AN INTRODUCTION TO STATISTICS

Statistics is all pervading. We come across it at work (how can I convince the Trust board I need more staff?); in the journals (why did they use that polysyllabic test?) and socially (am I abnormal if I have only kissed my mother?).

Though the subject is important, reading "How to" statistical articles is up there with root canal work as activities to avoid. Tragically though, like dental repair, it happens to all of us at some time in our life.

Part of the problem is the bad experiences incurred in trying to gain a grasp of a subject that is generally taught in a strange language divorced from the reality of emergency medicine. In an attempt to correct this we have written a series of articles from the perspective of an emergency clinician who knows nothing about the statistics.

The first four articles give an overview, explaining common terms and showing how statistics is used in everyday practice. After that we move into a fuller description of the common tests encountered in the journals. In doing this we aim to allow you to:

- Have a clearer grasp on what statistics can do
- Assess the tests selected by authors before accepting their results
- Understand what a statistician will be asking for when you see them about a study

Above all we want you to have a pain free time reading the articles. To this end numerous examples are provided along with a short quiz at the end of each article to test you own understanding of what has been written. If parts are unclear we want to hear from you.

An introduction to everyday statistics—1

P Driscoll, F Lecky, M Crosby

Objectives

- Define statistics
- Discuss the types of data commonly encountered in accident and emergency (A&E) work
- Describe the techniques used to summarise a single dataset

In covering these objectives we will deal with the following terms:

- Variable
- Discrete and continuous
- Frequency distribution
- Grouping
- Transformation

Statistics is defined as a process by which numerical data are transformed into a usable form for scientific interpretation. This entails manipulating data to summarise the findings (descriptive statistics). It can also be used to develop general conclusions from the data (inferential statistics) (fig 1).

It is useful to start this series by dealing with descriptive statistics because:

- You deal with them everyday in the A&E department.
- It will introduce the importance of knowing what type of data you are handling
- The majority of statistical analysis starts by summarising data

To demonstrate this consider the following problem:

Dr Canute is a consultant in A&E medicine at Deathstar General. A recent visit by the audit commission has found the waiting times for paediatrics are too long and need to be cut. Galvanised into action, he asks his new specialist registrar Egbert Everard to investigate. Happy to help out, Egbert lists a number of factors he considers important. These are known as variables because they are measurements that vary between individuals.[1] He asks for a print out of these variables on all paediatric

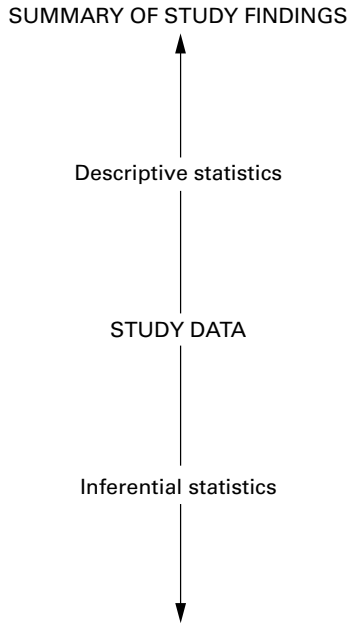


Figure 1 Relation between descriptive and inferential statistics.

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

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

Accepted 13 December 1999

Table 1 Paediatric A&E admissions in one day at Deathstar General

Case	Age (y)	Sex	Complaint	Triage category	Diagnosis	Arrival time	Waiting time (min)
1	3	M	Medical	Yellow	Poisoning	00:09	46
2	3	M	Medical	Yellow	Poisoning	00:10	45
3	4	M	Medical	Yellow	Poisoning	00:11	44
4	2	F	Trauma	Green	Head injury	00:24	26
5	2	F	Medical	Yellow	URTI	02:08	52
6	1	M	Orthopaedic	Green	Limp? cause	07:46	14
7	11	F	Trauma	Green	Bruised foot	08:51	19
8	6	F	Trauma	Green	FB finger	09:21	18
9	2	M	Surgical	Green	Non-specific abdominal pain	09:41	14
10	13	M	Trauma	Green	Thumb	09:42	34
↓	↓	↓	↓	↓	↓	↓	↓
48	4	M	Medical	Yellow	URTI	23:50	41

arrivals in the A&E department over the past 24 hours. A button is pressed and several yards of computer printout are produced (table 1).

Key point A variable is something whose value or quality can vary

Obviously there is a lot of information here but in its current form it is not very usable. To make this more manageable, and useful, Egbert needs to collate and summarise the data. This is best done systematically by:

Firstly, identify the types of data used to measure the variable

Secondly, summarise the data for each variable

Thirdly, compare the summaries of different datasets

In this, and the following article, we will see how this can be achieved.

Data identification

The first job in extracting useful information from a mass of figures is to identify what types of data you are dealing with. There are two main types, qualitative and quantitative. Both can be further divided into two subgroups (table 2).

Table 2 Types of data

Qualitative	Quantitative
Nominal	Discrete
Ordinal	Continuous

NOMINAL DATA

Here data are classified using numbers that are arbitrarily assigned to particular categories. For example, giving nationalities particular numbers such as 1 for Iceland, 2 for England, 3 for Wales would produce nominal data. Dichotomous (or binary) data are a special type of nominal data where there are only two categories. Examples of this include marking male subjects 1 and females 2; another is dividing the groups according to the presence or absence of a particular disease.

Even though numbers are applied to these characteristics, the ordinary arithmetic rules do not apply. This is because the numbers are simply labels and do not imply any inherent order. As a result we cannot add, subtract, divide or multiply the numbers because the result would be meaningless. It would be the same as trying to add (or subtract, divide or multiply) Iceland and England.

Beware nominal data are also referred to as categorical data in some statistical texts.

ORDINAL DATA

In contrast with nominal data, ordinal data deal with categories that can be organised in some logical sequence known as "rank order". The categories themselves could be either numeric (for example, the Glasgow coma score) or non-numeric. An example of the latter is seen in the work by Esposito *et al* in which they asked surgeons to grade how much they learnt about trauma from a variety of different experiences.[2] The grading system they used was:

A great deal
A fair amount
Very little
Nothing
Not applicable

It is important to note that with ordinal data the difference between each category does not have to be equal. For example, consider the Glasgow coma score. A patient who obeys commands (M = 6) has a better motor response than one who has abnormal flexion to pain (M = 3). However, it does not mean that response is twice as good. Similarly a deterioration from M = 6 to M = 5 is not necessarily the same amount of deterioration as seen with a fall from M =2 to M = 1. The important consequence of this is that the arithmetic processes of adding, subtracting, dividing and multiplying cannot be used on ordinal data.

QUANTITATIVE-DISCRETE DATA

These data commonly have up to 20 possible whole number values with the difference between each being equal. The number of compliments your department receives per week and the weekly number of cardiac arrests represent types of quantitative-discrete data. As the numbers can be listed in rank order, these data could be considered ordinal and are often analysed as such. However, because the difference between each number is equal, ordinary arithmetic rules apply. As a result additional information can be obtained using tests that cannot be applied to ordinal data. These will be described in detail later in this series but for now consider the common problem of determining the "average" result. When dealing with quantitative-discrete data the mean value can be calculated. As this is the sum of all the values divided by the number of data points, it provides information about the whole group. In contrast, with ordinal data the median is used. This is less informative because it is simply the midpoint of the data values once they have been put into rank order.

QUANTITATIVE-CONTINUOUS DATA

When there are more than the arbitrary figure of 20 possible values, the data are considered to be quantitative-continuous. These are similar to discrete data in that the difference between consecutive numbers is equal. However, they do not need to be whole numbers, instead they can be any value within a particular range.

Indeed, as the data are continuous, each value is only restricted by the accuracy of the measuring device. Consequently quantitative-continuous data have more than 20 possible values and are invariably associated with units of measurement. Everyday examples include Sao_2, heart rate, blood pressure and weight.

Key points

A discrete variable can only have a limited number of values
A continuous variable can have an infinite number of values
Nominal, ordinal and quantitative-discrete data are all inherently discrete

It is pertinent to note at this stage that not all statistical texts classify quantitative data as shown here. A traditional way is to separate it into ratio and interval groups. The difference between these two scales is best demonstrated in the following examples:

The difference between 60°F and 40°F is the same as that between 80°F and 60°F. However 40°F is not twice as hot as 20°F.

The difference between 60 kg and 40 kg is the same as that between 80 kg and 60 kg. Furthermore, 40 kg is twice as heavy as 20 kg.

The reason for this is that Fahrenheit has an arbitrary, rather than absolute zero. It is therefore an example of an interval scale because only the difference between the points is meaningful. In contrast, weight is an example of ratio data because it includes an absolute zero. Consequently differences on the scale can be compared as well as relative magnitudes. Other examples of ratio data are temperature measured in degrees Kelvin; blood pressure and heart rate.

The type of classification of quantitative data used is left to personal preferences and what is considered most appropriate for the information being collected.

Key points

There are different types of data
The major categories are nominal, ordinal and quantitative
There are two ways of sub-classifying quantitative data. Discrete and continuous or interval and ratio.

Using this four part classification of data, Egbert identified the following types of data in the computer print out (table 3).

Table 3 Types of data identified by Egbert

Type of data	*Variable*
Nominal	Sex
	Presenting complaint
	Discharge diagnosis
Ordinal	Triage category
Quantitative — discrete	Age (y)
Quantitative — continuous	Booking time
	Time to be seen (in minutes)

Table 4 Comparison of graphical and tabular formats

Graphical	*Tabular*
Allows quick assimilation of information	Takes longer to assimilate the data
Less precise because there is a limit to the amount of information given	Allows more information to be given so can be more precise
Type of presentation depends upon the data	Can be used with any data
Good for showing multiple observations on individuals or groups	Ideal for displaying a large number of variables at once

Summarise the data for each variable

Summarising data is important because it allows the information to be easily and quickly interpreted. It can be done graphically or in a tabular format depending upon the type of presentation (table 4). Nowadays this information can be gained at a touch of a keyboard. However, the method you use needs to take into account the type of data being summarised.

Key point You need to know the type of data to summarise appropriately

To demonstrate these points, consider figure 2 (A) and (B). Both show the same data from a study assessing the effect of training to encourage glove wearing during resuscitation. With ordinal or quantitative data the rank order should be used. However, as the type of team member is a nominal classification, the order of the groups on the horizontal axis is left to personal preference. Commonly though authors list the variables as going from most to least common.

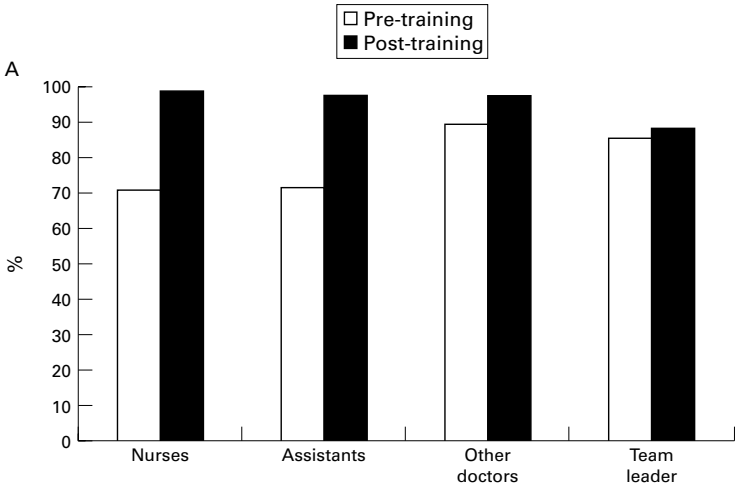
Although the graph gives an immediate picture of the results, only percentages are shown and you do not know how large a sample these are based on. This is because the sample size varied for each of the eight situations, a feature difficult to add to a graph. In contrast a table can have so much information that the key message may be lost in the data. It is therefore easier to interpret tabular results when they are accompanied by a sentence summarising the pattern shown. For example it is helpful to say "glove wearing increased after the initiative, particularly in nurses" rather than "the results are shown in the table".

TABULAR SUMMARY

These are commonly used to summarise nominal data but they can be applied to ordinal and quantitative varieties as well. The number within a particular category is called the frequency. Consequently, a frequency table lists the various numbers within different categories.

Key point A frequency table is a list showing the number of items in each of the possible categories that a variable can take

As shown in figure 2B, the first column lists the categories and the subsequent ones have the frequencies. Percentages or proportions can also be included and these allow easier comparisons between summaries obtained



B

	Before	After
Team leader	48/56 (86%)	63/71 (89%)
Other doctors	55/61 (90%)	78/80 (98%)
Nurses	79/111 (72%)	106/107 (99%)
Assistants	48/67 (72%)	91/93 (98%)
Total	230/295 (78%)	338/351 (96%)

Figure 2 Glove wearing by trauma team members before and after training. (A) Graphical presentation; (B) tabular presentation.

from different sized groups (fig 2B). Ideally a total, or average, row at the bottom should be added as it makes it easier to pick out high and low values (fig 2B). As percentages are usually rounded up to the nearest whole number when dealing with less than 100 subjects, the grand total may not be equal to 100.

Table 5 Optimal graphical representations of various data

	Nominal	*Ordinal*	*Quantitative — discrete*	*Quantitative — continuous*
Bar chart	√	√	√	×
Pie chart	√	√	√	×
Histogram	×	×	×	√
Frequency distribution	×	×	×	√

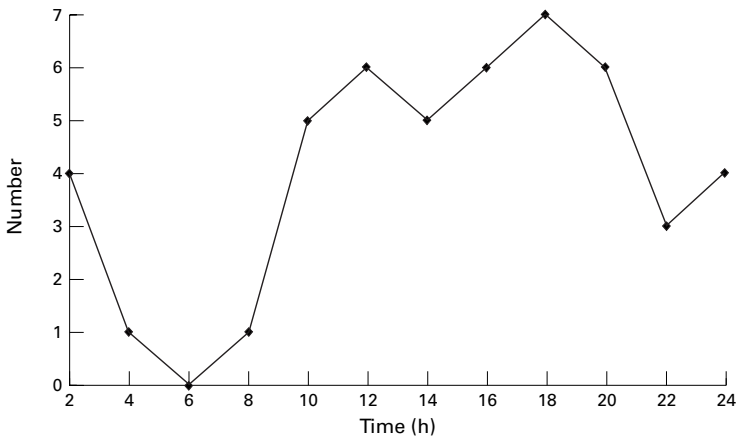


Figure 3 Change in paediatric admissions over time.

When quantitative-continuous data are used, simply listing all the cases provides little information. This is because most values only occur once making it difficult to discern any underlying pattern. The same applies when using ordinal or nominal data that have over 15 categories. To overcome this, groups are made either from combining separate categories or taking sections of a range of values. The total number (that is, frequency) of items lying within these groups is then listed. In "grouping" separate categories it is important that they do not overlap. For example, if Egbert divided the patient's age as:

0–4 y 4–8 y 8–12 y 12–16 y

he would be unsure in which group those aged 4, 8 and 12 years should lie.

Key points
When grouping categories in frequency tables ensure that:
All the categories are of equal width
There is no overlap
There are between 5 to 15 categories

When designing these tables bear in mind that it is easier to read down a column then across a row. Furthermore, avoid unnecessary precision. Extra decimal places are rarely important and make comparing figures more complex.

GRAPHICAL SUMMARIES

There are several ways of summarising your information graphically but the choice depends upon the type of data you are dealing with (table 5).

Line chart

These charts are commonly used to show the change in the data over time. When used in this way, time is recorded on the horizontal axis with the frequency on the vertical axis (fig 3). These data are sometimes referred to as longitudinal data. In contrast, cross sectional data measure the value of a variable at a fixed point in time. An example of this would be the systolic blood pressures of all patients in the emergency department at the time of the New Year.

Pie chart

This is an excellent method of presenting discrete data when you want to compare the frequency of each category with the total frequency (fig 4). The main limitation is that only one variable can be presented without the diagram getting too complex. Consequently, though the relative proportions of different variables can be shown in a series of pie charts, a bar chart is preferable.

Bar chart

These charts are used to show discrete data. The bars are of equal width and their height represents the frequency of individual cases in each category (fig 2A). As the data are discrete, the bars are separated by gaps of equal size. In cases of ordinal and quantitative-discrete data they are also arranged in the correct order

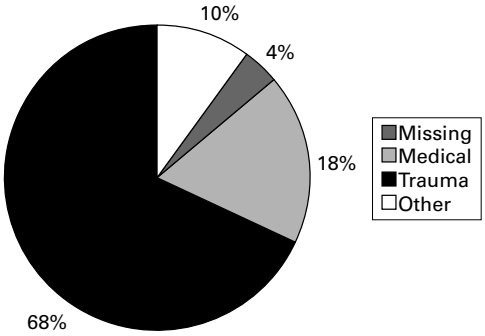


Figure 4 Type of complaint in paediatric A&E attendances.

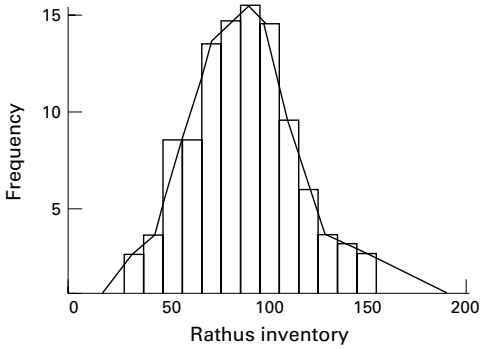


Figure 6 Assertive behaviour in 86 student nurses at the beginning of their Project 2000 training.

(fig 5). In contrast, the arrangement of categories of nominal data is arbitrary. In practice, however, they are usually positioned such that there is decreasing frequency from left to right.

Frequency histogram

This resembles a bar chart with the variable being divided into columns called intervals. These are equivalent to the categories and groups in the frequency distribution tables discussed previously. In contrast with the bar chart there are no gaps because they are continuous data (fig 6). The interval's width is dependent upon the range of values chosen but an optimum needs to be achieved between being too narrow or too wide as information can be lost. When a constant interval width is used, the height of the column represents the frequency. However, if it is decided to use columns with variable widths, the height needs to be adjusted. Consequently if the width of the column is doubled, the height needs to be reduced by a half. This ensures the area of the block remains proportional to the frequency and so a correct interpretation of the graph can be made.

Frequency distribution

The middle of the tops of the histograms intervals can be joined by straight lines to form a frequency polygon (fig 6). To close the polygon it is necessary to assume an imaginary category with zero frequency at each end of the distribution. When the histogram has many intervals the resulting frequency distribution gets smoother and tends to fit into one of the following types of curve (fig 7):

Normal distribution

This curve, which is sometimes called a "Gaussian distribution", has the characteristics of a single peak with an even distribution of values on either side (fig 7A). The vast majority of distributions in medical statistics have only one peak (that is, they are unimodal). However, some variables, such as enzyme concentrations, are not symmetrical nor do they have an obvious peak.

Statistics trivia[3]

Johann Karl Gauss (1777–1855) was born in Brunswick and an infant prodigy in mathematics. By the age of three he was helping to sort his gardener father's accounts. At 30 he was appointed Professor of Mathematics and Director of Observatory in Gottingen. While making many measurements at the observatory he proposed that observations were subject to a large number of small, independent errors and so would be "normally" distributed.

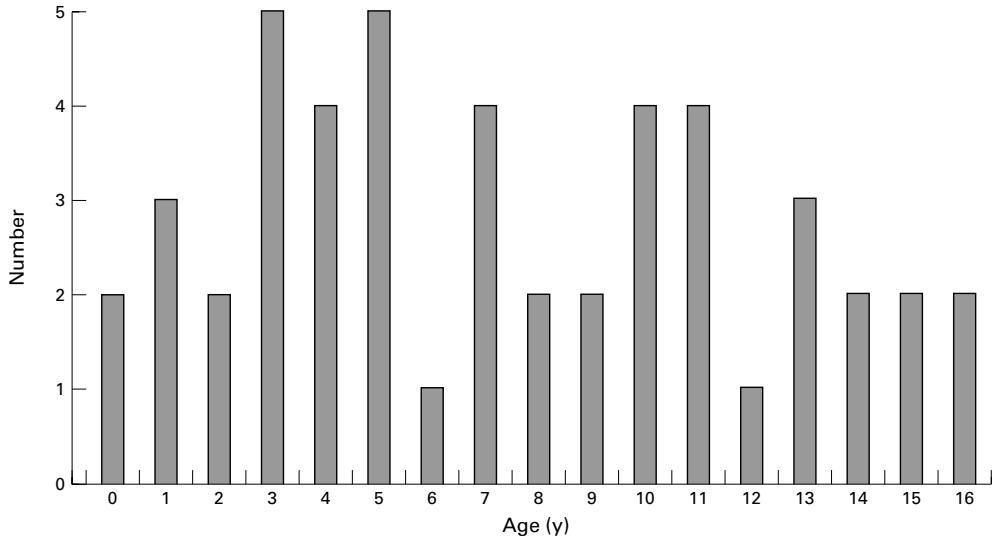


Figure 5 Age of paediatric A&E attendees.

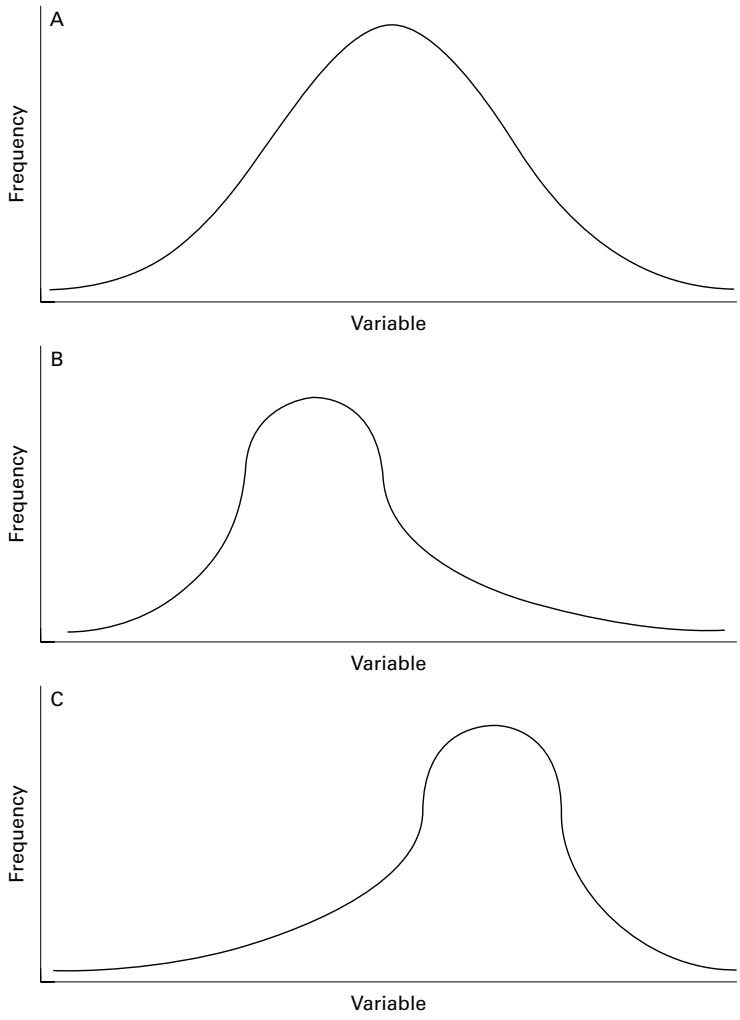


Figure 7 Various types of distribution. (A) Normal distribution; (B) right (positive) skewed distribution; (C) left (negative) skewed distribution.

Skewed distribution to the right (positive) or left (negative)

Here the peak is pushed to one side with an uneven distribution of values between both sides (fig 7 B, C). The nomenclature is based upon the location of the long tail, rather than the peak of the distribution.

Distributions that are positively skewed, but have a single peak, can sometimes be converted to a normal distribution by transforming the variable. As we will see later in this series, the rationale for this is to enable more information to be deduced from the data. A common transformation is to use a logarithmic rather than a linear scale. The effect is to compress the upper end of the range relative to the lower part. This is demonstrated in figure 8 where Egbert plotted the parent's income of the paediatric trauma victims attending Deathstar General. A frequency distribution shows there is a positively skewed distribution (fig 8A). After logarithmic transformation there is a more even spread of the data (fig 8B).

An important example where this type of manipulation is carried out is the pH scale. This is the negative logarithmic transformation of the hydrogen ion concentration. When considering logarithmic transformation it is important to bear in mind that logarithmic scales do not include zero or negative values. Consequently this type of transformation cannot be used with variables that can have such figures (for example, blood pressure and base excess).

Table 6 Data collected on 50 paediatric trauma patients attending Deathstar General

Patient	*Day of the week*	*Admission time*	*Site with the highest AIS*	*GCS*	*Respiratory rate (min)*	*Systolic blood pressure (mm Hg)*
1	Monday	12:35	Extremity	14	22	110
2	Monday	16:43	Head	3	4	70
3	Tuesday	18:54	Extremity	15	14	120
4	Thursday	22:51	Abdomen	13	20	106
5	Friday	04:35	Head	5	7	80
6	Friday	15:45	Extremity	15	30	114
7	Friday	16:17	External	15	21	111
↓	↓	↓	↓	↓	↓	↓
50	Monday	17:25	Extremity	14	18	129

Key point Logarithmic transformation cannot be carried out on variables that could have zero or negative values.

Polymodal distribution

This is when the distribution has more than one peak. These are not common but the trimodal distribution of death after trauma is one well recognised in emergency medicine.

Summary

Descriptive statistics are used to summarise numerical information so that it is in a more manageable form. There are a variety of ways of carrying this out depending upon what type of data we are dealing with. There is also a choice when presenting data. Graphical and tabular formats are possible but each have strengths and weaknesses. Selection therefore needs to take these into account along with the format of the presentation and the type of data.

Quiz

(1) Classify the following types of data:
   (a) Temperature in degrees centigrade
   (b) Heart rate
   (c) ISS
   (d) Pain score
   (e) PEFR

(2) Identify which of the following can be transformed logarithmically:
   (a) Blood pressure
   (b) Base excess
   (c) pH
   (d) Patient's weight
   (e) Triage category

(3) Identify which of the following variables are discrete and which are continuous:
   (a) Age (in 10 year blocks)
   (b) Systolic blood pressure
   (c) Sex
   (d) Plasma sodium concentration
   (e) Triage category

(4) What are the differences between a bar chart and a frequency distribution?

(5) This is one for you to try on your own.

Egbert has collected data on the last 50 paediatric trauma patients attending Deathstar General (table 6).

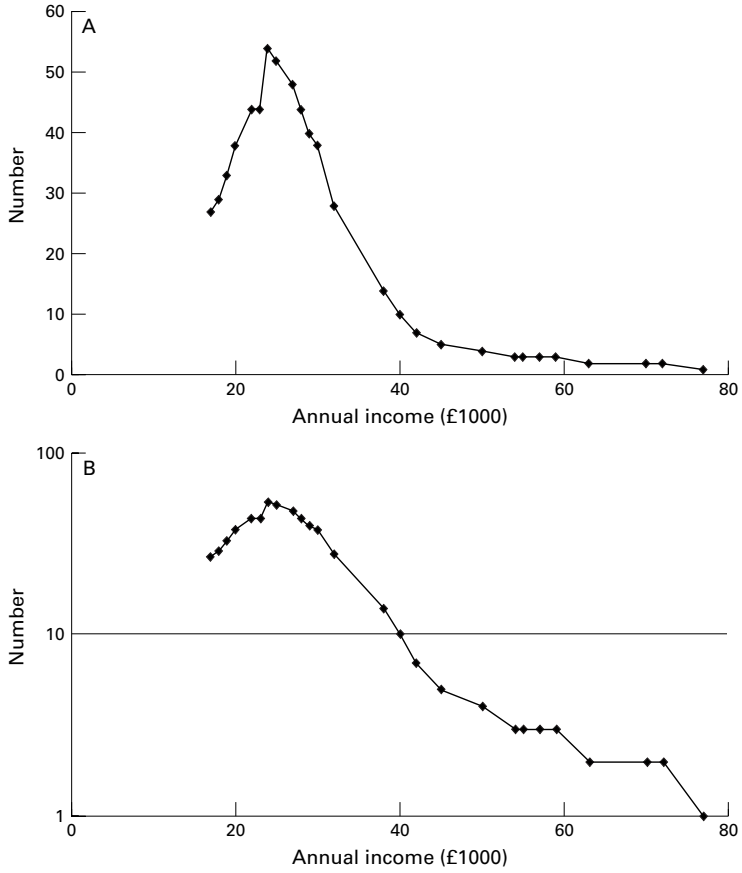


Figure 8 Income of parents of paediatric trauma victims. (A) Standard frequency distribution; (B) same data after logarithmic transformation.

(a) Describe the type of data each variable represents
(b) Identify which variables are continuous
(c) Describe the best graphical summary for each variable

Answers

1 (a) Quantitative-continuous (or interval) data
(b) Quantitative-continuous (or ratio) data
(c) The ISS represents an example of data that are difficult to categorise. ISS is usually treated in the literature as a continuous variable but it could be argued that it is not even ordinal data. After all there is no reason why a patient with an ISS of 18 from an extremity injury (AIS = 3) and a burn (AIS = 3) should be more at risk of dying than a patient with a single AIS 4 head injury (ISS = 16). However, we know from experience that there is a smooth relation between ISS and mortality. ISS therefore tends to be used as if it were a continuous variable.
(d) Ordinal data
(e) Quantitative-continuous (or ratio) data

2 (a) No (Has 0 in its scale)
(b) No (Has 0 and negative values in its scale)
(c) No (It is already a logarithmic scale)
(d) Yes
(e) No (Ordinal data)

3 (a) Discrete
(b) Continuous
(c) Discrete
(d) Continuous
(e) Discrete

4 There are gaps between the columns in the bar charts because it is used to demonstrate discrete data. In contrast, a frequency distribution does not have gaps between the columns because it is used to show continuous data.

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.
Funding: none.

Conflicts of interest: none.

1 Altman D, Bland J. Variables and parameters. *BMJ* 1999;**318**:1667.
2 Esposito T, Kuby A, Unfred C, *et al*. General surgeons and the advanced trauma life support course: is it time to refocus? *J Trauma* 1995;**39**:929–33.
3 Jones R, Payne R. Clinical investigation and statistics in laboratory medicine. London: ACB Venture Publications, 1997.

Further reading

Altman D. Types of data. In: *Practical statistics for medical research*. London: Chapman Hall, 1991:10–18.
Altman D, Bland J. Presentation of numerical data. *BMJ* 1996;**312**:572.
Bowlers D. Types of variables. *Statistics from scratch*. Chichester: John Wiley, 1996:14–31.
Bowlers D. Organising qualitative data. *Statistics from scratch*. Chichester: John Wiley,1996:37–64.
Coogan D. *Statistics in clinical practice*. London: BMJ Publications, 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.