1

IN BRIEF
- An understanding of the jargon and underlying concepts of statistics
- A description of sources of bias
- The interpretation of a confidence interval
- An outline of the statistical approach to hypothesis testing

Further statistics in dentistry
Part 1: Research designs 1

A. Petrie[1] J. S. Bulman[2] and J. F. Osborn[3]

This new series of articles is designed to supplement, rather than replace, the material contained in the earlier series on statistics in dentistry published in the British Dental Journal and subsequently made available in book form.[1] With the increasing availability of calculators, personal computers and computer statistical software packages, the need for details of the 'nuts and bolts' of statistical theory has diminished, but the need for an understanding of how and when such theory should be applied, and how studies should be designed to make best use of it, has probably increased. It is with these thoughts in mind that this new series is being offered.

[1]Senior Lecturer in Statistics, Eastman Dental Institute for Oral Health Care Sciences, University College London; [2]Honorary Reader in Dental Public Health, Eastman Dental Institute for Oral Health Care Sciences, University College London; [3]Professor of Epidemiological Methods, University of Rome, La Sapienza Correspondence to: Aviva Petrie, Senior Lecturer in Statistics, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD E-mail: a.petrie@eastman.ucl.ac.uk

DEFINITIONS

Inevitably every specialist subject has its own vocabulary, where often everyday words take on a special meaning. These new meanings are usually designed to put across a specific and frequently used concept, and therefore it is useful for the novice to be aware of and understand them. Unfortunately, statisticians, like specialists in other disciplines, have not avoided the use of jargon, or words apparently invented to make a simple idea incomprehensible to the outsider. Here, then, are some of the terms and ideas which are employed regularly in any statistical investigation.

VARIABLES, FACTORS AND EFFECTS

A **variable** is a quantity that can take different values for individuals in a study; it is called a **response variable** when it relates to the outcome of interest. A variable may be *quantitative,* that is, it takes a *numerical* value; or it may be *qualitative* or *categorical,* in which case it describes an attribute. For example, changes in body weight after specific periods on a prescribed diet are quantitative because, for each subject, the observation takes a numerical value. In contrast, a child may or may not show evidence of dental caries at a particular moment in time. In this case the observation describes the presence or absence of a characteristic, and it is therefore qualitative rather than quantitative. When the values for the variable for several individual subjects are aggregated, quantitative observa-

tions may be summarised to yield some sort of average value, for example the arithmetic mean. Qualitative data, particularly if related to disease or death, may be summarised by a proportion or a rate, such as the prevalence or incidence rate. All these summary measures which represent characteristics of the observations in the population are called **parameters.**

Although it is possible to define the term **factor** in a number of ways, it most commonly refers to a quantity which is a possible cause of variation of the outcome of interest in an investigation. If the factor is a categorical variable, then the different categories are called **levels** of the factor. For example, in a clinical trial the factor 'treatment' with two levels would divide the group under investigation into two subgroups, such as those who receive an active treatment and those who do not. (Note however, in statistics, the word 'treatment' does not necessarily mean a treatment as given by a doctor or dentist to a patient. It is used in a more general sense to mean any experimental procedure applied to any experimental unit). The factor 'gender' divides the subjects into males and females; the factor 'age' can be used to divide the study group into different age categories.

The **effect** of a factor can be determined by estimating the appropriate parameter, such as the average or proportion, for each level of the factor and then comparing the resulting estimates. This is just a clumsy way of saying, for example, that social class (the factor) has an

Interaction

An interaction exists between two factors when the effect of one factor varies for the different levels of the other factor

Confounding

Two factors are confounded when some or all of the effect of one factor is actually due to the other factor

effect on caries incidence rate (the parameter) if the latter increases with decreasing social class (each social class specifies a level of the factor). As another example, the effect of the factor 'topical fluoride' may be different in the average annual caries increment in those subjects who receive it (one level of the factor) and in those who do not (a second level of the factor). For a particular study, the effect of the factor 'gender' may be that males have twice the risk of getting the disease than females. If the factor of interest is a health intervention (eg a new treatment), then the effect of the health intervention is of primary interest; ie it is necessary to compare the response variable in those 'exposed' to the intervention with those who are not. Those subjects who do not receive the intervention are often called a 'control' or 'comparison' group, which may nevertheless comprise those who are on a standard treatment (or a placebo). Such comparisons are sometimes made by examining for each individual the response variable, say the area of an oral lesion, before and after the exposure to the intervention and then comparing these differences between groups. A study that has no control group is, in general, an unsatisfactory way of investigating the effect of a factor and is often called an **uncontrolled case series.**

INTERACTION

Very often in a study there will be factors, other than just the principal factor of interest, which have an effect on the response. Suppose a study is designed to investigate the effect of a particular health intervention. The study population is divided into two groups, those who are exposed to the intervention and those who are not; ie the factor 'intervention' has two levels. Suppose also that the study includes both males and females so that the factor 'gender' also has two levels. An **interaction** between the factors exists if the effect of the intervention in males differs from the effect of the intervention in females. Interaction is sometimes called 'effect modification' because the effect of one factor is modified or changed by the level of the other. If an important interaction exists, it does not make much sense to speak of the effect of the factor of interest without specifying the level of the other. An extreme example would be an intervention which is beneficial to one half of the population but harmful to the other half. It would not be very useful to report that on average the intervention had no effect. A more realistic example might be a study specifically designed to determine those sectors of a population who might gain greatest benefit, thus enabling appropriate targeting of the intervention.

CONFOUNDING

In 1948, the year the National Health Service was introduced in the United Kingdom, the crude death rate was 11.0 per 1,000 population in England and Wales. This implies that on average about eleven people died per annum for every thousand people in the population. Since

then, the rate has increased slowly and erratically, and for example in 1986 it was 11.6 per 1,000. In other words, the rate of dying seems to have increased since the introduction of the NHS. But the observed effect of the introduction of the National Health Service is distorted if there exists some other factor, for example age, which is associated with the death rate, and which is distributed unequally in the years being compared. Since the proportion of the population who are elderly increased during the 38 years between 1948 and 1986, it is not surprising that the crude death rate did not continue to decline from its minimum ever value of 11.0 in 1948. So although the apparent effect of the introduction of the National Health Service has been an increase in the crude death rate, this apparent effect has been distorted by the factor 'age', and the effect of the intervention is said to be **confounded** or confused by 'age'. Using the same reasoning, just because over a period of time per capita sugar consumption in a country decreased with a parallel decrease in dental caries, it should not be assumed that the reduced sugar consumption caused the decrease in caries. There might, for example, have been increased availability of fluoride dentifrices over the same period, and this could be the confounding variable which was the main reason for the fall in caries.

BIAS

The considerations which govern a well designed study rely, to a large extent, on ensuring that the study is free from bias. **Bias** is present when the results from the study are systematically distorted and so are consistently above (or below) what they should be. Biases may arise in a number of ways. Typical examples are:

· *Observer bias* – when one observer consistently over-reports (or under-reports) a variable. This may be resolved by training and calibration
· *Selection bias* – when the individuals in the study are not representative of the population of interest. This may be avoided by ensuring that a *random* method of selection is used rather than relying on purposive or judgement sampling, where investigators include in their samples those individuals who they believe are typical or representative of the population
· *Publication bias* – which is a tendency for journals to publish only 'significant' results
· *Recall* bias – when certain patients have a differential ability to remember details about their past
· *Allocation bias* – when treatment groups in an experimental study are not comparable with respect to the variables influencing the response of interest. Random allocation (also called *randomisation)* of the treatments to the patients is a way of avoiding this bias and is discussed in detail in the first paper on Clinical Trials which will appear later in this series
· *Assessment bias* – resulting from the manner

in which the responses to treatment are assessed, often because of the subjective nature of those responses and the preconceived notions of those assessing the response. Making the trial 'blind' so that the study personnel and perhaps the patients are unaware of which treatment each patient receives is a way of overcoming assessment bias. This is discussed in more detail in Part 3 of this series – Clinical Trials 1.

STATISTICAL INFERENCE

In many situations, investigators are interested in the effect of some 'intervention' or 'treatment' in a particular target population. For example, the question of interest might be, 'What effect would the introduction of water fluoridated at 1 ppm have on the prevalence of dental caries in Salisbury?' To answer this question, exactly as phrased, would involve studying the whole population of Salisbury before and after the introduction of fluoridated water. In practice, the investigators might decide to study the effect of the intervention on a small scale by looking at a representative sample of the population, rather than the whole population. If the effect of the intervention were beneficial in the small scale sample study, an **inference** might be made that the effect of the intervention would be beneficial in the whole population. Clearly such an inference might not be exactly correct; the effect in the whole population may be greater or less than the effect in the sample. Indeed, the observed effect in the sample will depend on precisely which sample is used for the study. Different samples will give slightly different estimates of the effect of the intervention in the population, although common sense will suggest that the larger the sample, the more likely will the effect approximate closely to the effect in the population. Thus any particular sample study is likely to have some degree of **sampling error** if the effect of the intervention in the sample is used to estimate the effect of the intervention in the whole population. Statistical inference is, to a large extent, concerned with **estimating the magnitude of an effect** in a population and assessing the sampling error of the estimate, using evidence obtained from a representative sample drawn from that population. The second major component of statistical inference involves *assessing the effects of interest,* achieved by performing **significance tests,** also called **hypothesis tests.**

Estimation

In estimation, the sample study provides an estimate of the effect of the intervention in the population and consideration of sampling error yields an interval, known as a **confidence interval,** which is reasonably certain to contain the (unknown) true population effect. For example, suppose a sample study implied that fluoridated drinking water in a given area would halve exactly the incidence of dental caries in 12-year-old children. If this sample result were

extended to the population there would clearly be some doubt as to whether the incidence of caries would be halved exactly. However, statistical analysis of the sample data might lead the investigators to be reasonably certain that the effect of fluoridated water would be to reduce the incidence in the population to between 40% and 60% of the pre-intervention level. This interval, which is believed to include the true effect of the intervention in the population, is called a confidence interval, and the degree of belief is measured by a probability expressed as a percentage, usually 95%. Thus a 95% confidence interval for the effect of an intervention is commonly interpreted to mean that there is a 95% chance that the interval will include the true (but unknown) population effect of the intervention.
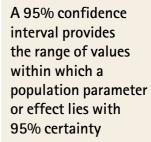
Hypothesis tests

A hypothesis test is concerned with answering a simple question about the effect of interest in the population, such as 'will a given intervention have an effect in a population?'. This is resolved in statistical terms by determining whether the evidence in the sample suggests that the statement 'the intervention has *no* effect in the population' is likely to be *false.* The statement is called the **null hypothesis**. So, rather than attempting to answer the original question and show that it is likely that the intervention has an effect, the statistical approach is to determine whether there is enough evidence in the sample to reject the null hypothesis that the intervention has no effect in the population. There are two possible conclusions that can be reached after the significance test has been performed. Either:

1. 'On the basis of the sample, there is good evidence to suggest that the intervention will have some effect in the population', or
2. The sample has not provided enough information to safely conclude that the statement 'the intervention has no effect in the population' is false'.

In the language of statistics, if the sample result implies that there would be an effect of the intervention in the population, the effect is said to be **statistically significant.** However, if the sample study does not provide sufficient information to reject the statement, the effect is said to 'Fail to reach statistical significance'. An effect that is statistically significant is unlikely to have arisen merely because of sampling error. Another way of explaining statistical significance is to say that there is a very small chance or *probability* of obtaining the observed results, or even more extreme results, *if the null hypothesis is true.* This probability is called the **P-value** of the test so that statistical significance is achieved if this P-value is small, typically less than the (arbitrary) cut-off value 0.05, called the **significance level.** The effect is then said to be significant at the 5% level.

Confidence interval

A 95% confidence interval provides the range of values within which a population parameter or effect lies with 95% certainty

P-value

The P-value is the probability of obtaining the sample results if the null hypothesis, that there is 'no effect' in the population, is true

Statistical significance

An effect is statistically significant if the null hypothesis is rejected, usually if $P < 0.05$

It is important to distinguish between the use of the word 'significant' in statistics with its use in everyday English. Colloquially, the word 'significant' is synonymous with 'important', perhaps in some clinical or public health sense. In statistical parlance, 'significant' means 'unlikely to be caused by sampling error.' Notice that a statistically significant effect could be small and of very little clinical importance. Conversely, an observed effect may be found to be not statistically significant at the 5% level because of inadequate sample size, even though, if the intervention were applied to the population, it might have a very important clinical effect. In this context it is crucial to note that a finding of 'not significant' does not 'prove' that the intervention will have no effect in the population; it merely indicates that there is insufficient evidence to say that it will have an effect. For this and other reasons, although until relatively recently statistical results in published medical and dental papers tended to report in terms of significance levels (eg $0.01 < P < 0.05$ or $0.001 < P < 0.01$), the modern trend is to quote exact P-values (eg $P = 0.03$ or $P = 0.007$) and concentrate on 95% confidence intervals. Many leading journals now will not accept papers which do not follow this trend.

The next paper in this series will take a further look at types of study design for statistical analysis.

1   Bulman J S, Osborn J F. *Statistics in Dentistry.* London: British Dental Association, 1989.