 3. In all cases the p value is 0.05 but the difference and sample size vary. For very large samples the difference only has to be small to produce a statistically significant result. The converse applies when the sample is small. As will be discussed later in this series, the p value is also affected by the standard deviation of the distribution.

TYPE II ERROR

These represent mistakes in falsely accepting the null hypothesis and is represented by β. If β is large there is a high chance of making a type II error. As this is the opposite of what we want, the reciprocal of the term is often used in instead. This is known as "power" and is equal to 1−β. Consequently a test with a high power has a low chance of making a type II error. Conventionally, a study is required to have a power of 0.8 to be acceptable. In other words the study should have an 80% chance of being able to detect if the null hypothesis did not apply.

Four factors effect the probability of making a type II error (box 2)

Table 2

	Reality	
	True difference	*No difference*
Statistical test		
NH rejected	Rejection correct	Rejection incorrect (type I error)
NH accepted	Acceptance incorrect (type II error)	Acceptance correct

*Table 3 The effect of sample size and difference on the p value when comparing two groups (assuming a constant standard deviation)**

Sample size of each group	*Difference (in units)*	*p Value*
4	10.0	0.05
25	4.0	0.05
400	1.0	0.05
2500	0.4	0.05
10 000	0.2	0.05

*Adapted from a theoretical study by Norman *et al* where they compared the difference between the IQ in the normal population and those reading their statistic book.

Key points

- It is possible to reduce the chances of making type I error by using a lower α level.
- The p value is affected by the size of the difference, the number of cases in the sample and the standard deviation.
- In a large study, you are very likely to get a "statistically significant" result.
- In a small study it is very hard to get a "statistically significant" result. Consequently a p value greater than 0.05 in this situation proves nothing.

Box 2 Factors affecting the power of a study

- Size of α
- Variability of the sample
- Size of the sample
- Point estimate

When α increases you are less likely to accept the null hypothesis. Consequently the chances of making a type II mistake fall. A balance therefore has to be struck so that the chances of both type I and II errors are kept as small as possible.

We have already discussed that inferential statistics are used to take account of variations in statistics when a parameter is being estimated. If the variations are large then possible values for the parameter will also cover a wide range. In such circumstances the chances of rejecting the null hypothesis are reduced. Conversely factors that decrease variability will increase the power of the study. Consequently, as increasing the sample size leads to a fall in variability, it will also reduce the chances of making a type II error.

Key point
Increasing the sample size is one of the commonest ways of increasing the power of a study.

The size of a difference between the study groups (for example the control and the experimental groups) will directly effect power. An increase leads to greater power as there is less chance of falsely accepting the null hypothesis.

To help understand these principles of null hypothesis testing, consider a follow up study carried out by Egbert. He was particularly concerned about how overweight the male personnel were in the emergency department. He therefore set up a study with the null hypothesis being that the mean weight of fit, healthy men and departmental men was the same. Having weighed all 40 of them, he found the mean weight was 87 kg. Figure 3 shows the normal probability distributions of the two populations. Population 1 all had the characteristic of being fit and healthy whereas population 2 were unfit couch potatoes. Egbert's finding of a mean of 87 kg could therefore lie in either distribution. The chances of a weight this big, or heavier and being part of the population 1 is shown by the darker shaded area. This represents α—that is, the probability of making a type I error and falsely rejecting the null hypothesis. Conversely the chances of being part of population 2 and having a weight this big, or lighter is shown by the lighter shaded area. This represents β—that is, the probability of making a type II error and falsely accepting the null hypothesis.

The four factors mentioned above are used in an equation to calculate the power of a study. However, if you are setting up a study you can set the power at a particular level (often 80%). Therefore, if the size of the other variables are known (that is, α, variability and point difference), the same equation can be used to determine how many subjects are required in the study.

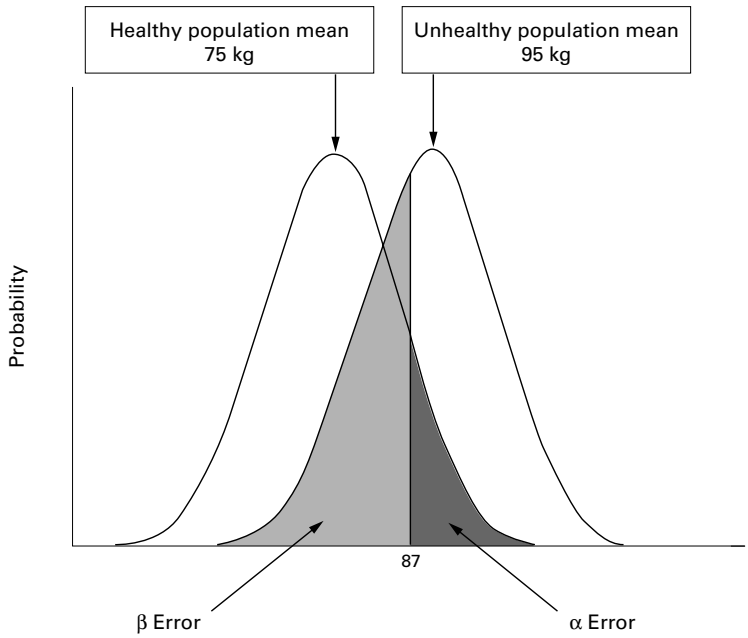

Figure 3 Distribution curves of the weights of healthy and unhealthy men.

CLINICAL VERSUS STATISTICAL SIGNIFICANCE

Consider a study comparing a new antihypertensive medication (A) with a standard one (B). The result of the trial shows that the blood pressures in patients receiving A were significantly lower than those on B ($p = 0.0001$). This means the probability that the difference found, or bigger, being attributable to chance is 1 in 10 000. In a well run study we would have no problem in accepting this as a statistically significant result. However, this does not mean it is clinically useful. In the same study the point estimate between the groups was 5 mm Hg. From a clinical point of view we may consider this to be too small to offset the difficulties, side effects and expense associated with the drug A.

As stated before, accepting a p value of 0.05 to reject the null hypothesis may not be appropriate in some clinical settings. Clinical considerations also have to be considered when accepting the null hypothesis if the p value is greater than 0.05. You need to take into account the type of study carried out, the number of subjects in each group and the weight of other published data. A further point that should be remembered is that in clinical practice we usually need to know the presence and size of any difference. p Values only inform you on the likelihood of a difference being attributable to chance (that is, normal variation).

Key points
- The p value answers the question, "Is there a statistically significant difference between the study groups?"
- Clinical issues need to be considered along with the size of the p value
- The size of any difference needs to be known
- Statistical significance does not necessarily mean clinical significance

In the majority of cases these limitations with the p value can be overcome by using confidence intervals. This will be discussed further in the following article.

Summary
Statistical inference is used to make comments about a population based upon data from a sample. In a similar manner it can be applied to a population to make an estimate about a sample. It is commonly seen in medical publications when the null hypothesis is being tested. This calculates the probability (p value) of a type I error—that is, that a particular finding is attributable to chance. It is also important to be aware of the chances of a type II error—that is, accepting the null hypothesis when it does not apply. Sample size, point estimate and variability are common factors that will affect the

chances of making these two types of errors. Interpreting results therefore needs to take these factors into account as well as the clinical relevance of the findings. Statistical significance does not necessarily mean clinical significance.

Quiz

(1) Complete the following phrase:
A parameter is to a —— as a ——- is to a sample.

(2) You are told that the probability of a female patient having a fractured femur is 0.3 and green eyes is 0.4. Assuming these are independent of one another, what is the probability that she has:
Both a fractured femur and green eyes
Either a fractured femur, or green eyes or both

(3) Name three factors that will affect the chances of making a type I and II error.

(4) A new thrombolytic "Dyno-coronary" has been developed. Though very expensive and toxic it is thought to produce coronary patency quicker than standard treatment. If you were to design a study to assess this what α level would you choose—0.05 or 0.01?

(5) One for you to do one your own. Formulate the null hypothesis for the study by Ireland *et al* that investigated whether supine oblique views provide better imaging of the cervicothoracic junction than a swimmer's view.[7] Consider the conclusion drawn with respect to statistical and clinical relevance.

Answers

(1) A parameter is to a statistic as a population is to a sample.

(2) $0.3 \times 0.4 = 0.12$
$0.3 + 0.4 - 0.12 = 0.58$

(3) Point estimate, sample variability, sample size

(4) You would be aiming to minimise the chances of making a type I error, consequently an α level of 0.01 would be preferable.

The authors would like to thank John Heyworth, Sally Hollis, Jim Wardrope and Iram Butt for their invaluable suggestions.
Funding: none.

Conflicts of interest: none.

1 Driscoll P, Lecky F, Crosby M. An introduction to everyday statistics—1. *J Accid Emerg Med* 2000;**17**:205–11.
2 Stiell I, Greenberg G, Wells G, *et al*. Prospective validation of a decision rule for the use of radiography in acute knee injuries. *JAMA* 1996;**275**:611–15.
3 Altman D. Theoretical distributions. In: *Practical statistics for medical research.* London: Chapman and Hall, 1997:48–73.
4 Coggon D. Probability. In: *Statistics in clinical practice.* London: BMJ Publishing Group, 1995:52–61.
5 Koosis D. Populations and samples. In: *Statistics— a self teaching guide.* New York: John Wiley, 1995:40–76.
6 Sunde K, Eftestol T, Askenberg C, *et al*. Quality assessment of defibrillation and advanced life support using data from the medical control module of the defibrillator. *Resuscitation* 1999;**41**:237–47.
7 Ireland A, Britton I, Forrester A. Do supine oblique views provide better imaging of the cervicothoracic junction than swimmer's views? *J Accid Emerg Med* 1998;**15**:151–4.

Further reading

M, Gaddis G. Introduction to biostatistics: Part 1, basic concepts. *Ann Emerg Med* 1990;**19**:143–6.
Glaser A. Inferential statistics. In: *High-yield statistics.*. Philadelphia: Williams and Wilkins, 1995:9–30.
Normal G, Streiner D. Inferential statistics. In: *PDQ statistics.* 2nd ed. London: Mosby, 1997:17–36.

Correction

We regret that two errors occurred in the statistics paper published in July (2000;**17**:274–81). On page 277, 2nd column, +/− signs were omitted from 2SD in the penultimate paragraph. On page 278, 1st column, the equation was incorrect; it should have read: $SEM=SD/\sqrt{n}$