Elementary statistics in medicine

B. K. Dassanayake, MBBS (Peradeniya)

Department of Surgery, Faculty of Medicine, University of Peradeniya.

Key words: Statistical software; Medical statistics.

Introduction

This series aims to explain the basic statistical methods commonly used in medical and surgical research. A working knowledge of statistics would require solving three problems:

a. Which statistic to use on data, and where
b. A basic knowledge of the principles behind the statistics used
c. How to use statistical software to get this analysis done

The first component - which statistic to use - is the stumbling block which generally limits a medical professional's knowledge of the other two issues. The content will thus focus primarily on which statistics to use in data analysis, and will comprise five sections dealing in order with the following areas:

1. The history of medical statistics, types of data, types of statistics
2. Descriptive statistics
3. Inferential statistics-1: sample size, testing a hypothesis, comparing categorical data groups
4. Inferential statistics-2: comparing ratio/interval data, correlation, regression, ANOVA
5. Summary of inferential statistics, multifactorial analysis, survival data

The history of medical statistics

Although mathematical methods have been used sporadically - sometimes to great effect - in Western Medicine for over 3 centuries, the advent of 'Medical Statistics' is a 20th century concept.

In 1747, a Scottish Physician named James Lind picked twelve sailors with scurvy and fed two each with various mixtures including citrus fruit, sulphuric acid and sea water. Those fed with citrus recovered, establishing a cure for scurvy. Such directness would have modern medical ethics committees up in arms, but this is the first recorded use of mathematical method employed in modern medicine. This system of using numbers to prove a point was named 'The Numerical Method' by Pierre CA Louis (1835). The Numerical Method was put to great use by John Snow - the father of Epidemiology - during the Great London Cholera Outbreak of 1854. Using a simple street map as a frequency table, Snow was able to trace the source of infection to a single contaminated water pump.

'Real' progress in Medical Statistics as a field, however, can be attributed to a series of seminal papers titled Principles of Medical Statistics published in The Lancet by Austin Bradford Hill in 1937. The icing on the cake was arguably Professor Archibald Cochrane's Effectiveness and Efficiency: Random Reflections on Health Services (1972). This was the seed that resulted in modern Evidence Based Medicine, whose greatest proponent, The Cochrane Collaboration, is thus named in his honor.

What is statistics?

Laplace once noted that "all knowledge is uncertain, and therefore probabilistic in nature". Statistics may be viewed as a method of hanging a number on that probability. The Oxford Dictionary of Statistical Terms defines statistics as the study of the collection, organisation, analysis, interpretation and presentation of Data. Therefore a knowledge of statistics first requires a knowledge of Data.

What are the Types of Data?

This is a critical concept - it is the Type of Data that dictates the type of statistic to be used.

Stevens (1946) organised data into 4 types:

1. Categorical/nominal data
2. Ordinal data
3. Interval data
4. Ratio data

Let us look at these data types in detail

1. Categorical/nominal data

This is data that can only be divided into groups. Each group will have a frequency, or number of observations, attached to it.

Eg: The site of breast cancer may be categorized as occurring in the four quadrants of the breast. There will be a fre-

Correspondence: Buddika. K. Dassanayake
thraless@gmail.com

quency of occurrence in each quadrant.

Note that there is no order between the categories.

2. Ordinal data

Ordinal data is more advanced because this type of data can be arranged according to a rank. This may include data with numerical values, or even categories which have can be arranged in an order.

Eg: Tumor differentiation on histology: poor, moderate, good

Intelligence Quotient (IQ) values

Visual Analog Pain Scale values

Glasgow Coma Scale values

Note that the first example (tumour differentiation) is different from Categorical Data since, the categories can be set in order. If analysed as categorical data regardless of rank, it will yield less information.

In the next three examples, the data has a numerical value. But comparing these values quantitatively is difficult. For example, a rise in the visual analog pain scores from 0 to 3 or is not the same as a rise from 7 to 10 (the maximum possible value). Comparing an IQ of 130 with one of 100 is not the same as comparing an IQ of 100 with one of 70. Therefore, in Ordinal Data, comparing the numerical intervals is meaningless. However, we can easily say that one value is higher than the other.

3. Interval data

Unlike Ordinal data, the intervals between values are meaningful in Interval data.

Eg:Celsius scale: 130°C is 30°higher than 100°C. This is the same difference as between 70°C and 100°C.

However, this kind of data is limited in analytic ability by a shortcoming - the measurement has no true zero. Thus, in interval data, ratios between values do not make sense. But intervals between values do.

Hence, for example, 100°C is not 'half as hot' as 200°C.

4. Ratio data

This is the most robust of data types, allowing many types of analysis to be performed. The value scale has a true zero, and hence, ratios between values make sense.

Eg: Kelvin temperature scale: 100°K is half as hot as 200°K.

Length/height: 100 cm is twice as long as 50 cm

Weight: a 30kg person weighs half as much as a 60 kg person

The first step in using statistics is therefore to establish which type of data you are dealing with.

Types of statistics

The second question is which type of statistics to use on your data. Generally, most research requires two types of statistics:

1. Descriptive statistics

Descriptive statistics is the simpler type, and simply 'describes' or 'summarises' the data. This employs measurements that can be directly verified from the data that we have collected. Any data we collect can be described or summarised using two measures:

a. Measures of central tendency: A single value which will give a good idea of all the data that has been collected. Eg: mean, median etc.
b. Measures of spread: A method of describing how close or far from the Measure of Central Tendency the actual data points are. Eg: range, variance, standard deviation

2. Inferential statistics

This attempts to take our data a step further than simply description. The main idea here is to use the data we have collected as a 'sample' and make a prediction using it regarding all such data (the 'Population').

Various kinds of predictions can be made:

A hypothesis can be tested (are 10 year old male children different in height from females?),

A value can be predicted for a population parameter (what is the average height of a ten year old male?).

We cannot reach any statistically valid conclusions based on Descriptive data alone. Hence most research attempts to first describe the data and then use inferential statistics to reach conclusions. The next section will deal with Descriptive statistics, and discuss which Descriptive Statistics to employ in each type of data.

References (also to be used as further reading)

1. Bradford Hill, A. (1971). Principles of medical statistics. The Lancet Ltd.
2. Cochrane, A. L. (1972). Effectiveness and efficiency: random reflections on health services (Vol. 1971). London: Nuffield Provincial Hospitals Trust.
3. Thattil, R.O. and L.H.P. Gunaratne, (2000). Basic Statistics for Advanced Level Students. Godage & Bros. Colombo, Sri Lanka.
4. Dodge, Y. (2003) The Oxford Dictionary of Statistical Terms, Oxford University Press.
5. Stevens, S. S. "On the Theory of Scales of Measurement." Science, New Series, Vol. 103, No. 2684 (Jun. 7, 1946), pp. 677-680