Singapore Med J 2003 Vol 44(6) : 280-285

Biostatistics 101: Data Presentation

Y H Chan

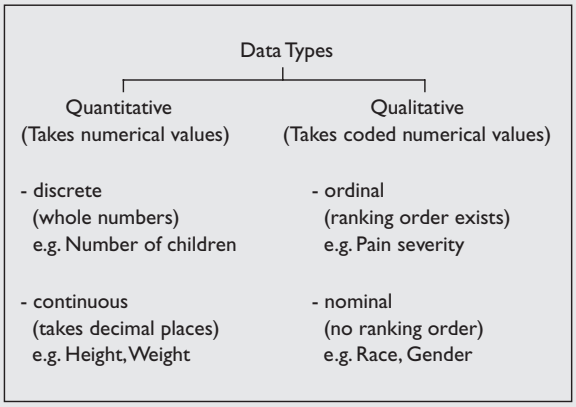
INTRODUCTION

Now we are at the last stage of the research process[1]: Statistical Analysis & Reporting. In this article, we will discuss how to present the collected data and the forthcoming write-ups will highlight on the appropriate statistical tests to be applied.

The terms **Sample & Population; Parameter & Statistic; Descriptive & Inferential Statistics; Random variables; Sampling Distribution of the Mean; Central Limit Theorem** could be read-up from the references indicated[2-11].

To be able to correctly present descriptive (and inferential) statistics, we have to understand the two data types (see Fig. 1) that are usually encountered in any research study.

Fig. 1



There are many statistical software programs available for analysis (SPSS, SAS, S-plus, STATA, etc). SPSS 11.0 was used to generate the descriptive tables and charts presented in this article.

It is of utmost importance that data "cleaning" needed to be carried out before analysis. For quantitative variables, out-of-range numbers needed to be weeded out. For qualitative variables, it is recommended to use numerical-codes to represent the groups; eg. 1 = male and 2 = female, this will also simplify the data entry process. The "danger" of using string/text is that a small "male" is different from a big "Male", see Table I.

Table I. Using Strings/Text for Categorical variables.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	38	50.0	50.0	50.0
	male	13	17.1	17.1	67.1
	Male	25	32.9	32.9	100.0
	Total	76	100.0	100.0	

Researchers are encouraged to discuss the database set-up with a biostatistician before data entry, so that data analysis could proceed without much anguish (more for the biostatistician!). One common mistake is the systolic/diastolic blood pressure being entered as 120/80 which should be entered as two separate variables.

To do this data cleaning, we generate frequency tables *(In SPSS: Analyse – Descriptive Statistics – Frequencies)* and inspect that there are no strange values (see Table II).

Table II. Height of subjects.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.30	20	26.3	26.3	26.3
	1.40	14	18.4	18.4	44.7
	1.50	28	36.8	36.8	81.6
	1.60	10	13.2	13.2	94.7
	1.70	3	3.9	3.9	98.7
	3.70	1	1.3	1.3	100.0
	Total	76	100.0	100.0	

Someone is 3.7 m tall! Note that it is not possible to check the "correctness" of values like subject number 113 (take note, all subjects must be key-coded; subjects' name, i/c no, address, phone number should not be in the dataset; the researcher should keep a separate record – for his/her eyes only) is actually 1.5 m in height (but data entered as 1.6 m) using statistics. This could only be carried out manually by checking with the data on the clinical record forms (CRFs).

Clinical trials and Epidemiology Research Unit
226 Outram Road
Blk A #02-02
Singapore 169039

Y H Chan, PhD
Head of Biostatistics

Correspondence to:
Y H Chan
Tel: (65) 6317 2121
Fax: (65) 6317 2122
Email: chanyh@cteru.gov.sg

DESCRIPTIVE STATISTICS

Statistics are used to summarise a large set of data by a few meaningful numbers. We know that it is not possible to study the whole population (cost and time constraints), thus a sample (large enough[12]) is drawn. How do we "describe" the population from the sample data? We shall discuss only the descriptive statistics and graphs which are commonly presented in medical research.

Quantitative variables

Measures of Central Tendency

A simple **point-estimate** for the **population mean** is the **sample mean**, which is just the average of the data collected.

A second measure is the **sample median**, which is the ranked value that lies in the middle of the data. E.g. 3, 13, 20, 22, 25: median = 20; e.g. 3, 13, 13, 20, 22, 25: median = (13 + 20)/2 = 16.5. It is the **point** that divides a distribution of scores into two equal halves.

The last measure is the **mode**, which is the most frequent occurring number. E.g. 3, 13, 13, 20, 22, 25: mode = 13. It is usually more informative to quote the mode accompanied by the percentage of times it happened; e.g, the mode is 13 with 33% of the occurrences.

In medical research, mean and median are usually presented. Which measure of central tendency should we use? Fig. 2 shows the three types of distribution for quantitative data.

It is obvious that if the distribution is normal, the mean will be the measure to be presented, otherwise the median should be more appropriate.

How do we check for normality?

It is important that we check the normality of the quantitative outcome variable as to allow us not only to present the appropriate descriptive statistics but also to apply the correct statistical tests. There are three ways to do this, namely, graphs, descriptive statistics using skewness and kurtosis and formal statistical tests. We shall use three datasets (right skew, normal and left skew) on the ages of 76 subjects to illustrate.

Graphs

Histograms and Q-Q plots

The histogram is the easiest way to observe non-normality, i.e. if the shape is definitely skewed, we can confirm non-normality instantly (see Fig. 3). One command for generating histograms from SPSS is *Graphs – Histogram (other ways are, via Frequencies or Explore)*.

Another graphical aid to help us to decide normality is the Q-Q plot. Once again, it is easier to spot non-normality. In SPSS, use *Explore or Graphs – QQ plots* to produce the plot. This plot compares the quantiles of a data distribution with the quantiles of a standardised theoretical distribution from a specified family of distributions (in this case, the normal distribution). If the distributional shapes differ, then the points will

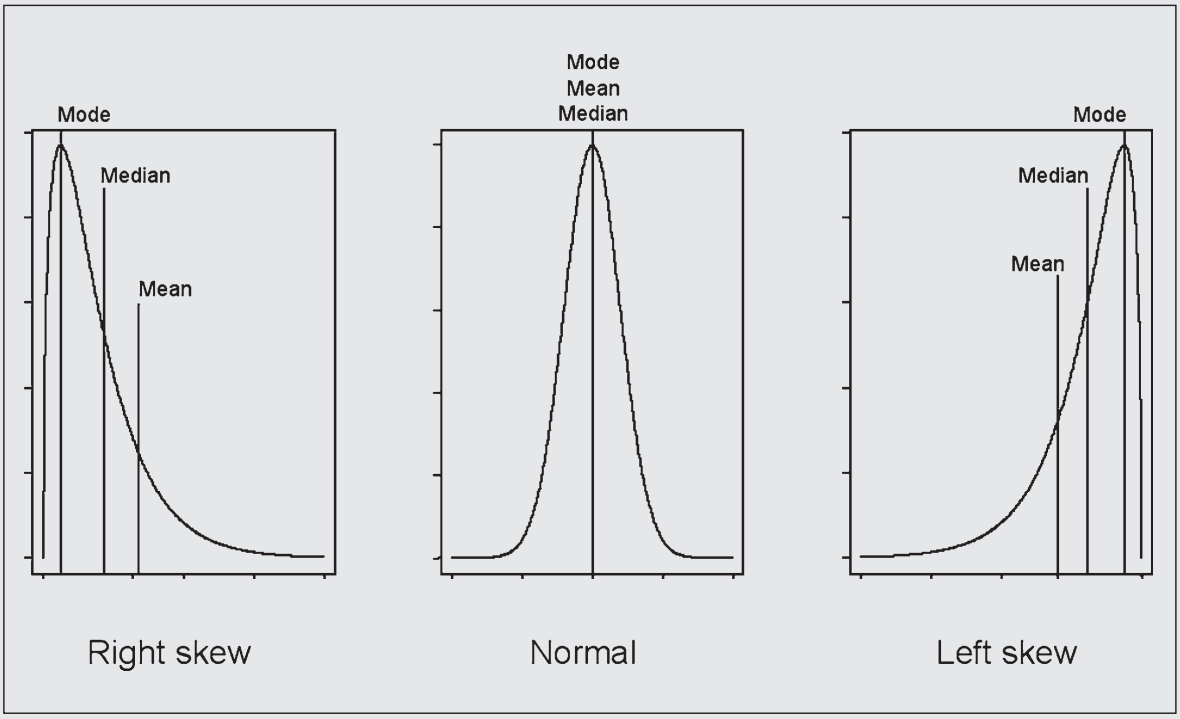


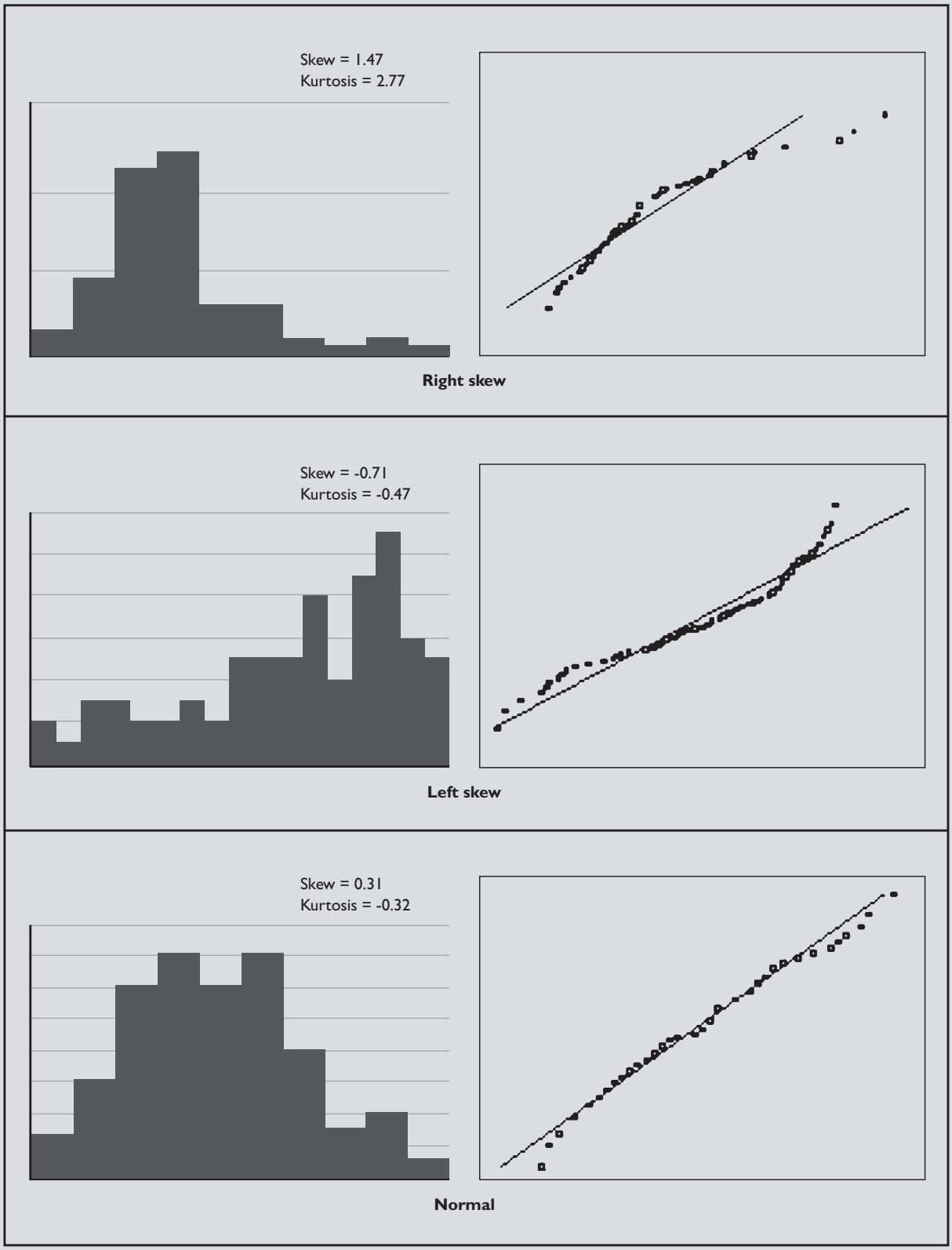
Fig. 2 Distributions of Quantitative Data.

plot along a curve instead of a line. Take note that the interest here is the central portion of the line, severe deviations means non-normality. Deviations at the "ends" of the curve signifies the existence of outliers. Fig. 3 shows the histograms and their corresponding Q-Q plots of the three datasets.

Descriptive statistics using skewness and kurtosis

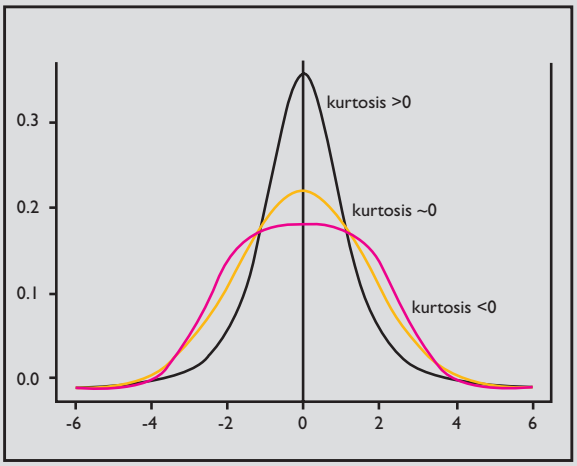
Fig. 3 shows the three types of skewness (right: skew >0, normal: skew ~0 and left: skew <0). Skewness ranges from -3 to 3. Acceptable range for normality is skewness lying between -1 to 1. Normality should not be based on skewness

Fig. 3



alone; the kurtosis measures the "peakness" of the bell-curve (see Fig. 4). Likewise, acceptable range for normality is kurtosis lying between -1 to 1. The corresponding skewness and kurtosis values for the three illustrative datasets are shown in Fig. 3.

Fig. 4



Formal statistical tests – Komolgorov Smirnov one Sample test and Shapiro Wilk test

Here the null hypothesis is: Data is normal

Table III. Normality tests.

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Right Skew	0.187	76	0.000	0.884	76	0.000
Normal	0.079	76	0.200	0.981	76	0.325
Left skew	0.117	76	0.012	0.927	76	0.000

From the p-values (sig), see Table III, both Right skew and Left skew are not normal (as expected!). To test for normality, in SPSS, use the Explore command (this will also generate the QQ plot). One caution in using the formal test is that these tests are very sensitive to the sample sizes of the data.

For small samples (n<20, say), the likelihood of getting p<0.05 is low and for large samples (n>100), a slight deviation from normality will result in the rejection of the null hypothesis! Urghh... I know this is so confusing! So, normal or not? Perhaps, Table IV will give us some light in our checking for normality. Take note that the sample sizes suggested are only guidelines.

Table IV. Flowchart for normality checking.

1. Small samples* (n<30): always assume not normal.
2. Moderate samples (30-100).

   If formal test is significant, accept non-normality otherwise double-check using graphs, skewness and kurtosis to confirm normality.
3. Large samples (n>100).

   If formal test is not significant, accept normality otherwise Double-check using graphs, skewness and kurtosis to confirm non-normality.

* Reminder: not ethical to do small sized studies[12].

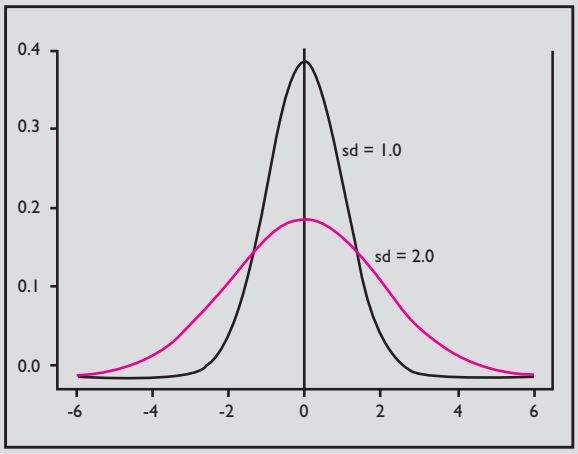
Measures of Spread

The measures of central tendency give us an indication of the typical score in a sample. Another important descriptive statistics to be presented for quantitative data is its variability – the spread of the data scores.

The simplest measure of variation is the **range** which is given by the difference between the maximum and minimum scores of the data. However, this does not tell us what's happening in between these scores.

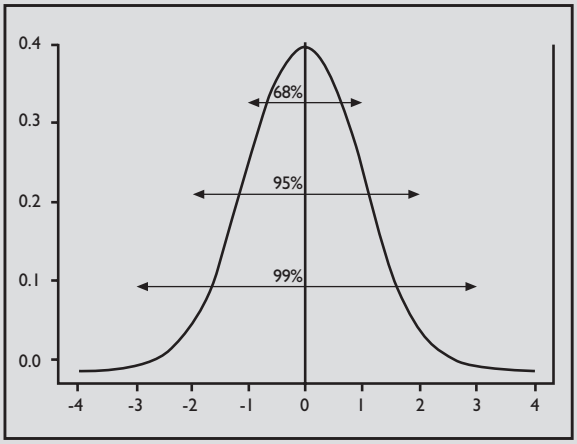
A popular and useful measure of spread is the **standard deviation (sd)** which tells us how much all the scores in a dataset cluster around the mean. Thus we would expect the sd of the age distribution of a primary one class of pupils to be zero (or at least a small number). A large sd is indicative of a more varied data scores. Fig. 5 shows the spread of two distributions with the same mean.

Fig. 5 Measures of Spread: standard deviations.



For a normal distribution, the mean coupled with the sd should be presented. Fig. 6 gives us an indication of the percentage of data "covered" within one, two and three standard deviations respectively.

Fig. 6 Distribution of data for a normal curve.



Here comes the million dollar question? Does a small sd imply good research data? I believe most of you (at least 90%) would say yes! Well, partly you are right – it depends.

For the age distribution of the subjects enrolled in your research study, you would not want the sd to be small as this will imply that your results obtained could not be generalised to a larger age-range group. On the other hand, you would hope that the sd of the difference in outcome response between two treatments (active vs control) to be small. This shows the consistency of the superiority of the active over the control (hopefully in the right direction!).

Interval Estimates (Confidence Interval)

The accuracy of the above point estimates is dependent on the sampling plan of the study (the assumption that a representative sample is obtained). Definitely if we are allowed to repeat a study (with fixed sample size) many times, the mean and sd obtained for each study may be different, and from the theory of the Sampling Distribution of the Mean, the mean of all the means of the repeated samples will give us a more precise point estimate for the population mean.

In medical research, we do not have this luxury of doing repeated studies (ethical and budget constraints), but from the Central Limit Theorem, with a sample large enough[12], an **interval estimate** provides us a range of scores within which we are confident, usually a 95% Confidence Interval (CI), that the population mean lies within.

Using the Explore command in SPSS, the CI at any percentage could be easily obtained. For a simple (large sample) 90% or 95% CI calculation for the population mean, use

$$\text{sample mean} \pm c * \text{sem}$$

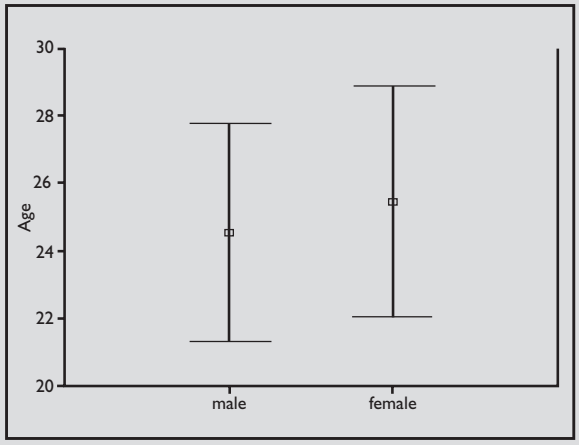
where c = 1.645 or 1.96 for 90% or 95% CI respectively and sem (standard error of the mean) = $\text{sd}/\sqrt{n}$ (where n is the sample size).

For example, the mean difference in BP reduction between an active treatment and control is 7.5 (95% CI 1.5 to 13.5) mmHg. It looks like the active is "fantastic" with a 7.5 mmHg reduction but from the large confidence interval of 12 (= 13.5 - 1.5), it could possibly be that the study was conducted with a small sample size or the variation of the difference was large. Thus from the CI, we are able to assess the quality of the results.

When should the usual 95% CI be presented. Surely for treatment differences, it should be specified. How about variables like age? There's no need for age in demographics but if we are presenting the age of risk of having a disease, for example, then a 95% CI would make sense.

The error bar plot is a convenient way to show the CI, see Fig. 7.

Fig. 7 Error bar plot.



Qualitative variables

For categorical variables, frequency tables would suffice. For ordinal variables, the "correct order" of coding should be used (for example: no pain = 0, mild pain = 1, etc). Graphical presentations will be bar or pie charts (will not show any examples as these plots are familiar to all of us).

CONCLUSIONS

The above discussion on the presentation of data is by no means exhaustive. Further readings[2-11] are encouraged. A recommended "Table for demographic" in an article for journal publication is

	Group A	Group B	p-value
Quantitative variable (e.g. age)			
Mean (sd)			
Range			
Median			
Qualitative variable (e.g. sex)			
Male	n_1 (%)	n_2 (%)	
Female	n_3 (%)	n_4 (%)	

We shall discuss the statistical analysis of quantitative data in our next issue (Biostatistics 102: Quantitative Data – Parametric and Non-Parametric tests).

REFERENCES

1. Chan YH. Randomised controlled trials (RCTs) — Essentials, Singapore Medical Journal 2003; Vol 44(2):60-3.
2. Beth Dawson-Saunders, Trapp RG. Basic and clinical biostatistics. Prentice Hall International Inc, 1990.
3. Bowers D, John Wiley and Sons. Statistics from scratch for Health Care professionals, 1997.
4. Bowers D, John Wiley and Sons. Statistics further from scratch for Health Care professionals, 1997.
5. Pagano M and Gauvreau K. Principles of biostatistics, Duxbury Press. Wadsworth Publishing Company, 1993.
6. Lloyd D Fisher, Gerald Van Belle, Biostatistics. A methodology for the health sciences. John Wiley & Sons, 1993.
7. Campbell MJ, Machin D. Medical statistics —A commonsense approach. John Wiley & Sons, 1999.
8. Bland M. An introduction to medical statistics. Oxford University Press, 1995.
9. Armitage P and Berry G. Statistical methods in medical research. 3rd edition, Blackwell Science, 1994.
10. Altman DG. Practical statistics for medical research. Chapman and Hall, 1991.
11. Larry Gonick and Woollcott Smith. The cartoon guide to statistics. HarperCollins Publishers, Inc. 1993.
12. Chan YH. Randomised controlled trials (RCTs) — Sample size: the magic number? Singapore Medical Journal 2003; Vol 44 (4):172-4.