Statistics II: Central tendency and spread of data

Anthony McCluskey BSc MB ChB FRCA
Abdul Ghaaliq Lalkhen MB ChB FRCA

Central tendency

Examining the raw data is an essential first step before proceeding to statistical analysis. Thereafter, two key sample statistics that may be calculated from a dataset are a measure of the central tendency of the sample distribution and of the spread of the data about this central tendency. Inferential statistical analysis is dependent on a knowledge of these descriptive statistics. In the first article of this series, types of data and correlations were discussed.[1]

Different measures of central tendency attempt to determine what might variously be termed the typical, normal, expected or average value of a dataset. Three of them are in general use for most types of data: the mode, median, and mean.

The mode

Literally, the mode is strictly a measure of the most popular (frequent) value in a dataset and is often not a particularly good indicator of central tendency. Despite its limitations, the mode is the only means of measuring central tendency in a dataset containing nominal categorical values. For example, in a survey of 10 senior house officers (SHOs) asked which form of continual professional development (CPD) activity they preferred in preparation for a forthcoming examination, the following responses were obtained: viva practice 5, tutorial 3, in-theatre teaching 2, lecture 0. The mode of this dataset is viva practice as it is the largest (most popular) category. We might say that a 'typical' SHO prefers viva practice and plan the CPD time accordingly.

The mode may also be used for ordinal categorical data and for interval data, although the median or mean are more useful in these circumstances. For example, suppose a pilot study is undertaken to determine the severity of pain on injection of propofol in 10 patients and an ordinal verbal pain score system between 0–3 is used. The pain scores observed are: 0, 1, 1, 2, 2, 2, 2, 3, 3, 3. The mode of this dataset is a pain score of 2.

The median

The median is defined as the central datum when all of the data are arranged (ranked) in numerical order. As such, it is a literal measure of central tendency. When there are an even number of data, the mean (see below) of the two central data points is taken as the median. For the distribution of pain scores described above, the median pain score is again 2.

The median may be used for ordinal categorical data and for interval data. When analysing interval data, the median is preferred to the mean when the data are not normally (symmetrically) distributed, as it is less sensitive to the influence of outliers.

The mean

The mean is used to summarize interval data. As the mean may be influenced by outlying data points, it is best used as a measure of central tendency when the data is normally (symmetrically) distributed. Although several different means are defined, the arithmetic mean is most commonly used. The arithmetic mean is calculated by adding all the individual datum values in a dataset ($x_1 + x_2 + \ldots + x_n$) and dividing by the number of values (n) in the dataset.

The mean of ordinal categorical data is often reported in the literature (together with its associated measure of data spread, SD). For example, in the sample of verbal pain scores above, the mean score is 1.9. The use of mean and SD for ordinal data is controversial. For example, what does a pain score of 1.9 actually mean when using a categorical scale?

Measurement of spread of data (variability)

Once again, the first step in assessing spread of data is to examine it in either a table or an appropriate graphical form. A graph often makes clear any symmetry (or lack of it) in the spread of data, whether there are obvious atypical values (outliers) and whether the data is

Key points

Two key sample statistics are measures of central tendency and the spread of data about this central tendency.

The mode, median, and mean are used to describe, respectively, the central tendency of categorical data, interval data that is not normally distributed, and normally distributed interval data.

The most important distribution in stastistical analysis is the normal (Gaussian) distribution, defined by its mean and SD.

The normal distribution is characterized by its unimodal, symmetrical, bell-shaped frequency curve.

Anthony McCluskey BSc MB ChB FRCA
Consultant
Department of Anaesthesia
Stockport NHS Foundation Trust
Stepping Hill Hospital
Stockport SK2 7JE
UK
Tel +44: 0161 419 5869,
Fax: 0161 419 5045,
E-mail: a.mccluskey4@ntlworld.com
(for correspondence)

Abdul Ghaaliq Lalkhen MB ChB FRCA
Specialist Registrar
Department of Anaesthesia
Royal Lancaster Infirmary
Ashton Road
Lancaster LA1 4RP
UK

doi:10.1093/bjaceaccp/mkm020
Continuing Education in Anaesthesia, Critical Care & Pain | Volume 7 Number 4 2007

skewed in one direction or the other (a tendency for more values to fall in the upper or lower tail of the distribution).

Range

The simplest measure of data variability is the range, defined simply as the interval between the highest and lowest values in a distribution. It is of limited practical use in statistical analysis as it is obviously profoundly influenced by extreme outliers.

Percentiles

When a dataset is arranged in order of magnitude, it may be divided into 100 separate cut-off points (percentiles). The *x*th percentile is defined as a cut-off point such that *x*% of the sample has a value equal to or less than the cut-off point. For example, the 35th percentile splits the data up into two groups containing, respectively, 35 and 65% of the data.

Quartiles are used most commonly, i.e. lower (25th percentile), middle (50th percentile or median), and upper (75th percentile). They split the data into four equal groups. The interquartile range (IQR) is often quoted when referring to interval data that is not normally distributed. If the 25th percentile value (lower quartile) of a dataset is 10 and the 75th percentile value (upper quartile) is 40, the IQR may be expressed as either 10–40 or simply as 30. Percentiles and quartiles may be estimated from a cumulative frequency curve.

A useful graphical representation of the distribution of interval data is the box and whisker plot. For example, the box and whisker plot in Figure 1 has been produced from the marks obtained by a cohort of candidates taking an examination. The upper and lower limits of the box (hinges) represent the upper and lower quartiles, respectively. The horizontal line inside the box is the median and the whiskers represent extreme values, in this case the 10th and 90th percentiles. Any further outliers are represented by asterisks.

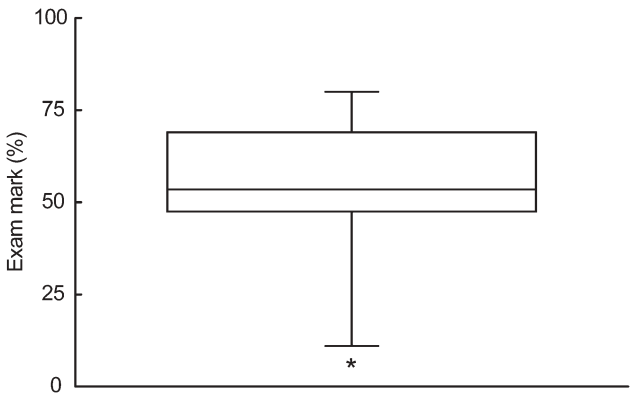

Fig. 1 Box and whisker plot of the examination marks. The data do not follow a normal distribution.

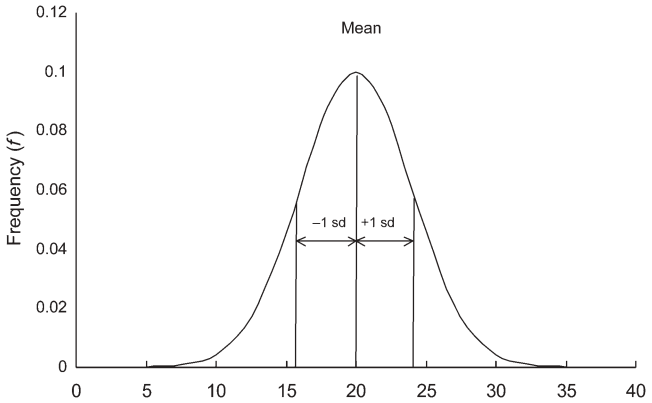

Fig. 2 The normal distribution [mean (*SD*) = 20 (4)].

The normal distribution

The most important and useful distribution of data in statistical analysis is the normal or Gaussian distribution (Fig. 2). It is also often referred to as a parametric distribution because two key parameters which fully describe its shape can be defined (the mean and SD). A normal distribution is characterized by a unimodal, symmetrical, bell-shaped curve when interval data are represented by a histogram or line graph.

Much biological data such as height and mean arterial blood pressure in healthy adults are normally distributed. In clinical trials, sample data drawn from such a population will also follow a normal distribution provided the sample size is reasonably large (e.g. >100 in each group). However, it is not actually necessary for sample data to follow a normal distribution in order to subject the data to parametric statistical analysis (which is often the case with the smaller sample sizes described in clinical studies). Rather, it is necessary for the sample data to be compatible with having been drawn from a population, which is normally distributed. Thus, although visual inspection of the frequency curve is useful and should always be undertaken, with smaller sample sizes it may not be obvious that the sample data is compatible with normally distributed population data. Data can be subjected to formal statistical analysis for evidence of normality using a variety of tests (e.g. Shapiro–Wilkes test, D'Agostino–Pearson omnibus test). However, these tests should be used with caution for very small samples as they may then give false positive results. If in doubt as to whether data is normally distributed or not, it is safer to use non-parametric inferential statistical analysis which is not based on any assumptions about the shape of the frequency distribution curve.

The normal distribution shown in Figure 2 has a mean of 20 and a SD of 4. The mean positions the frequency curve on the x-axis. The spread (width) of the curve around the mean is determined by its SD. The shape of any normal distribution frequency curve is entirely described by these two parameters. The centre of the distribution occurs at the zenith and all three measures of

central tendency (mode, median, and mean) are equal and described by the zenith.

For a population, the SD is calculated by summing the squares of all the individual differences of each datum value from the mean, then by calculating the mean of this value, and finally by calculating the square root. The process of squaring the differences followed by taking the square root results in all of the differences being converted into positive values. Otherwise, the SD would always be zero.

$$SD = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \mu)^2}{n}},$$

where μ = population mean and n = population size.

As clinical research rarely deals with populations but with (random) samples drawn from the population, the SD of a sample is:

$$SD = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}{n-1}},$$

where $\bar{x}$ = sample mean and n = sample size.

It is expressed in the same units as those of the mean from which it is derived.

For parametric (normally distributed, symmetrical) data, the mean and SD are the appropriate measures of central tendency and variability of the data. For non-parametric data, the median is the appropriate central tendency measure and the IQR is the appropriate measure of the variability of the data.

In a normal distribution, approximately two-thirds of the data (68%) lie within ± 1 SD of the mean, approximately 95% of the data lie within ± 2 SD of the mean (sometimes quoted more accurately as ± 1.96 SD) and approximately 99.7% of the data lie within ± 3 SD of the mean.

Several other sample statistics related to the sample SD are often used in statistical analysis, e.g. sample variance is defined as SD^2, standard error of the mean (SEM) = $SD/\sqrt{n}$

The standard normal distribution

The standard normal distribution is defined as having a mean = 0 and SD = 1. Any normal distribution may be converted into the standard normal distribution according to the formula: $z = (x_i - \mu)/\sigma$, where x_i is a datum value from the original distribution, μ is the mean of the original normal distribution, and σ is the SD of original normal distribution. The standard normal distribution is therefore sometimes referred to as the z-distribution. A z-value indicates the number of SDs, a datum value is above or below the mean.

The standard normal distribution may be useful in comparing different normal distributions. For example, suppose we wish to consider two candidates for a job by comparing their performance in a qualifying examination set by different examination boards. They might both have scored 60% but it would not be surprising if the mean and SD of the marks obtained in each of the examinations was different. The raw marks obtained by candidates may be converted into z-values, equivalent to the number of SDs that the scores are from the mean of zero, according to the equation: z-value = $(x_i - \mu)/\sigma$, where x_i = a candidate's raw score, μ = mean mark of the population of all the candidates in each particular examination, and σ = SD of the population of all the marks obtained in each particular examination.

If candidate 1 scored 60% in an examination with a mean mark of 70% and SD of 10%, the corresponding z-value mark is -1. His performance was 1 SD below the mean for that examination, i.e. his mark was at the level of the 16th percentile [50 − (68/2)] for his cohort. If candidate 2 scored 60% in an examination with a mean mark of 40% and SD 10%, the corresponding z-value mark is $+2$. His performance was 2 SD above the mean for that examination, i.e. his mark was at the level of the 97.5 percentile [50 + (95/2)] for his cohort.

Even with the same pass mark, candidate 2 is seen to have performed better. In this context, the z-values are equivalent to the candidates' standard normal scores in their examinations. Strictly speaking, candidate 2 performed better in relation to his cohort, not to an absolute standard. If the two cohorts were of widely differing general ability then our conclusion that candidate 2 was more able based on his better standard normal score might be invalid.

Standard deviation *vs* SEM

When the results of statistical analyses are reported, the SD and SEM are sometimes used inappropriately. For example, authors may quote the range of their sample data as mean ± SEM rather than mean ± SD. The temptation to do this follows from the fact that by definition, the SEM decreases as sample size increases (SEM = $SD/\sqrt{n}$). When statistics comparing two or more groups are quoted in this way, any differences between them appear more significant than when the SD is quoted. The SD should always be used when describing the variability of the actual sample data. The SEM is used specifically to describe the precision of the sample mean, i.e. how far is the sample mean from the population mean. The 95% confidence interval (see future article) for the population mean = sample mean ± 2 SEM.

Skewness and kurtosis

Two terms may be defined with reference to the shape of a frequency curve, kurtosis and skewness. Kurtosis describes the peakedness of the curve whereas skewnesss describes the symmetry of the curve. A positive kurtosis indicates a frequency distribution with a sharper peak and a longer tail than a normal distribution; a negative kurtosis indicates a wide, flattened distribution. The standard normal distribution has a kurtosis of zero. If a frequency curve has a longer upper tail, the data is positively skewed; if the data has a longer lower tail, it is negatively skewed.

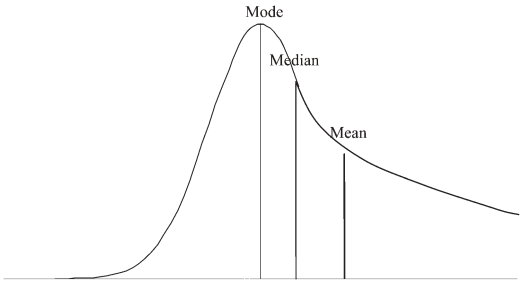

Fig. 3 Positively skewed data.

The distribution shown in Figure 3 is positively skewed. It is evident from the frequency curve that the mean has been shifted to the right by the skewed data; although the median has also shifted, it has been influenced less by the outliers on the upper tail and is the appropriate measure of central tendency for skewed distributions. Biological data not infrequently follows such a distribution, examples being adult weight, white cell count of healthy individuals, and serum triglyceride concentrations. Negatively skewed data is relatively uncommon.

When faced with a sample that comprises non-normally distributed (skewed) data, there are two choices: to accept the distribution as it is and use non-parametric inferential statistical analysis or to attempt to transform the data into a normal distribution. Common methods of transforming skewed data into a normal distribution are logarithmic, square root, and reciprocal transformations.

Acknowledgements

The authors are grateful to Professor Rose Baker, Department of Statistics, Salford University for her valuable contribution in providing helpful comments and advice on this manuscript.

Bibliography

1. McCluskey A, Lalkhen AG. Statistics I: Data and correlations. *CEACCP* 2007; **7**: 95–99
2. Bland M. *An Introduction to Medical Statistics*, 3rd Edn. Oxford: Oxford University Press, 2000
3. Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC, 1991
4. Rumsey D. *Statistics for Dummies*. New Jersey: Wiley Publishing Inc., 2003
5. Swinscow TDV. Statistics at Square One. http://www.bmj.com/statsbk (accessed 14 June 2007)
6. Lane DM. Hyperstat Online Statistics Textbook. http://davidmlane.com/hyperstat/ (accessed 14 June 2007)
7. SurfStat Australia. http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html (accessed 14 June 2007)
8. Elwood M. Critical Appraisal of Epidemiological Studies and Clinical Trials, 2nd Edn. Oxford: Oxford University press, 1998

Please see multiple choice questions 14–17