

## Statistics for Clinicians

### 2: Describing and displaying data

JB CARLIN<sup>1,4</sup> and LW DOYLE<sup>2-4</sup>

<sup>1</sup>The Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital Research Institute, <sup>2</sup>the Division of Newborn Services, the Royal Women's Hospital, the Departments of <sup>3</sup>Obstetrics and Gynaecology, and <sup>4</sup>Paediatrics, the University of Melbourne, Parkville, Victoria, Australia

Before starting any study or collecting any data, a researcher should ensure that data management, including data collection, data 'cleaning' and analysis, is an integral part of the study design. Methods for guaranteeing that data are collected and managed in a reliable way should be implemented before the study begins. However, whether or not this ideal is achieved, it is always important to examine data in an exploratory, descriptive, manner. Such examination should include graphical displays as well as descriptive summaries using frequencies, means, and other *summary statistics*. These descriptive analyses serve at least two major purposes:

1. Errors in the data can be detected and corrected before too much time has been wasted on analyses giving incorrect results. This process is known as *data cleaning*, and is often carried out most effectively using graphical techniques. For example, a child's age entered as 120 years, instead of 12 years, is quickly detected as an error.

2. The main patterns of group differences or associations in the data may be apparent, providing a check with the results of later more formal analyses. This type of data analysis is often called 'exploratory' but it should be closely guided by the research questions underlying the study, and not just be a 'fishing expedition'.

The methods to be discussed in this article thus span a range of ideas that are useful both in preliminary analysis and data cleaning, as well as in the final presentation of results. It is important to distinguish these two purposes, since there is usually a great deal of work behind the scenes that does not appear in the concise summaries required when reporting research findings.

To illustrate methods of descriptive analysis, we will use data collected in the study of a consecutive cohort of very low birthweight (VLBW), birthweight 500–1499 g, infants born in the Royal Women's Hospital, Melbourne, in 1992 (Table 1). There were 202 live births, and 165 (81.7%) survived their primary hospitalization. The survivors have been assessed at 2 and 5 years of age. There are data on over 800 individual variables. Data on these infants were entered onto a personal computer and

arranged in a rectangular array, where each row represents an individual infant, and each column an individual variable. Each variable has a unique name, and each infant has a unique identifying number (recorded in a variable named IDNUM). This rectangular structure of subjects (rows) by variables (columns) is a format required for all standard statistical analysis. As an aside, it should not be assumed that the best means of *entering* the data is directly into a spreadsheet or statistics package; this can sometimes be done more efficiently and with less potential for error by using a database program such as the freeware package Epi-Info (Center for Disease Control and Prevention, Atlanta, GA, USA),<sup>1</sup> or more powerful products such as Microsoft Access (Microsoft Corporation, Redmond, WA, Australia).

It is difficult to gain an understanding of the meaning of the data by examining the entire rectangular array; part of any analysis must be the *reduction* of information to meaningful summaries. The important point when displaying and describing data is to convey a message in the simplest and clearest way. The appropriate level of simplicity and clarity may vary with the sophistication of the audience, e.g. it is important to remember that most readers of the journal are not statisticians! Describing the data may involve tables or diagrams, which in turn may represent raw data or summary values of various kinds. Graphical displays are often most useful to analysts themselves, and may not necessarily be the best means for conveying results to others, including journal readers. There are several ways the data in our rectangular array can be examined and described, depending upon the type of variable(s) being considered.

#### CATEGORICAL (DISCRETE) VARIABLES

Variables whose values are limited to a small number of categories (e.g. gender, disease status) can be summarized simply by counting and displaying the frequencies with which subjects fall into each category. Such frequencies are generally best displayed in *relative* form, e.g. as percentages (or proportions), although care must be taken to define the denominator for such calculations. Moreover, frequencies may be displayed either in tables or in diagrams. As an example, in Table 2 we display a frequency tabulation of gender in the cohort of VLBW infants.

The relationship between two separate categorical variables is best shown in a *cross-tabulation*, as illustrated in Table 3. This table provides an example where the final presentation in an article would probably differ from the table examined as part of the exploratory analysis. The latter would have the detailed form as shown in Table 3 (and possibly other

Correspondence: Associate Professor LW Doyle, Department of Obstetrics & Gynaecology, The University of Melbourne, Parkville, Victoria 3052, Australia. Fax: (03) 9347 1761; email: l.doyle@obgyn-rwh.unimelb.edu.au

JB Carlin, PhD, Statistician. LW Doyle, MD, MSc, FRACP, Paediatrician.

Accepted for publication 25 January 2000.

**Table 1** Structure of data file for a very low birthweight cohort, 1992. The data file itself should be accompanied by a *codebook* that documents the definitions of all variables, and includes necessary codings (e.g. the variable 'MUMSRES' identifies the mother's place of residence, with codes 1 = 'Victoria', 2 = 'elsewhere').

IDNUM	HOSPCARE	HOSPBIRT	BOOKED	MUMSRES	COUNTRYB	-
70049	7	7	1	1	1	-
70051	7	7	2	1	1	-
70056	7	7	2	2	1	-
70057	7	7	1	1	11	-
70058	7	7	3	1	11	-
70059	7	7	2	1	1	-
-	-	-	-	-	-	-

**Table 2** Gender distribution of very low birthweight cohort, 1992

Gender	Frequency	% of Total
Male	102	51
Female	100	49
Total	202	100

**Table 3** Relationship between gender and requiring assisted ventilation after birth in survivors. Values shown are counts (with percentages of row totals)

Gender	Assisted ventilation		Total
	Yes (%)	No (%)	
Male	56 (74.7)	19 (25.3)	75
Female	68 (75.6)	22 (24.4)	90
Total	124 (75.2)	41 (24.8)	165

percentage calculations as well), while in a final presentation it would be redundant to report the column of 'No' percentages as well as the column of 'Yes' percentages. Notice that we present percentages by row since it is likely that the main interest here would be a comparison between males and females in the proportions receiving assisted ventilation.

*Diagrams* can also be useful in summarizing categorical variables. In particular, most readers will be familiar with *pie charts* and *bar charts*. A pie chart represents the relative proportions of the subgroups defined by a categorical variable by the areas of the slices of a pie. Although still popular in the media and other non-scientific presentations, it has been shown that pie charts do not provide strong visual information for comparing proportions between categories.<sup>2,3</sup> Better visual signals are conveyed when comparisons are represented as linear distances, as in a bar chart (Fig. 1).

When there are several subcategories within a variable, and when the variable is contrasted across another variable defining several subgroups, multiple bar diagrams or stacked bar charts may be useful. In general, however, bar charts tend to be overused and alternative methods using dot plots, or points and lines are often more effective, again because they allow better perception of comparisons. Examples are discussed in excellent treatises on graphical representation of data by Tufte<sup>2</sup> and Cleveland.<sup>3</sup>

The choice between a table and a diagram depends largely on the purpose of the display. Generally, diagrams are better for quick visual exploration and for oral presentations, to emphasize where differences and similarities lie across subgroups. For a journal article, a diagram such as Figure 1 is an inefficient use of space since it really only compares four numbers. Furthermore, a limitation of a diagram is that it may be difficult to determine the exact percentages of subgroups. Tables (including cross tabulations) are often better for written presentations because they convey summary results using less space and in more explicit detail, enabling readers, for example, to combine frequencies over two or more subgroups.

Adding the exact percentages and/or numbers to bar charts can combine the best features of tables and diagrams, although care must be taken to avoid cluttering the visual information.

**CONTINUOUS (METRIC, QUANTITATIVE) VARIABLES**

By definition, continuous variables take on a large number of distinct values, so it is only possible to summarize them either by grouping the data or by using various *summary statistics*, such as means, medians, etc. Conversely, the variation in the raw values often makes graphical displays very informative, especially for exploratory purposes.

**Frequency distributions and histograms**

Sometimes values of a continuous variable are grouped to create categories or ranges, so that the data may be presented as a *frequency distribution*, just as for a categorical variable. A *histogram* is the best-known graphical representation of a frequency distribution. In a histogram the area of an individual bar represents the relative frequency of values within the specified range. For this reason, the vertical axis (ordinate) should always start at zero. The number of subgroups (often referred to as 'bins') is usually between 10 and 20, although the choice of number of bins can have a dramatic effect on the appearance of the histogram, especially if the overall sample size is only modest (Fig. 2).

A *frequency polygon* is produced by using a line to join the midpoints of the top of each bar in a histogram, and may sometimes be effective in presenting two or more frequency distributions alongside each other. Neither histograms nor

frequency polygons, in fact, provide ideal ways to display frequency distributions obtained from continuous data. An underutilized alternative is the cumulative distribution plot, which graphs the cumulative frequency of values observed either above or below each point on the abscissa (*x*-axis). This graph is especially effective for making comparisons between distributions and has, for example, been promoted as a tool in examining serological conversion rates from immunisation studies<sup>4</sup> (Fig. 3).

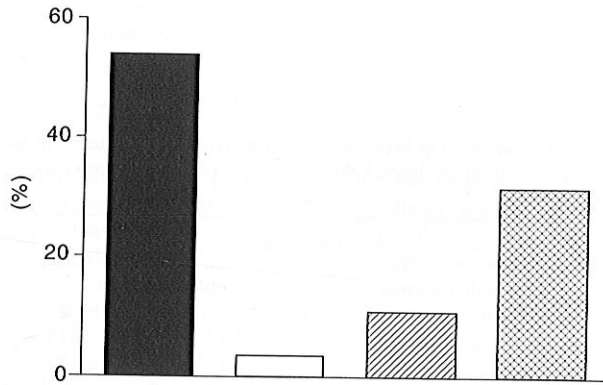


Fig. 1 A bar chart showing the distribution of mode of delivery among infants in the very low birthweight cohort. (■), Vaginal; (□), elective caesarean section; (▨), emergency caesarean section, no labour; (▩), emergency caesarean section, labour.

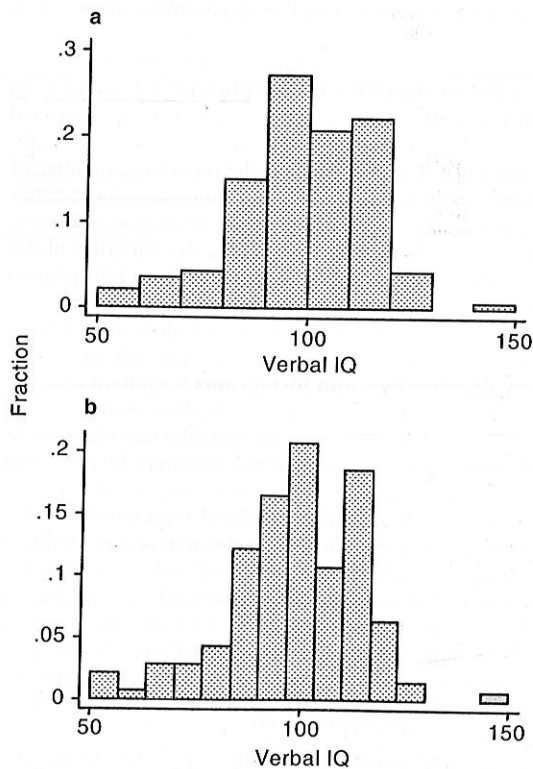


Fig. 2 Histograms of verbal intelligence quotient (IQ) at age 5 years, using two different choices of bin width. (a) 10 bins, (b) 15 bins (*n* = 140).

**Dot/bubble plots**

Where subjects in a study can be divided into two or more groups, *dotplots* or *bubbleplots* provide another useful way of comparing a continuous variable between the groups. Since this display shows every data value, it is an especially revealing (and honest!) way of presenting results, although it can be difficult to see subtle differences between groups which may still turn out to be important (and even statistically significant) (Fig. 4).

**Box plots**

For similar intergroup comparisons, *boxplots* (sometimes called *box-and-whisker* plots) are often preferable when there

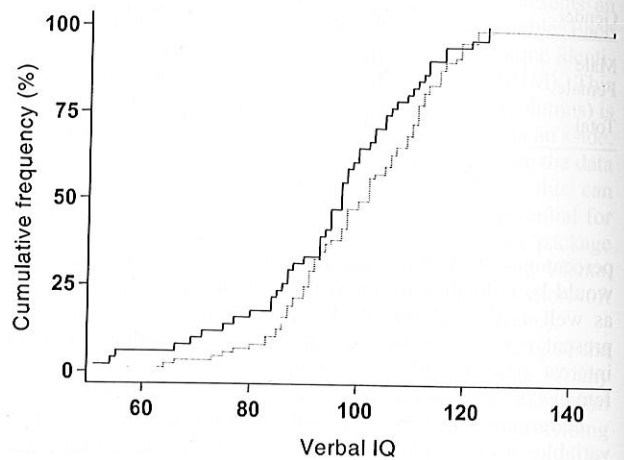


Fig. 3 Cumulative frequency distribution of verbal intelligence quotient (IQ) at age 5 years, by two birthweight groups. (—), < 1000 g; (---), 1000–1499 g.

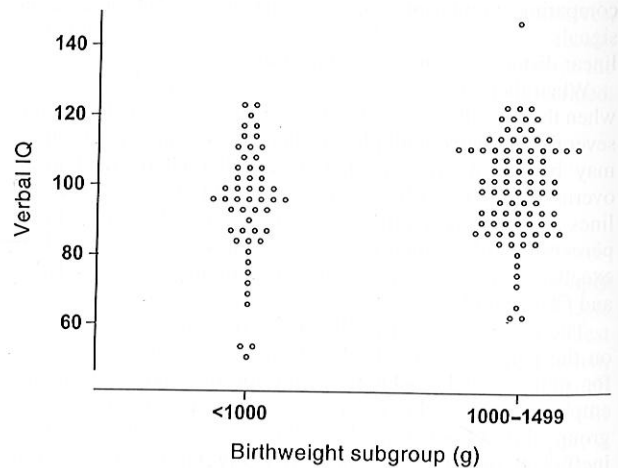


Fig. 4 Dotplot (or bubbleplot) of verbal intelligence quotient (IQ) at age 5, by two birthweight groups. In comparison with Figure 3, it is not quite as easy to see that the lower birthweight group produces lower scores (across the range of data); on the other hand, we easily notice a high outlier in the heavier group, and three very low values in the lighter group.

is a large number of individual observations, or a larger number of groups to be compared. Although there are minor variations between statistical programs, a typical box plot displays the interquartile range as the upper (75th centile) and lower (25th centile) ends of the box. The median (50th centile) is shown as a solid bar inside the box, a range representing 1.5 times the interquartile range is indicated with 'whiskers' extending from the box, and any more extreme values are marked separately outside this range ('outliers') (Fig. 5).

**Two continuous variables: Scatterplot**

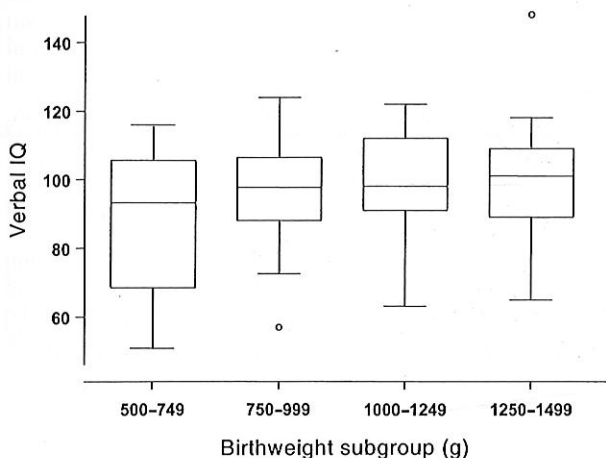
The relationship between two continuous variables can be shown in a scatterplot, and these sometimes provide a sobering antidote to over-enthusiastic interpretation (Fig. 6)! It has been shown that the visual interpretation of a scatterplot (the perceived strength of correlation) is affected by the scales used on the axes,<sup>5</sup> and it is best to ensure that the scales leave a small amount of white space on all sides of the cloud of plotted points.

The following general points should be observed when presenting results in tables or figures:

- It is important that the caption of every table or figure presented in an article should contain a full description, so that readers can understand the data without necessarily referring to the text.
- Totals should be indicated.
- The denominator (base) for any percentages should be clearly stated.
- The units for any measurements should be included in the table or in axis labels if using a graph.

**SUMMARY STATISTICS**

When we summarize variables for presentation in tables, there is often a choice of *statistics* (plural: 'single number summaries of data') that might be used. Categorical data are best described



**Fig. 5** Boxplots of verbal intelligence quotient (IQ) at age 5 years, by four birthweight groups. This diagram clearly indicates that the lowest birthweight category (< 750 g) has the widest variability in outcome, and also suggests a trend towards increasing verbal IQ with birthweight.

in frequency tables with their corresponding proportions or percentages, as outlined earlier. For continuous variables, more sophisticated summaries are required. These are traditionally divided into two main categories, *measures of location or central tendency*, which give a 'central' or 'typical' value to represent the distribution of actual values, and *measures of spread or dispersion*, which summarize the variation that the values exhibit.

**Measures of location**

Most of us are familiar with the terms *mean*, *median* and *mode*.

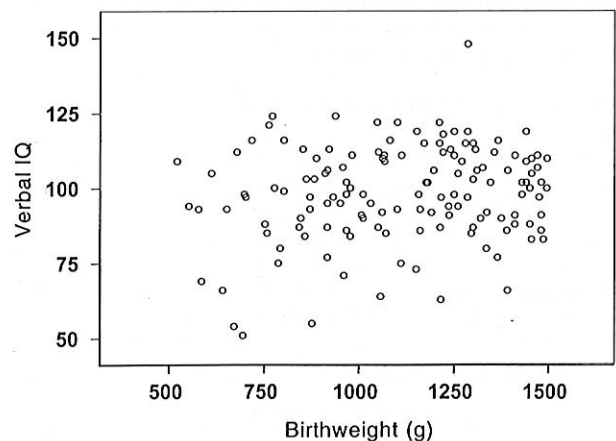
The *mean* (strictly called the *arithmetic mean*, but referred to as the *average* in everyday language) is calculated by adding all the numbers and dividing by the number of individual observations (the *sample size*).

The *median* is the middle observation of the group, also called the *50th centile*. With an odd number of observations, it is the value with index number  $(n + 1)/2$  after the data have been sorted in ascending order; with an even number of observations the average of the two middle values is conventionally reported.

The *mode* is the most common value in a series of observations.

Of the three location measures, the *mean* is the most commonly used since it is best understood, and it turns out that statistical inference is easier to perform for means than for other population summaries (see later in this series). There are times, however, when the *median* provides a better summary of a distribution, particularly when data have a *skewed* distribution. The *mode* is rarely used as a summary statistic, but most of us have heard of the term *bimodal*, which indicates that there might be subgroups within a particular sample.

We illustrate with the distribution of behaviour score at age 5 years in the VLBW cohort (Fig. 7). Since the distribution is skewed to the right, the mean (25.3) is higher than the median (22). The mode is actually 12 (a value recorded by eight children, 5.6% of the cohort), but this does not provide a useful summary of a variable that takes a nearly continuous range of values.



**Fig. 6** Scatterplot of verbal intelligence quotient (IQ) at age 5 years with birthweight ( $n = 140$ ). The same weak trend as in Figure 5 may be seen, but the scatterplot highlights the large variation between individual outcomes.

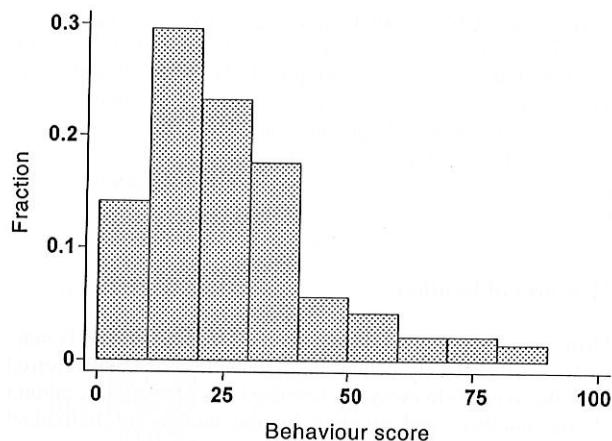


Fig. 7 Histogram of behaviour scores at age 5 ( $n = 142$ ), illustrating skewed distribution (mean = 25.3, median = 22).

Another measure of location that is sometimes encountered is the *geometric mean*. The geometric mean is defined as the  $n^{\text{th}}$  root of the product of all observations (related to averaging in the logarithmic scale), and will be further discussed later in the series when we deal with transforming data. With positively skewed distributions it is worth knowing that it is often close in value to the median. For the behaviour score, the geometric mean was 19.7 (omitting one zero value).

### Measures of spread

How spread out are the data? How much *variability* is there around the central value?

The *range* is the difference between the highest and lowest value. It is useful in data editing or cleaning, for checking if any values fall outside the range of possible values. However, it is a poor indicator of spread since it is determined by the two most extreme values and these will vary greatly between different samples even from the same population.

The *interquartile range* is the difference between the *lower* and *upper quartiles*. These terms were discussed earlier under *boxplots* (Fig. 6). The interquartile range is particularly useful when combined with the median in describing the spread of skewed data.

The *standard deviation* is the most commonly used measure of spread since it complements the mean, and because of its usefulness in statistical inference, as we will see in subsequent issues of the series. Standard deviation is commonly abbreviated to SD in publications, but statisticians often use the shorthand 's'. The computation of the standard deviation looks confusing to those unfamiliar with algebra, but in English it is calculated by taking the mean away from each individual value, squaring the result, summing these values for all observations,

dividing by one fewer than the sample size, and then taking the square root. It thus represents a kind of average variation of values around the mean value, where the squaring (and subsequent square root) may be regarded as a trick to remove the sign that represents the direction of variation above or below the mean. A statistic that is closely related to the SD is the variance, which is the SD squared. We will meet this term again later in the series, since it is used in some of the techniques of statistical inference.

The mean and SD are useful as summaries of normally distributed (we will discuss the *normal distribution* later in the series) data since approximately two-thirds of data points lie within 1 SD either side of the mean, 95% lie within 2 SD either side of the mean and almost all lie within 3 SD either side of the mean.

A term that sometimes appears in the medical literature is the *coefficient of variation (CV)*, which is the ratio of the SD to the mean, usually expressed as a percentage. The CV is often used to summarize the precision of a laboratory value: the smaller the coefficient of variation the more precise is the measurement, relative to its mean value. (Note, however, that expressing variability as a ratio of the mean value may not always be the best approach; see later in this series.)

Related to all these definitions, it is important to remember that precision is only one component of *accuracy* in any measurement. The other component is *bias*, which means any systematic difference between a set of measured values and the underlying true values they are supposed to represent. A measurement is accurate only when it is both free from bias and precise. Statistics can help with imprecision, since a larger sample size will reduce the error around an estimate that we derive from the data, and statistical theory shows how to estimate the level of (im)precision of our estimates (see next article in this series). In contrast, an increase in sample size has no effect on a biased measurement, and there is usually no information in the data themselves about the extent to which bias may affect our measurements.

In the next article in the series we will discuss the basic concepts of statistical reasoning or *inference*, which are used to draw appropriate conclusions from data, using the idea of population and sample and the tools of probability.

### REFERENCES

- 1 Epi-Info. Center for Disease Control and Prevention, Atlanta, GA. <http://www.cdc.gov/epo/epi/epiinfo.htm>.
- 2 Tufte ER. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, 1983.
- 3 Cleveland WS. *The Elements of Graphing Data*, 2nd edn. Hobart Press, Summit, New Jersey, 1994.
- 4 Reed GF, Meade BD, Steinhoff MC. The reverse-cumulative distribution plot: A graphic method for exploratory analysis of antibody data. *Pediatrics* 1995; **96**: 600-3.
- 5 Chambers JM, Cleveland WS, Kleiner B, Tukey PA. *Graphical Methods for Data Analysis*. Chapman and Hall, New York, 1983.