

## 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".<sup>1</sup> 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<sup>2</sup>**

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

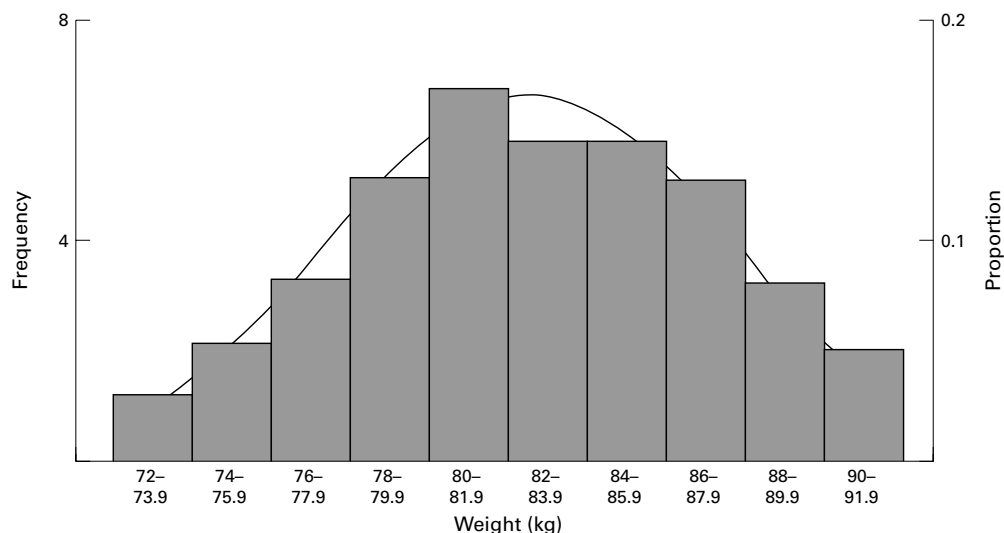


Figure 1 Weight of staff in the emergency department of Deathstar General.

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  $\times$  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.<sup>3</sup> To demonstrate this assume the probability of losing 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 losing 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.<sup>4</sup> 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.

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 2x2 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.<sup>5</sup> 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.<sup>5</sup>

Table 1

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

## 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 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.

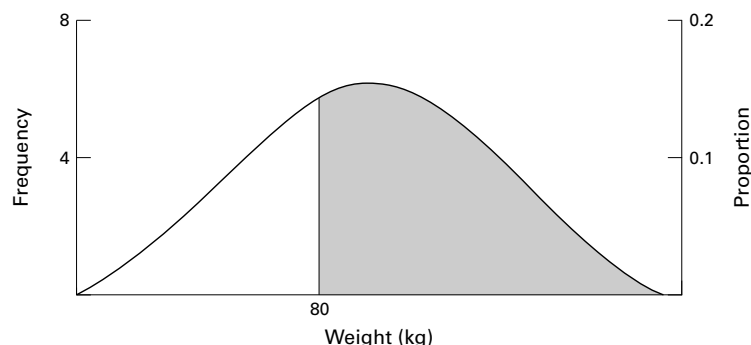


Figure 2 Distribution curve of the proportion of staff with a weight greater than or equal to 80 kg.

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.<sup>6</sup> 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 2x2 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  $\alpha$ . 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  $\alpha$  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  $\beta$ . If  $\beta$

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  $\alpha$  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.

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 instead. This is known as “power” and is equal to  $1-\beta$ . 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)

Box 2 Factors affecting the power of a study

- Size of  $\alpha$
- Variability of the sample
- Size of the sample
- Point estimate

When  $\alpha$  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  $\alpha$ —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  $\beta$ —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

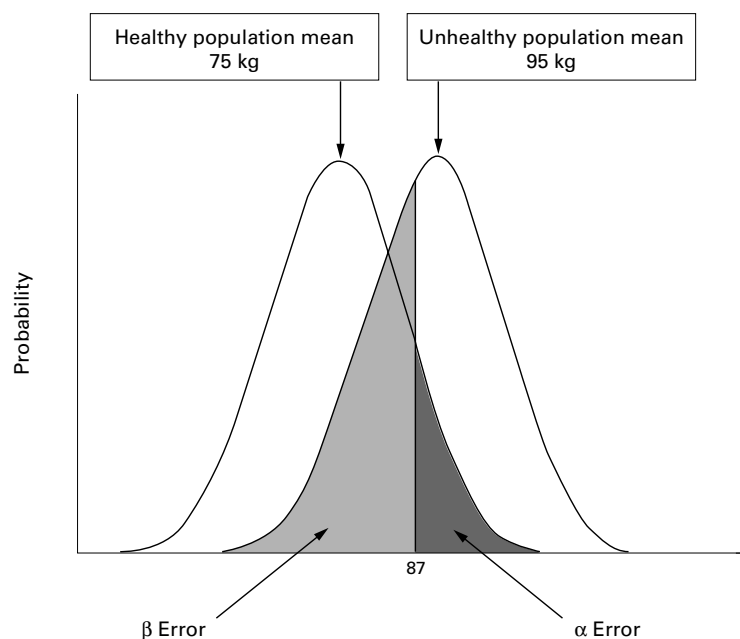


Figure 3 Distribution curves of the weights of healthy and unhealthy men.

variables are known (that is,  $\sigma$ , variability and point difference), the same equation can be used to determine how many subjects are required in the study.

**CLINICAL VERSUS STATISTICAL SIGNIFICANCE**

Consider a study comparing a new anti-hypertensive 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  $\alpha$  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.<sup>7</sup> 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  $\alpha$  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:  $SEM = SD/\sqrt{n}$