A concise guide to... descriptive statistics

By Gill Marshall and Leon Jonker

Introduction

It has been reported that the inability of healthcare staff, including nurses and allied health professionals, to understand statistics is a frequent barrier that stops them from locally applying the latest interventions, outcomes and methods identified through research. This handicap ultimately lessens the chance of healthcare professions undertaking evidence base practice[1-3]. Our aim here is to demonstrate that knowledge of statistics can be gained through self-teaching and by accessing relevant literature, and to introduce readers to statistical terminology and the basic elements of the application of descriptive statistics.

Relevance of understanding statistics

Radiography became an all degree profession in 1992, and has striven to become an autonomous profession. This means it must 'carve out a knowledge base that is dynamic and forward thinking'[4] and therefore radiographers need an understanding of statistics when reading papers or indeed applying them to their own studies. 'Today's medicine is only as good as the research 15 years ago' demonstrates the significant time taken between reporting findings of research and them affecting clinical practice[5].

If a radiographer is undertaking a research project, he/she must be statistically literate in order to work meaningfully with other researchers and statisticians to ensure that a project has the optimal methodological design. To ensure quality improvement[6], radiographers need to be able to measure parameters.

It is documented in literature that statistical teaching of radiography and radiology students is inadequate, and that students forget what they were taught, particularly if it is not applied[7,8]. Telling someone how to read and dissect statistics is most effective when the statistical concepts are taught first, before starting to teach 'how to undertake' the statistics[9].

Types of data

Statistics use numerical output to demonstrate the meaning of data, eg, the patient's blood pressure, or assign numbers to qualitative attributes such as eye colour. Descriptive statistics can be used to illustrate the characteristics of a group of observations, ie, the raw data[10]. In simple terms, they allow one to determine what the average measurement is: for example, the average volume of a tumour when a few hundred tumours are measured in different patients (measurement of central tendency). They also allow one to see if there is a small or wide variety in the measurement of a certain variable (measurement of dispersion), eg, by determining the smallest and largest tumour measured. In this manner, one can research if there are trends over time in terms of tumour volume, or if perhaps a patient's age is a factor in the size of a tumour.

Descriptive statistics are not capable of showing causality; this requires the use of inferential statistics. However,

they are an essential platform from which successful inferential statistics are carried out.

It is important to know what kind of data is to be collected or what data has been collected. The type of data determines how it can be summarised and also what subsequent inferential statistics can be carried out. There are different levels of measurement for scoring variables, with two main categories: categorical data and continuous data.

Within categorical data there are four sub-categories, although the last sub-category (interval/ratio data) progresses into continuous data. The different types of categorical data are:

◆ **Binary data:** only two outcomes or measurements are possible, eg, if the patient's survival is recorded there can only be two outcomes – survival or death. Another example is whether or not a radiological report has been filled in correctly. Such data is usually summarised using proportions or odds.

◆ **Nominal data:** this is the least robust type of data for categorising data into mutually exclusive categories, without ranking, and can be useful. An example could be blood groups, ie, O, A, B, AB, or distinguishing groups of professions, eg, radiographers, radiologists and nurses. Binary data and nominal data can be represented or stored by allocating numbers to categories (radiographers = 1, radiologists = 2, nurses = 3), where the numbers have no numerical significance[10]. With this data, the measure of central tendency (the category with the most cases), is known as the mode.

◆ **Ordinal data:** this has a clear order or hierarchy but not on a calibrated scale, eg, strongly agree, agree, etc, from a Likert scale with a statement provided[11]. An example is the level of pain caused by mammography, eg, no pain, mild discomfort, moderate pain, severe pain. The categories are mutually exclusive, whilst the numeric values attributed to them are not absolute measurements but do order the data. In ordinal data, the measure of central tendency is called the median, so the category which is the middle of the rank-ordered description is referred to as the median.

Binary, nominal and ordinal scales are considered discrete variables because the data is classified into discrete non-overlapping variables[12].

◆ **Interval/ratio data:** this is the strongest type of data, with ratio data being the stronger of the two because it has a true zero value[12]. Such data is achieved by the use of, for example, calibrated scale to provide quantitative measurements, density readings from a densitometer, weight in kilograms, or blood pressure.

Ratio or interval data may be summarised by the mean or the median (as a measure of central tendency) depending on the distribution. To illustrate the point, an example of interval data would be an IQ test score. Ratios of these measurements cannot be applied – an IQ of 140 in subject 1 versus 70 in subject 2 does not mean subject 1 is twice as clever as subject 2 but, on the other hand, examples of ratio data show that they all have a constant scale which includes a zero, eg, 200 metres is twice as long as 100 metres.

A potentially confusing element is that interval data can be both categorical and continuous; ratio data is classed as continuous. For interval/ratio data from a continuous scale, the range, inter-quartile range, and standard deviation are used to report the spread or width of the data. Data from interval or ratio scales which is continuous data provides information on continuous variables, because the data represents an underlying continuum where there are potentially an infinite number of values.

Table 1 summarises the different levels of measurement and measures of central tendency and dispersion tend to be presented. Standard software such as Excel, as well as specialist statistics software such as SPSS are capable of calculating the mode, median and mean.

Table 1: Descriptive statistics for each level of measurement.

Level of measurement	Description	Measure of central tendency	Measure of dispersion
Nominal	Classification	Mode	Frequency distribution
Ordinal	Relative rankings	Median Mode	Frequency distribution Percentile Maximum and minimum Range
Continuous	Rank ordering with equal intervals	Mean* Median Mode	Frequency distribution Percentile Maximum and minimum Range, inter-quartile range Standard deviation

* If the distribution of data is not normal,eg, there are a number of outliers that influence the mean dramatically (making the data skewed), it is sometimes better to present the median to give a better reflection of the value of most of the data points.

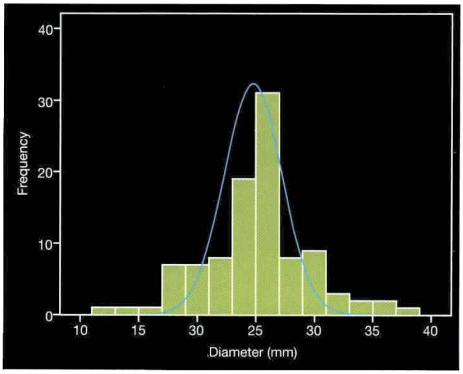

Figure 1. Example of data with a normal distribution. A hypothetical measurement of an artery in mm of 100 patients, presented through a histogram, with Gaussian bell curve highlighting the normal distribution pattern.

Distribution of data

Descriptive statistics use numerical procedures or graphical techniques, eg, bar charts, histograms, frequency polygons and pie charts, to organise, present and describe the characteristics of a sample. Descriptive statistics seek to describe the mid-point of a spread of scores, the measure of central tendency, and the spread of scores – the dispersion – of which variance is an example[13].

If measurements are taken from a large random sample, eg, of the weight of adult patients having contrast enhanced MRI, and a frequency polygon is plotted of the results, it is likely that a bell shaped curve will be produced which shows that the variables of a sample are normally distributed. This bell shape is called a normal or Gaussian distribution (Figure 1). The word 'normal' here means that the data complies with a distribution pattern that mathematically allows parametric statistical tests to be applied.

In radiography, normal distribution of measurements occurs when plotting the sizes and volumes of anatomical structures, such as the eye ball, optic nerve or, as shown in Figure 1, vascular structures. These measurements do not differ much between people. However, a normal distribution may not be achieved; outliers commonly occur which will give

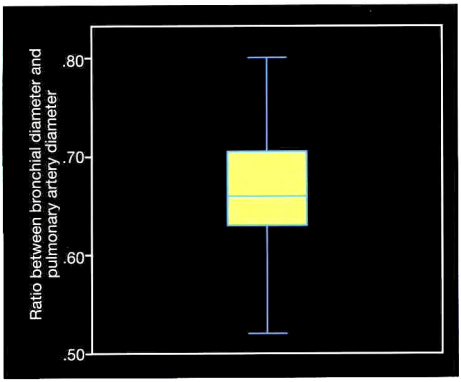

Figure 2: Data for the ratio between bronchial and pulmonary artery diameters summarised in box plot format.

the bell curve a 'tail' at either the left hand or right hand side.

There are ways to correct for skewed distribution of data when one wishes to apply subsequent inferential statistics[14]. The importance of normal distribution relates to whether a so-called parametric or non-parametric test can be applied. Our article on inferential statistics, to follow in next month's *Synergy Imaging & Therapy Practice*, will discuss this in more detail.

Standard deviation is a measure of how spread the data is, ie, the variance of it. The narrower the standard deviation, the closer to the midpoint of the data all results will be[15]. The standard deviation expresses variance using the same units as used for the observations or measurements. Approximately two thirds of all observations or measurements will lie within one standard deviation of the mean (the *top* of the distribution graph in Figure 1), and 95% lie within two standard deviations of the mean. The inter-quartile range is becoming a more common way to report descriptive statistics from continuous data. This statistic represents the middle 50% of the sample showing its dispersion, and is not influenced by outliers[16].

Collection and presentation of data

At the design phase, the researcher must know what data must be collected. It is a common failing not to pay attention to this, and it can result in the research question remaining unanswered because the collected data does not provide the necessary information. The level of measurement must be identified, so that the statistical procedure to be used can be chosen and a decision made on the sample size.

Descriptive statistics are the most straightforward statistics to undertake and interpret. They usefully summarise data and provide a description of the sample. Figure 2 shows a common example of how to summarise data; the hypothetical ratio between two measurements of the bronchial and pulmonary arteries was recorded for 100 patients. A box plot (or box and whisker plot) can present a number of statistics. It shows the 25th and 75th percentile values with the bottom and top bottom of the box respectively; the bar inside the box represents the 50th percentile (median) and the 'whiskers' attached to the box represent the lowest and highest values (the range) of the data, bar extreme outliers or extreme values.

Conclusion

Statistics provide a way of describing numerical data so that the data can be reviewed and analysed. Such observations can then be used as a starting point to probe the reasons behind these characteristics and trends, usually by using inferential statistics.

It is essential that the type of data collected and its analysis is appropriate so that the research question can be answered meaningfully. Consequently, it is vital to consider what type of data will be collected and presented as soon as the research question is identified.

Accreditation by cpd now the College of Radiographers

Test Yourself

Below are some questions for you to answer which you can then count towards your CPD. Note down your answers and any other observations and put them in your CPD folder. If you record this activity in CPD Now, remember that you can scan your paperwork and attach it electronically to your CPD record. The answers will be available online from 1 October, under 'Synergy resources', at: *www.sor.org/members/pubarchive/synergy.htm*

1. What is the name of the most frequent response to a question?
2. What is the median of a sample?
3. Is a percentage a descriptive or inferential statistic?
4. What is meant by standard deviation?
5. What is a normal distribution?
6. What is the central tendency of a sample?
7. What is the mean value of a sample?
8. What is meant by dispersion?
9. Give an example of nominal data.
10. What is binary data?

Checklist for understanding and applying descriptive statistics:

✓ Do you just want to describe your data and summarise it? If so, descriptive statistics will suffice.

✓ What measure of central tendency is appropriate to the data?

✓ Are your variables discrete or continuous?

✓ Is your data of nominal, ordinal or ratio/interval variety?

✓ Once you have undertaken descriptive statistics, is there more information to be gleaned from the data? If so, inferential statistics may be appropriate and allow inferences/conclusions to be drawn from the data.

About the Authors

Professor Gill Marshall is senior fellow and national teaching fellow of the HE Academy, and research development lead at the University of Cumbria.

Dr Leon Jonker is senior research fellow at the University of Cumbria.

Readers who wish to read a more comprehensive paper on this subject are recommended to access *Descriptive Statistics: a review and practical guide*, Marshall G & Jonker L, published in *Radiography*, and available online at: *www.radiographyonline.com/article/S1078-8174(10)00002-7/fulltext*

References for this article are under 'Synergy resources' at *www.sor.org/members/pubarchive/synergy.htm*

To comment on this article, please write to *racheld@synergymagazine.co.uk*

Glossary

Bar charts: used where the variable plotted along the horizontal axis is nominal or ordinal, ie, the categories shown along this axis have no quantitative or numerical value (nominal), or clear order or hierarchy but not on a calibrated scale (ordinal). The vertical bars are drawn separately, eg, the frequency distribution of the number of qualified radiographers in the UK categorised by age group of the radiographers.

Cohort: a defined group of people, eg, all women in the Morecambe bay area who have been screened for breast cancer in the last year.

Continuous variables: the data represents an underlying continuum where there are potentially an infinite number of values.

Data: numbers or values collected as a result of measurements. They could be counts or frequencies or actual numerical values or scores.

Descriptive statistics: those that can be used descriptively to illustrate the characteristics of a group of observation, ie, the raw data.

Discrete variables: when the data is classified into discrete not overlapping variables.

Frequency polygon: an alternative to a histogram and preferred when the variable is continuous. It is produced by placing dots at the tops of the centres of the bars of the equivalent histogram, and then joining the dots with straight lines. It is often used to compare two frequency distributions.

When measurements are taken from a large random sample and a frequency polygon is produced, the shape of the distribution is always the same. This is a bell shaped curve called a normal or Gaussian distribution.

Histograms: where the categories are measured on a numerical scale. The vertical bars are drawn touching, eg, the frequency distribution of the number of children per family living in Lancaster. Histograms can be used for discrete or continuous data. If used for continuous data, it is partitioned into ranges (the size of which depends on the sample size) to enable an optimum number of bars to be displayed.

Inferential statistics: those that can be used to infer generalisations from the sample group that can be applied to a wider population.

Interval data: stronger data than nominal or ordinal data that can be achieved by the use of a calibrated scale to provide quantitative measurements. The difference between interval and ratio data is that ratio data has a true zero, thus interval data is weaker.

Likert scale: commonly used in questionnaires, and the most widely used scale in survey research. It is used for ordered category data. When responding to a Likert questionnaire item, respondents specify their level of agreement to a statement. The scale is named after Rensis Likert who published a report describing its use[11].

Measurements of central tendency: (see also Table 1).

Mode – the numerical value with the greatest frequency

Median – the middle score of a rank ordered distribution

Mean – the average score

Measurements of dispersion: (see also Table 1). Note that variance is an example of dispersion.

Frequency distribution – the number of cases per category

Range – the distance between the highest and lowest score

Inter-quartile range – the range within which the middle 50% of the scores fall

Standard deviation – the root-mean-square deviation from the mean

Variance – the square of the standard deviation

Nominal data: the least robust data which categorises, but does not hierarchically rank, data into mutually exclusive categories.

Normal distribution: when a large number of measurements are made at random of one particular variable, the results usually fall into a pattern. Most of the measurements will lie close to the mean value, with few values lying at the extremes. When a frequency distribution is plotted, a familiar bell shaped curve is produced which represents a normal or Gaussian distribution.

Ordinal data: where the data has a clear order or hierarchy but not on a calibrated scale. The ordinal data can be ranked (such as from tallest to smallest) or ordered (eg, data categorised via the level of agreement with statements on a Likert scale, such as very painful, painful, etc). Both numerical and non-numerical can constitute ordinal data.

Parameter: a 'true' measurable characteristic of the population that cannot, in practice, be known with certainty, eg, the mean breast thickness of all women eligible for mammography breast screening. A 'statistic', ie, the mean value in a sample of patients is measured and used as an estimate of the true mean of the population.

Parametric tests: are selected for data that is normally distributed.

Pie charts: used to demonstrate distribution ratios because it is easy to compare the relative sizes of the various components.

Population: a complete set of individuals, objects or measurements having some observable characteristic in common, eg, all women who have had one child.

Ratio data: the strongest data, and has a true zero value. It can be achieved by the use of a calibrated scale to provide quantitative measurements.

Sample: a subset of the population, selected to participate in a study.

Standard deviation: figure that gives an indication of how closely distributed measurements or observations are around the mean of all the measurements. It gives an indication of the variance of the data. A small standard deviation will result in a steep-curved bell curve when presented in a graph. Oppositely, a large standard deviation will result in a flat bell curve.