Descriptive statistics (Part 1)

A Sheikh and A Cook

Descriptive statistics are widely used by researchers to summarise results in a concise yet intelligible manner. Descriptive, or summary, statistics should provide sufficient information to allow distributions of important variables to be visualised, thus allowing a clear mental picture of the group studied. Researchers must select summary statistics with care, taking into consideration both the type and distribution of a variable. Good interpretation and use of descriptive statistics requires little statistical knowledge.

TYPES OF VARIABLE

The first step in summarising data should always be to determine the variable type. The two main types of variable are numerical and categorical (Table 1). Numerical variables are either continuous or discrete; continuous data take values somewhere on a scale, such as measurements of peak expiratory flow, height or weight. Discrete data are usually limited to the positive integers, for example, the number of asthma exacerbations may be 0, 1, 2, *etc*.

Categorical variables comprise two or more categories into one of which each study participant can be placed, for example, asthmatic or non-asthmatic, male or female, or the type of inhaler device used. Ordinal variables are a subset of categorical variables, where the categories possess a clear natural ordering such as non-smoker / light smoker / heavy smoker.

The distinction between the variable types is not always clear as numerical variables may sometimes be transformed into categorical variables. For example, data on the numbers of cigarettes smoked could be treated as numeric, with typical values between 0 and 100 per day. Alternatively, categories of light and heavy smoking could be defined, and each individual then categorised as non-smoker / light-smoker / heavy smoker.

SUMMARISING DATA

Numerical variables

Data are summarised by an average value, together with a measure of how spread out observations are around this value. The best-known averages are the mean and the median, while the most widely used measures of dispersion are the standard deviation and the interquartile range. The choice of which average to use, and which measure of dispersion, depends on whether the data are Normally distributed. Figure 1 is an example of data that are Normally distributed; the symmetrical bell-shaped curve is characteristic.

Whether data are Normally distributed can often be ascertained using histograms as in Figure 1. However, with small samples a Normal plot or a formal test of Normality may be more conclusive. The best known test of Normality is that of Shapiro–Wilk, which is available with most statistical packages. Determining whether the data in a small sample are Normally distributed can be difficult, the shape of the distribution may not be apparent from a histogram and the Shapiro–Wilk test may have insufficient power to detect a departure from Normality. In such cases treating data as non-Normal is the safest approach.

For Normally distributed data, the mean and standard deviation are the usual parameters since they use all of the data and have other useful properties (Table 2). They become misleading, however, for

Table 1 Types of variable

Numerical	**Continuous**	Observations may take any value Usually generated from measurements
	Discrete	Observations limited to certain values Usually counts of an event occurring
Categorical	Each subject placed into one of the categories Usually characteristic attributes of study subjects	
	Ordinal	Categories possess a clear natural ordering

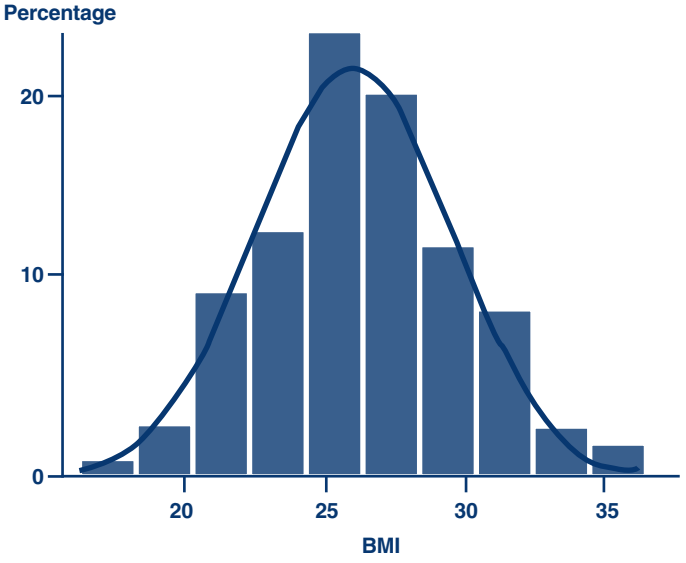

Figure 1 Body mass index (BMI) of British men

Men over 65 years of age ($n = 553$)

Aziz Sheikh
Clinical Research Fellow

Adrian Cook
Statistician

Department of General Practice and Primary Health Care, Imperial College School of Medicine, Norfolk Place, London W2 1PG, UK

Correspondence to:
Dr A Sheikh
aziz.sheikh@ic.ac.uk

Date received: 07/0799
Date accepted: 23/07/99

Asthma Gen Pract
1999;7(3):32–4

non-Normal data which are either skewed or contain a small number of highly unusual observations, known as 'outliers'. For such data the median is a more appropriate average and the interquartile range a better indicator of dispersion.

Figure 2 shows the distribution of daily air pollutant levels (PM_{10}) over three years in Santiago, Chile. Concentrations of PM_{10} were 50–100 $\mu g/m^3$ on almost half of the days, while levels on over 90% of days exceeded the European target of 50 $\mu g/m^3$. The distribution is positively skewed, with very high concentrations noted on occasions.

Table 2 Key statistical definitions

Normal distribution	Symmetric bell-shaped distribution	
	95% of observations expected to lie within 2 standard deviations of the mean	
	99% of observations expected to lie within 3 standard deviations of the mean	
Averages	Mean	Sum of values divided by number of observations
	Median	Middle ranking observation
Indicators of dispersion	Standard deviation	Average deviation of an observation from the mean
	Absolute range	Lowest to highest observation
	Interdecile range	Lowest to highest, excluding top and bottom 10% (deciles)
	Interquartile range	Lowest to highest, excluding top and bottom 25% (quartiles)
Confidence interval	Range believed to contain the population mean, derived from the sample mean and standard deviation	

The data show why the mean is a poor indicator of location for skewed data. The mean concentration is 117 $\mu g/m^3$, while the median is 99.8 $\mu g/m^3$, 15% lower. The difference occurs as a result of the mean being pulled upwards by the high levels in the right tail of the distribution. The magnitude of the effect depends on the degree of skewness, highly skewed data producing the greatest discrepancies. Outlying observations have a similar effect, pulling the value of the mean toward them.

When the mean is a poor indicator of location, the standard deviation should not be used as it is a measure of variation around the mean. Better indicators of dispersion are the interquartile and interdecile ranges. These give a measure of variability as well as some indication of the level of skewness. The absolute range is also suitable for skewed data, but should not be used for data containing outliers. It is unusual, although on occasion informative, to present more than one range.

Returning to Figure 2, the data have an absolute range of 19–380 $\mu g/m^3$, an interdecile range of 59–197 $\mu g/m^3$, and an interquartile range of 74–152 $\mu g/m^3$. From the median and the interdecile range, it is possible to visualise the data as skewed in their distribution with a peak at 100 $\mu g/m^3$ and 80% of observations between 60 and 200 $\mu g/m^3$. The interdecile range is, therefore, perhaps the most informative measure of dispersion in this instance.

Categorical variables

Data are usually presented as the numbers of subjects, together with percentages, in each category. If there are many categories, with some containing few subjects, merging categories may simplify presentation. Consider a symptom variable with four ordered categories: none, mild, moderate or severe. It may happen that most subjects fall into one of the two extreme categories. With such data the main features are perhaps best communicated to the reader by merging the middle two groups into a new category of 'some symptoms', and presenting numbers and percentages for the three groups.

WHEN NOT TO SUMMARISE

Data cannot always be usefully summarised. For a variable with more than one peak in the distribution both the

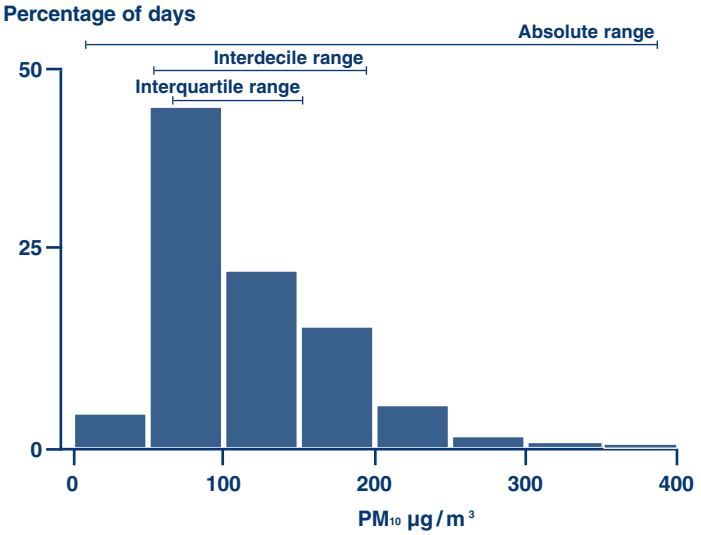

Figure 2 Daily particulate (PM_{10}) levels
Santiago, Chile (1992–4)

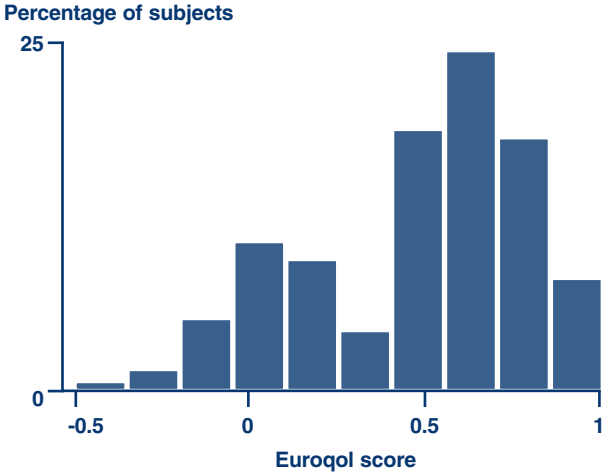

Figure 3 Euroqol quality of life scores ($n = 1811$)

Key messages

- Descriptive statistics allow important information about data to be conveyed in a concise manner
- Categorical variables can be summarised using counts and percentages
- Normally distributed variables should be summarised with the mean and standard deviation
- Non-Normal variables should usually be summarised with the median and a measure of range
- Continuous variables should always be plotted to ensure summary statistics are suitable

mean and median can be grossly misleading. Such situations are rare, but again show the importance of plotting variables.

Figure 3 shows the scores of subjects given a quality of life questionnaire (Euroqol). The distribution is bimodal, having two distinct peaks. The mean and median are 0.48 and 0.53 respectively. Both values lie in the group of scores to the right, and both fail to describe the observed data. These data can only be described in detail by showing the figure.

An alternative approach, and one normally used with data from this particular questionnaire, is to categorise subjects into 'cases' and 'non-cases' – 'cases' being the group to the left of the divide with the lower quality of life scores. Techniques appropriate to categorical data are then used.

POINTERS TO POOR SUMMARY STATISTICS

The inappropriate use of the mean is a common mistake. Suspicions of skewed data should be aroused if the standard deviation is greater than the mean for a variable limited to positive values, such as age or exposure (for example, pollen count). In such cases, the median and a measure of range would provide a more accurate summary.

It is also worrying to see a variable summarised by the mean and range. If the data are non-Normally distributed the mean will usually lie some distance from the centre of the range and the median would be more appropriate. If the mean is appropriate then the standard deviation is the best measure of dispersion, remembering that the range can still be estimated by adding and subtracting three standard deviations to the mean.

CONCLUSION

In this article, we have shown the importance of appropriately and adequately summarising data. Researchers need to thoroughly investigate data before attempting a data summary, aiming to provide sufficient information for the distribution of the data to be visualised. Used well, indicators of location and dispersion convey almost as much information as a figure. Used badly they serve only to mislead and further the belief that there are 'lies, damned lies, and statistics'. ■

Acknowledgments

We would like to thank Professor Brian Hurwitz, Dr Mark Levy, Dr John Salinsky, Dr Sangeeta Dhami, Bernadette Alves and Claire Cook for their helpful comments on this manuscript.

Recommended reading

Altman D. *Practical statistics for medical research*. London: Chapman & Hall, 1997:19–38

Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1997:46–72

Campbell MJ, Machin D. *Medical statistics: A common sense approach*. New York: John Wiley & Sons, 1990:40–53

Gardner M, Altman D. *Statistics with confidence*. London: BMJ Publishing Group, 1993:93–4

The next article in this series will discuss methods for summarising observed differences between two groups of subjects.