Descriptive statistics: Simply telling a story

***Laura Delaney** introduces the principles of descriptive statistical analysis and presents an overview of the various ways in which data can be presented by researchers.*

Undertaking statistical analysis of data can be a daunting task. Many people may have had bad experiences with mathematics in school and so dread having to deal with numbers again in their lives. Some may have heard the adage of 'lies, damned lies and statistics' and so wonder what is the point of using statistics. The following article has been written with gentle statistical hands to ease worries and concerns. It is suggested that anxious readers do remember to breathe during the reading of this article.

The following is a comprehensive overview of undertaking descriptive statistical analysis. Descriptive statistical analysis is the simplest statistical analysis available. The purpose of this type of analysis is to simply describe the sample group from which the data was collected.

The more advanced analysis is called inferential statistical analysis, which includes carrying out descriptive analysis but then drawing conclusions about the larger population from which the sample was taken. It is only descriptive statistical analysis which will be discussed here.

The purpose of undertaking descriptive analysis is to closely examine the collected data in order to describe its salient features. The process involved allows for gaining a sense of order in the data so 'the story' of the sample group can be told.

Laura Delaney is currently teaching statistical analysis at Holmesglen Institute of Tertiary and Further Education and Swinburne University of Technology, Melbourne Australia.
Email: laurad2008@live.com.au

KEY WORDS
- statistical analysis
- level of measurement
- histogram

Accepted for publication 19 April 2009

What follows is an overview of the basic tools that are available to researcher to tell that story. These tools include graphs (or the pictures in the story, if you like) and data summarizing techniques. Instead of the term 'data summarizing techniques' the word statistics could sometimes be used because that is primarily the purpose of a statistic—to summarize data. A statistic is simply a number that is used to summarize data. Not that scary after all.

Level of measurement

Before undertaking any quantitative research it is crucial to examine the nature of the questions being asked of the sample group. In particular the level of measurement of each question needs to be determined: that is, the type of data that will be collected with each question. There are four levels of measurement: nominal, ordinal, interval and ratio levels.

Nominal level

Numbers or letters are used to identify categories. For example with a question on gender '1' may be used to represent female and '2' to represent male or alternatively 'f' for female and 'm' for male.

Ordinal level

Numbers are used to rank according to a particular attribute or level of importance. This could be asking the sample group to rank in order of importance some aspects of their hospital stay, such as the food, the friendliness of the staff and the information received during their stay from 1–3 where '1' is their most important aspect and '3' their least important.

The ordinal level does not have equal distances between the rankings. That is, we cannot say that the distance between the 1st and the 2nd level of importance is the same distance as between the 2nd and the 3rd level of importance. In other words there is no numerical unit of measurement at this level.

There are various rating scales that are commonly used. A common scale is the Likert Scale which can be used to measure attitude. For example the respondent is given a serious of statements and asked to indicate their level of agreement to the statements on a response scale including Strongly Agree/Agree/Undecided/Disagree/Strongly Disagree. There is plenty of literature on the 'do's and 'don't's of constructing such scales.

Interval level

This can be more difficult to understand. The interval level is similar to the ordinal level in that there is a ranking involved but unlike the ordinal level there is equal distance between the points. For example asking the sample group to indicate on a scale of 1–5 how they feel about the quality of care they received during their hospital stay is based on an interval scale (*Table 1*). The distance between 1 and 2 is equal to the distance between 2 and 3 and so on.

Table 1. Interval level example

The quality of care was	1	2	3	4	5
	Poor				Excellent

LEARNING AIMS

- To understand the aim of descriptive statistical analysis
- To be aware of the different ways in which data may be presented for analysis
- To understand the features to look for when analysing data in graphs

Zero at this level of measurement does not mean that the attribute or quality does not exist. Temperature is an interval level of measurement in that when the temperature is 0°C it does not mean there is no temperature, but rather it is very cold.

The interval scale has great relevance in the field of psychology as most psychological constructs are measured using standardized tests based on this level of measurement. Zero values and negative values can be valid at this level of measurement.

As in the ordinal level, the Likert Scale can also be used at an interval level to undertake attitude measurement. It must be noted however that in order for a researcher to use a Likert Scale at this level of measurement there must be justification that there is equal distance between the points of the scale. Once again there is plenty of literature available on Likert Scales and using them at the two different levels of measurement.

The obvious advantage of using a Likert Scale at the interval level instead of the ordinal level is the associated data summary techniques and graphs that can be used with the former (*Table 2*). The interval level data summary techniques and graphs can make the description of the data more interesting, that is, the story is more interesting.

Ratio level

Absolute (or exact) numbers are used to measure a characteristic. For example the individual ages of the sample group in absolute terms (i.e. 23 years, 47 years) or individual weights in kilograms. Unlike the interval level, zero or negative values are not valid at this level of measurement.

The level of measurement for each question asked of the sample group will determine what data summary techniques and graphs will be appropriate for the data

Table 2.
Data summarizing techniques and graphs for each level of measurement for a single question

Level	Data summarizing techniques	Graph
Nominal	Frequency table	Bar chart Pie chart
Ordinal	Frequency table	Bar chart Pie chart
Interval	Measures of central tendency Measures of dispersion	Histogram Boxplot Stem and leaf plot
Ratio	Measures of central tendency Measures of dispersion	Histogram Boxplot Stem and leaf plot

Frequency tables

Frequency tables can be easily constructed in the Statistical Package for Social Sciences (SPSS) (*Table 3*). Reading the table from the left the first two column show the four categories of diagnosis and the actual number (or frequency) of patients within each category. There are a total of 230 patients in the sample. The 13 people coded as missing represents those in the sample for which there was no diagnosis available at the time of interview. When there are missing data researchers tend to use the valid percentages in the discussion of the findings because they exclude the missing data. The cumulative frequency is useful for determining the percentage less than a particular category or between two categories.

Table 3.
Frequency table

Patient diagnosis		Frequency	Percent	Valid percent	Cumulative percent
Valid	Anorexia nervosa	97	42	45	45
	Anorexia with bulimia nervosa	36	16	17	61
	Bulimia nervosa after anorexia	56	24	26	87
	Atypical eating eisorder	28	12	13	100%
	Total	217	94	100%	
Missing	99	13	6		
Total	230	100%			

Bar chart

A bar chart allows for a comparison of the categories in a question (*Figure 1*). Obviously if there are many categories in a question the overall picture may be lost in the details of too many categories. A bar chart can also be displayed horizontally. From this bar chart we can see that 10% of the sample group are currently smoking and not trying to quit compared with 40% who are currently smoking and trying to quit.

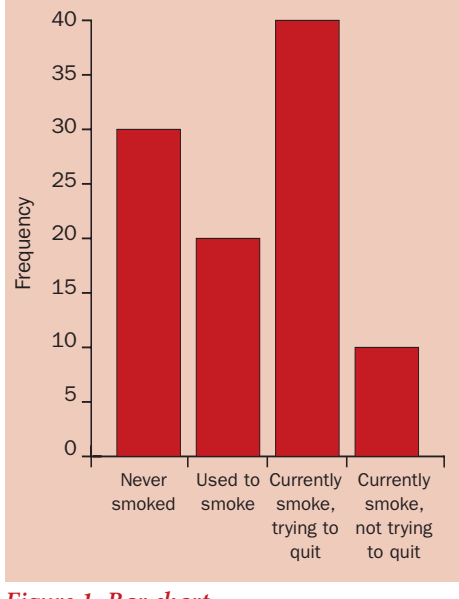

Figure 1. Bar chart

Pie charts

Pie charts are a common way of showing the categories of a question as proportional segments (*Figure 2*). It is important to find a balance between having too many segments in the pie and too few segments. A rule of thumb is to have between three and ten segments. Any less than three and the chart can be uninteresting, for example, showing gender as 50:50 in a pie chart. More than ten and the chart can become too busy to get the overall picture across.

In this pie chart we can see that finance and health issues were the main problems in the last 12 months for the sample group.

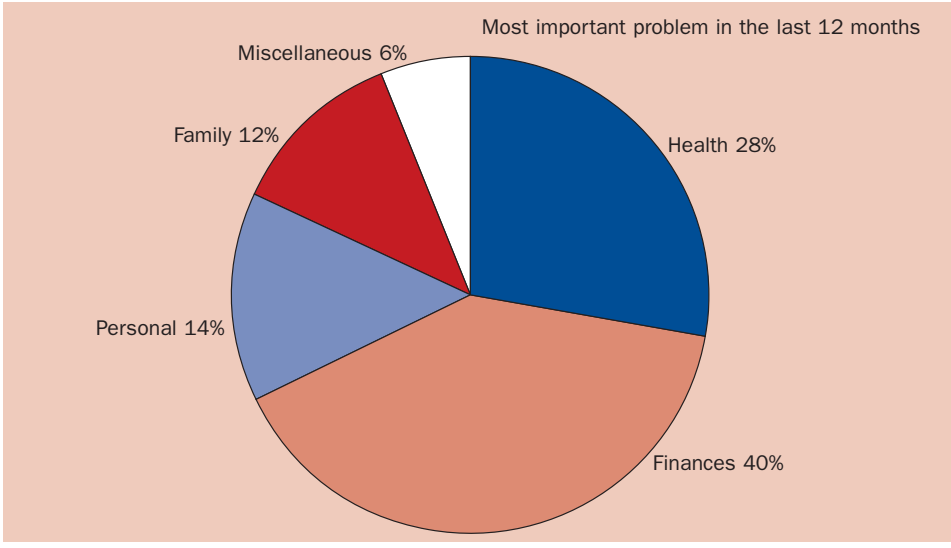

Figure 2. Pie chart

Histograms

A histogram is the most common graph to get an overall picture of ratio and interval data. To use a histogram to summarize data it is important to closely examine some of the features of the graph. To aid researchers in how to examine these features it can be useful to think of the features as 'SOCS': shape, outliers, centre and spread.

Shape

What shape does the histogram have? Or in other words how is the data distributed?

Is there a peak in the data? Where is the peak (or peaks)? Is the peak in the centre or to one or other side of the distribution? If there is a single peak with an even trailing off either side this shape or distribution is called symmetric or a normal distribution (*Figure 3*). In a symmetric histogram, values to the right of the peak are greater than the value of the peak and values to the left of the peak are less than the value of the peak.

The peak in a histogram is known as the mode of the distribution, that is, the value in the data which occurs the most frequently. If a histogram has two peaks of similar height we say the distribution is bimodal (*Figure 4*).

When a histogram is constructed using the SPSS package there is by default a box to the right of the graph which includes three summary values to aid the researcher in telling the story of the sample group. The three values are the mean, the standard deviation and *n*, which indicate a general measure for the data, a measure of difference or spread in the data and the size of the sample group, respectively. The mean and the standard deviation will be dis-

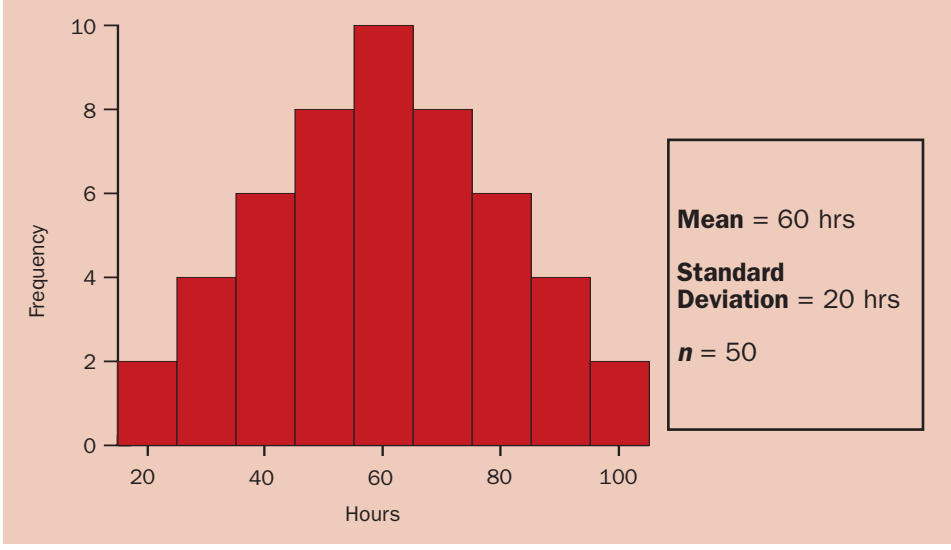

Figure 3. Symmetric histogram

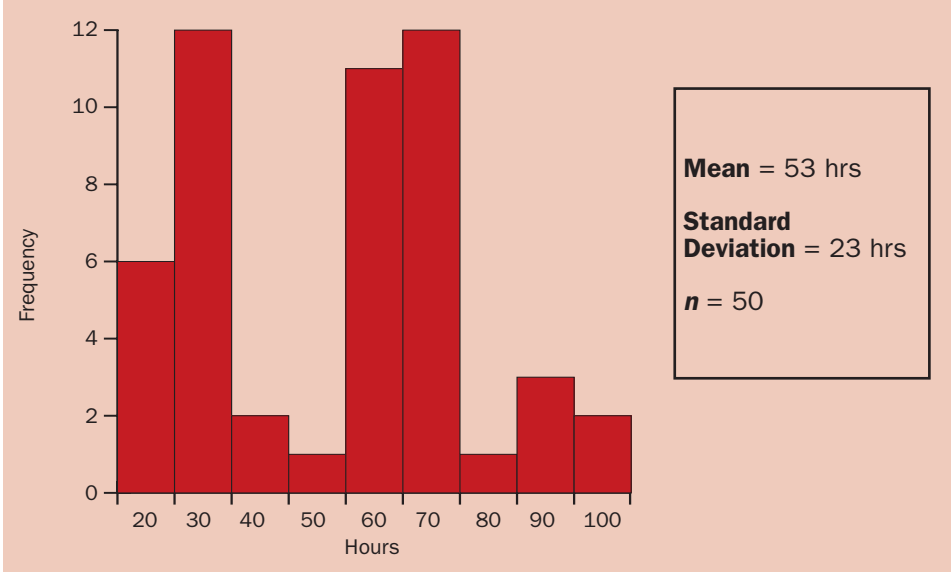

Figure 4. Bimodal histogram—there two mode values or two peaks at 30 and 70 hours, both with a frequency of 12

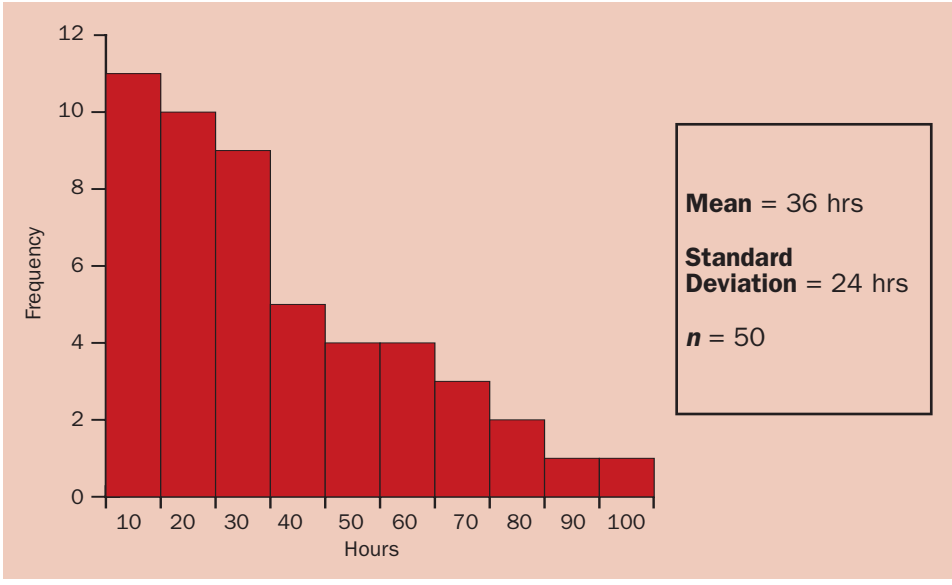

Figure 5. Positively-skewed histogram

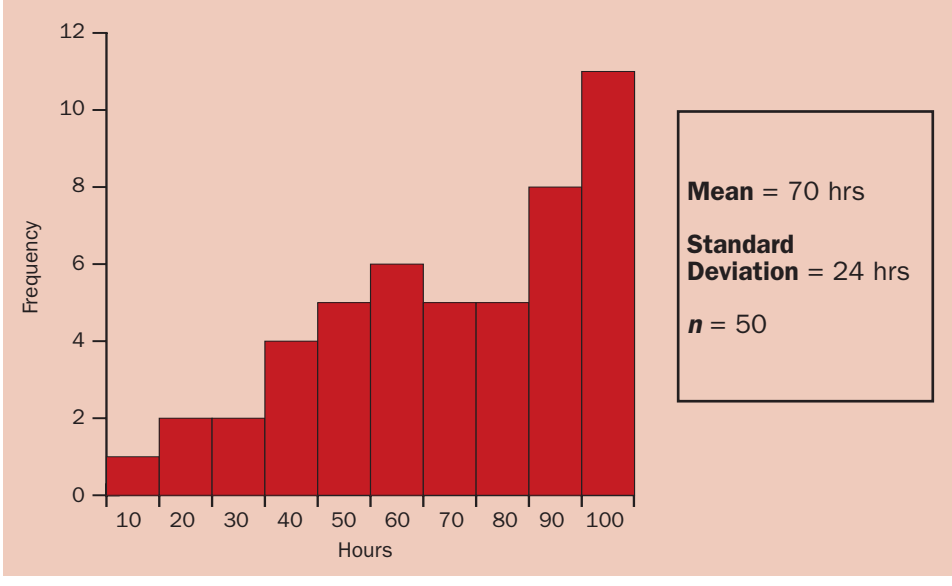

Figure 6. Negatively-skewed histogram

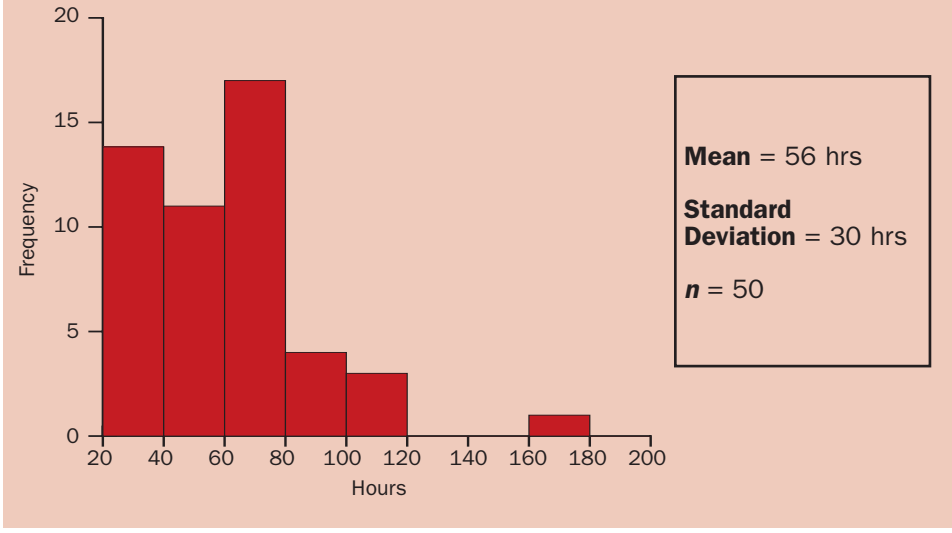

Figure 7. Histogram with outlier. In this histogram we can see that there is an outlier value between 160 and 180 hours

cussed in more detail later. The size of the sample or number of people in the sample is self-explanatory.

Sometimes a histogram may have a distinct 'tailing off' shape. If a histogram tails off to the right, we say that the distribution is positively skewed (*Figure 5*). It is positive because the tail incorporates the higher values of the data.

If the histogram tails off to the left we say the distribution is negatively skewed. It is negative because the tail incorporates the lower values of the data (*Figure 6*).

In an attempt to help consolidate this concept of skewed data, it can be useful to remember that 'the direction of the skew is where there are few', rather poetic.

Outliers

Outliers are data values that stand out from the main body of the data—data values that are atypically high or low (*Figure 7*). Outliers will usually need further examination to ascertain whether they represent an error in the data or if indeed they are legitimately different from the main body of data.

Centre

When discussing 'the centre' of a histogram we are interested in the middle value. The middle value may correspond with the highest peak of the graph, but as seen in skewed data the middle value does not always correspond with the highest peak.

With ratio and interval data the concept of centre also involves calculating a value to get an overall picture of the data. The most common value that gives this sense of the data is the mean value, also known as the average value.

The mean, the median and the mode are the three measures of central tendency. They are used to produce an overall picture of the data.

As stated earlier when a histogram is constructed using the SPSS package there is a box to the right of the graph which gives a couple of data summary values or 'statistics'. One of these statistics is the mean value of the data.

The mean is calculated by adding up all the individual values and then dividing by the number of values. Since each individual value in the data is factored into the calculation the mean value will be 'pulled' towards any outliers that might exist.

If outliers exist in the data it is advisable to calculate the median value instead of the

mean to get a general sense of the data. The median is the data value that represents the 50% point in the data. Most numeric/statistical software packages locate the median value easily.

The mode is simply the value in the data set which occurs the most frequently or the value with the highest peak in a histogram.

When a distribution is symmetrical the value of the mean, median and mode are all equal.

Spread

By spread we mean how varied or how different the data values are.

In a histogram we note if the values in the distribution are tightly clustered indicating a narrow area of spread (*Figure 8*), or if they are distributed over a larger area indicating a broader area of spread (*Figure 9*).

With ratio and interval data the concept of spread also involves calculating a value to get some idea of the level of difference in the data. These calculations are also known as measures of dispersion (dispersion being another word for spread, variation etc)

A simple measure of spread is the range. The range is difference between the highest and lowest value in the data.

A common measure of spread is the standard deviation. The standard deviation sounds scarier than it is. To get the standard deviation the mean value must first be calculated. Then each individual data value is compared with the mean value of the data. The amount of difference between each value and the mean values is noted and then a mean of the differences is calculated. The higher the standard deviation value, the greater the degree of difference between the values and the mean value. This point is clearly illustrated by comparing the standard deviation values of the narrowly-spread and broadly-spread histograms.

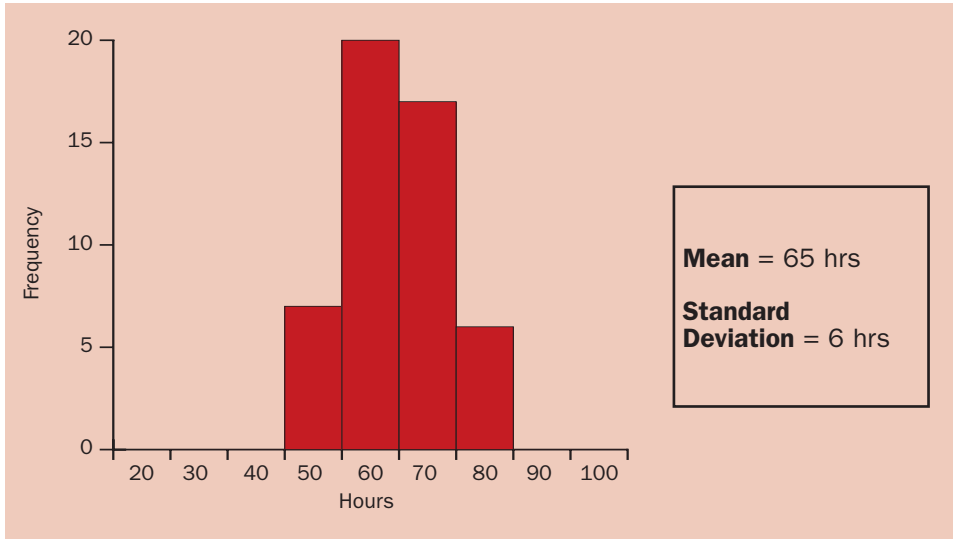

Figure 8. Narrowly-spread histogram (narrow clustering between 50 and 80 hours)

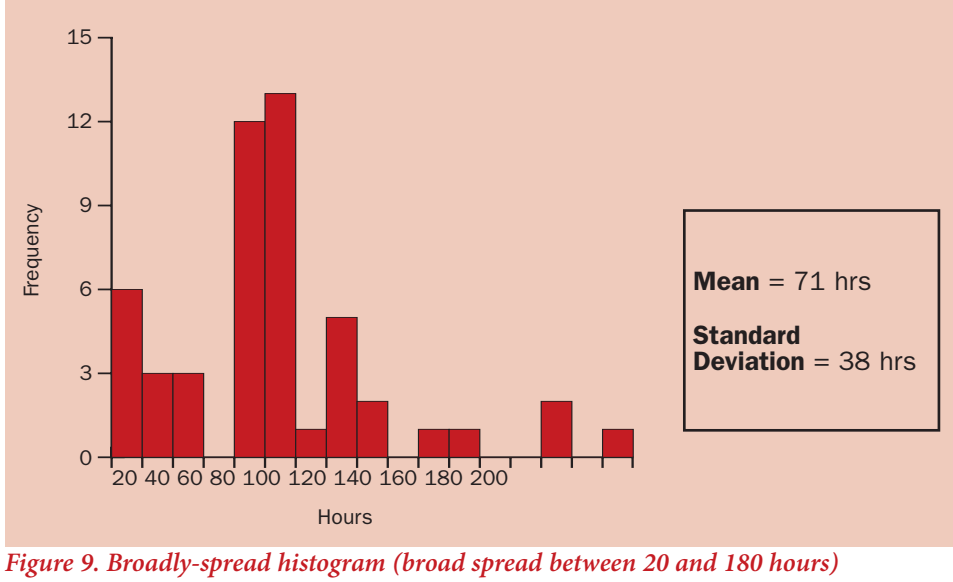

Figure 9. Broadly-spread histogram (broad spread between 20 and 180 hours)

Stem and leaf plots

Stem and leaf plots can be used as an alternative to a histogram. They can be useful in displaying small sets of data (up to 50 data values) and have the advantage of showing all original data values (*Figure 10*). The stem of the plot represents the leading digit for the data values and the leaf represents the second digit of the value. In the example, the first stem has '1' as the leading digit and a leaf of '0', which represents the data value of 10, followed by another 10, 11, 11, 12, 13 and so forth. The largest data value is 87. As with a histogram the shape of the distribution and measures of central tendency and dispersion should be discussed.

Descriptive techniques for level of measurement combinations

Often it is appropriate to examine data from two questions at the same time. For example a researcher might be interested in describing the differences between males and females and how they felt about their hospital stay or examining the relationship between age and weight of a sample of group. *Table 4* outlines some of the techniques available for some level of measurement combinations.

Boxplot

Boxplots can be easier to read if they are constructed with the ratio data shown on the horizontal axis and the nominal data shown on the vertical axis. The five points of interest in a boxplot have been noted in *Figure 11* to aid its use as a descriptive tool. The five points include the minimum and maximum values, the first quarter point (also called the quartile) in the data or the

Hours – Stem width: 10 hours

```
1 . 001123489999
2 . 0001233347
3 . 0034588
4 . 0014567
5 . 01
6 . 00005
7 . 0000
8 . 07
9 . 0
10. 0
```

Figure 10. Stem and leaf plot

Table 4.
Data summarizing techniques and graphs for some level of measurement combinations

Levels	Data summarizing techniques	Graph
Nominal and ratio	Median Interquartile range	Boxplot
Nominal and nominal/ ordinal/interval	Crosstabulation	
Ratio and ratio	Correlation	Scatterplot

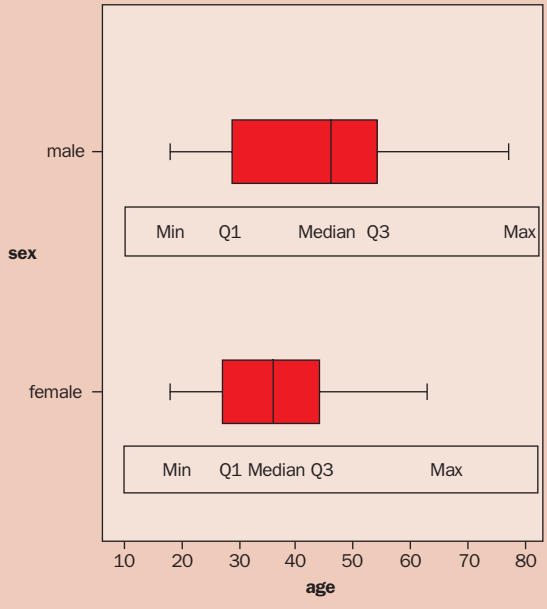

Figure 11. Box plot

Table 5.
Crosstabulation of the variable 'sex' by 'satisfied with information received during hospital stay'

Satisfied with the information received during hospital stay	Respondents sex		
	Females	Males	Total
Yes	412 65%	547 62%	959 64%
No	220 35%	329 38%	549 36%
Total	632 100%	876 100%	1508 100%

25% point in the data, followed by the median point or the 50% point of the data and then the third quarter point or 75% point in the data.

From this sex-by-age boxplot the females have less range in their ages with no female in the sample older than mid-60s. Fifty percent (from the median point) of the males are older than 45 years.

Sometimes it is appropriate to describe the data in the middle fifty percent of the distribution as the majority of a sample group often lies within this area. The middle fifty percent is the red boxed area in the plot. The range of this area can be calculated by noting the value at the Q3 point (75%) and the Q1 point (25%) and subtracting the two values. This difference between Q3 and Q1 is known as the interquartile range.

Crosstabulations

A two-way crosstabulation is simply two frequency tables merged into one table. *Table 5* shows the crosstabulation of data on the sex of respondents and whether the respondents were satisfied with the information received during their hospital stay. The table shows that 65% of the females and 62% of the males in the sample were satisfied.

Crosstabulations are easier to read if the independent data (the influencing data) is placed in the column of the table and the dependent data (the effected data) is placed in the row. The independent data tends to be more important in the relationship between the two sets of data.

Scatterplot

A scatterplot enables the researcher to determine whether a relationship exists between two sets of ratio data. *Figure 12* shows a scatterplot for the pulse rate of a sample group before undertaking an activity and the pulse rate after the activity.

In a scatterplot the independent data should be represented on the horizontal axis (X axis) and the dependent data should be represented on the vertical axis (Y axis). It is the shape and lean of the scatterplot that is described. If there is a line shape (also known as a linear shape) to the plots it is noted that there is a linear relationship between the pulse rate prior and the pulse rate after the activity.

When the scatters have a right-hand lean it is said that the relationship between

pulse prior and pulse post is positive. A positive relationship in the data means low values of the X axis data tend to correspond with low values of the Y axis data and high values of the X axis data tend to correspond with high values of the Y axis data. In the pulse rate data a positive linear relationship exists, which means low values in pulse before tend to correspond with low values in the pulse rate after the activity and high pulse before tend to correspond with high values in the pulse rate after the activity.

Alternatively, if there is a negative relationship the lean of the line is in the opposite direction to the positive lean. That is, there is a left-hand lean with low values of the X axis data corresponding with high values of the Y axis data and high values of the X axis data tending to correspond with low values of the Y axis (*Figure 13*).

If a linear relationship exists in the scatterplot the researcher can describe the relationship further by examining the strength of the relationship between the two sets of data.

Examining the strength is done by calculating the correlation coefficient, also known a Pearson's r (just a fancy way of saying a number to determine the strength of the relationship). The formula for the correlation coefficient (*r*) is a bit scary looking but fortunately statistical calculators and software packages such as Microsoft Excel and SPSS do it with the click of a mouse.

Table 6 shows that an *r* value will range from -1 to +1. If the scatterplot has a positive lean the value for *r* should be positive and alternatively if the scatterplot has a negative lean the value for *r* should be negative.

There are different cut-off points to determine the strength of the relationship ranging from weak, moderate and strong. The closer the scatterplots form a linear shape, the stronger the relationship.

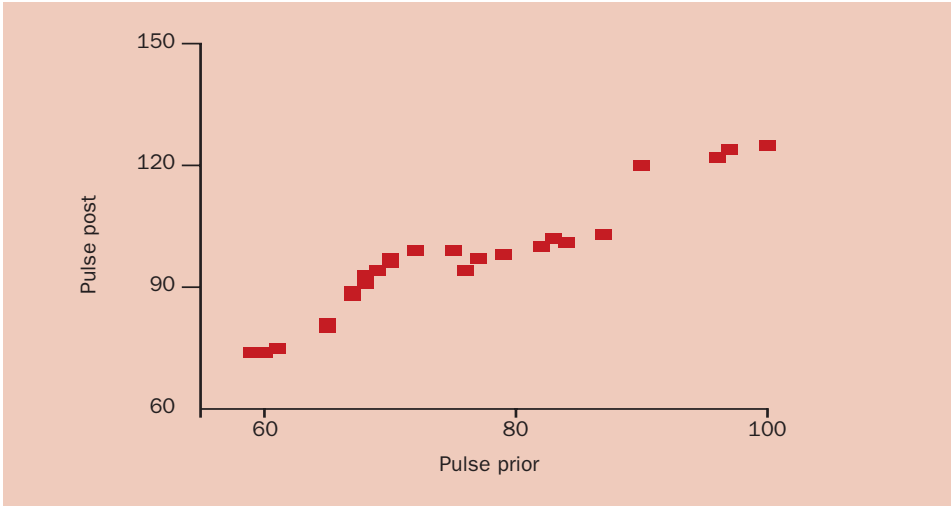

Figure 12. Positive linear scatterplot

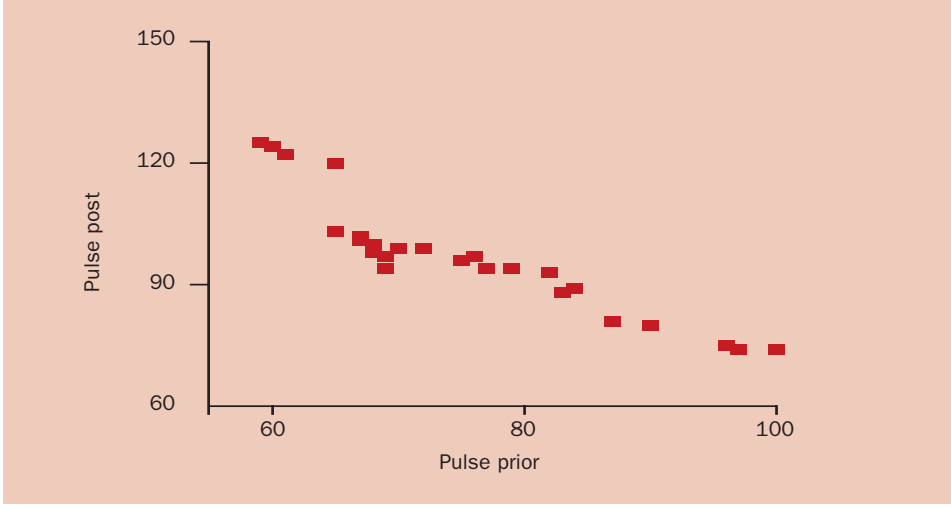

Figure 13. Negative linear scatterplot

Conclusions

It is important to remember that the above graphs and data summary techniques are simply tools for the researcher to tell the story about the sample group. The more understanding and insight the researcher has of these tools, the more interesting and accurate the story will be. BJCARDN

Coming soon … inferential statistical analysis

Table 6.
Correlation coefficient values

Negative relationship											Positive relationship									
-1	-0.9	-0.8	-0.7	-0.6	-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	+1
	Strong			Moderate			Weak					Weak			Moderate			Strong		

Copyright of British Journal of Cardiac Nursing is the property of Mark Allen Publishing Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.