STATISTICS REVISITED: A REVIEW FOR CONTRIBUTORS AND READERS

From samples to populations: Estimation and hypothesis testing

Jill Mollison, BSc, CStat, *Lecturer in Medical Statistics, Department of Public Health, University of Aberdeen, Aberdeen, UK;* **Julie A Simpson**, PhD, CStat, *Biostatistician, Department of General Practice and Primary Care, University of Aberdeen, Aberdeen, UK;* **Phil C Hannaford**, MD, FRCGP, *Grampian Health Board Chair of Primary Care, Department of General Practice and Primary Care, University of Aberdeen, Aberdeen, UK*

Correspondence: *Jill Mollison, Department of Public Health, University of Aberdeen, Medical School, Powarth Building, Foresterhill, Aberdeen AB25 2ZD, UK. E-mail: j.mollison@abdn.ac.uk*

(Accepted 31st January 2002)

The Journal of Family Planning and Reproductive Health Care 2002: **28(2)**: 101-104

Introduction

We previously described how data from family planning research can be presented and summarised.[1] In this paper we will describe how to use data obtained from a *sample* to answer questions about the *population*. For example, we might use data from a survey to estimate how many women purchase hormonal emergency contraception over-the-counter from pharmacies. Or we might seek to answer the question, does taking the oral contraceptive pill (OCP) increase the risk of cervical cancer?

From sample to population

In order to make inferences about a particular population based on data obtained from a sample of individuals, the sample must be representative of the population. There is no statistical test for assessing whether a sample is representative of the population; instead we need to consider the way in which the sample was selected and the response rate. A non-random sampling method[2] or a low response rate could result in the sample being non-representative. If we believe that the sample is representative, then values relating to the sample (e.g. mean, proportion) provide the 'best guess' of what the corresponding population value might be. Indeed the true population value is unknown and can only be estimated using data obtained from a sample. Sample values, however, are unlikely to exactly match population values. In other words, the sample values may be less than, or greater than, the true population value. This uncertainty about the sample value is quantified by the *standard error*. The key when making inferences about a population on the basis of a sample is to get some idea of how closely the sample is likely to relate to the population. There are two approaches: estimation and hypothesis testing.

Estimation

Estimation involves calculating an interval around the sample value (e.g. mean), using data from the sample. This interval provides information about the uncertainty with which the sample value rightly represents the corresponding population value. If a sample is large, the sample is likely to be close to the population value and so the interval will be narrow. Conversely, if the sample includes data from only a small number of subjects we will be more uncertain that the sample value lies close to the population value and the interval will be wide. The interval is called the *confidence interval* since it represents, for a given degree of confidence, the range of values within which the true population value is likely to lie. Confidence intervals, therefore, not only tell us about the likely population values, they also provide, at a glance, some information about the amount of data upon which we are making our inference. Sample values with narrow confidence intervals are sometimes referred to as more precise point estimates, and those with wider confidence intervals less precise point estimates. Different levels of confidence can be set to reflect the degree to which the interval is likely to include the actual population value. Commonly, 95% confidence intervals are used. These give the range of values in which the true population value is likely to lie on 95% of occasions. In other words, the true value will lie outside the 95% confidence interval on 5% of occasions. If we want to be more certain of including the true population value, we tend to calculate 99% confidence intervals. The formula for a 95% confidence interval is: sample value (e.g. mean) ± 1.96 x standard error of sample value.

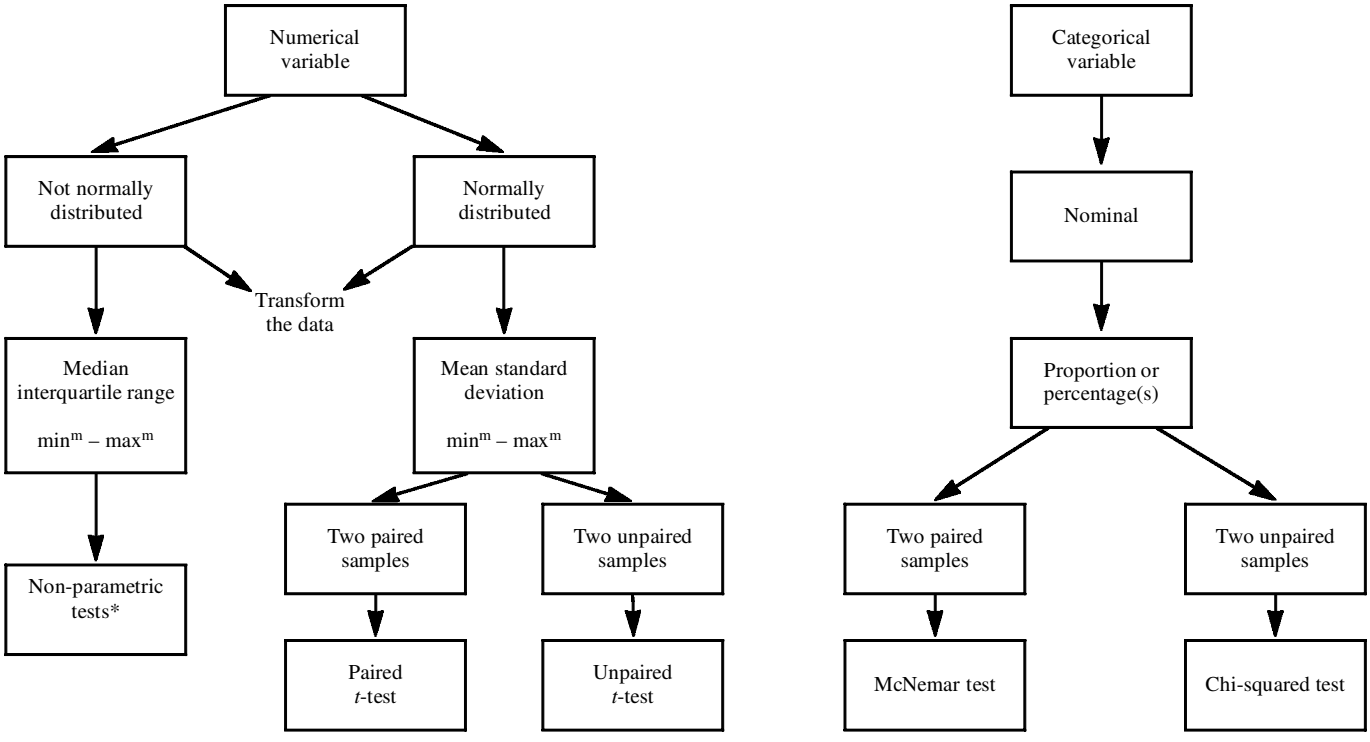
Confidence intervals can be computed for sample values such as the mean or proportion, as well as the difference between two means and the difference between two proportions. Many statistical computing packages calculate confidence intervals. There is also a book and accompanying computer program that allows the straightforward calculation of confidence intervals.[3]

Hypothesis testing

A complementary approach when making inferences about a population on the basis of a sample is hypothesis testing. Hypothesis testing is where a statistical test is performed to assess whether the observed values could have arisen purely by chance. For example, when assessing the association between a risk factor (e.g. smoking) and disease (e.g. coronary heart disease), we need to know whether the association observed in the sample is likely to be real (i.e. exists in the population) or whether it has occurred purely by chance.

The type of data available and the study design will dictate which statistical test is appropriate (Figure 1). The main considerations are whether the variable being compared is categorical (nominal/ordinal) or numerical (continuous/discrete) and whether the data are paired or unpaired. For example, a study to assess the effect of OCP use on heavy/painful periods would collect data from the same woman before and after commencement of OCP. The

Figure 1 *Choosing the appropriate statistical test*



*Non-parametric tests are discussed in more detail in Altman.[5]

data in this example are paired. Conversely, in a randomised controlled trial where patients are randomly allocated to treatment groups, the data obtained from the two randomised groups are not paired, i.e. they are independent. We will not give formulae for each of the tests, as there are numerous statistical packages available that can be used to perform the calculations.

In hypothesis testing we start with the assumption that there is 'no association' between exposure and disease and 'no difference' in outcome between treatments. This is known as the null hypothesis. The appropriate statistical test is then applied to the sample data, to quantify whether there is sufficient evidence to refute the null hypothesis. Each hypothesis test produces a probability (p value). In statistical terms, sufficient evidence to refute the null hypothesis is where a very small probability value (e.g. $p < 0.05$) is found, indicating that it is highly unlikely that the observed sample data arose purely by chance. We would therefore rule out chance as being a potential explanation for the results found; we would reject the null hypothesis and we would infer that there is an association in the population. Conversely, if the probability (p value) is large (e.g. $p > 0.05$), we would take this as insufficient evidence to reject the null hypothesis, i.e. we would infer there is no association in the population. By convention, a cut-off value of $p < 0.05$ is used for statistical significance, i.e. our rejection of the null hypothesis will be wrong on only 5% of occasions. However, if we wish to be more rigorous in our assessment, we can take a more stringent p value, such as $p < 0.01$, in which case the rejection of the null hypothesis would be wrong on only 1% of occasions.

The principles of estimation and hypothesis testing are illustrated below using a number of examples, all common to researchers in family planning and reproductive health. The data have been taken from a postal survey of women in the Royal College of General Practitioners' (RCGP) Oral Contraception Study conducted in 1994–1995,[4] although in some instances information from only a subset of women has been used.

Example 1: Estimating prevalence

Estimating the prevalence of disease or risk factor in a population is common in family planning research. Of the 10 073 responders to the 1994–1995 survey, 4519 (44.9%) reported that they had ever smoked. Assuming that the sample of women participating in the RCGP study were representative of all women in the UK, our figure of 44.9% should provide a good estimate of the level of smoking nationally in this group of women. However, our sample 'best guess' may underestimate or overestimate the true level of smoking in the population, purely because of sampling error. The sampling error will reduce as the size of the sample increases. In our example, the standard error was 0.5%, which resulted in a 95% confidence interval from 44.0% to 45.8%. In other words, we can be 95% confident that the true value of the proportion of smokers in the population is somewhere between these two values (i.e. it may be as low as 44% or as high as nearly 46%). The narrowness of the confidence interval also tells us that the results were based on a large amount of information.

Example 2: Comparison of two means (unpaired)

Is smoking status at recruitment related to age at diagnosis of coronary heart disease?

The research question involves assessing whether observed differences between groups in age at diagnosis indicates a real association or has occurred purely by chance. Expressed in terms of the null hypothesis, can we refute the hypothesis that there is no difference in age at diagnosis between smokers and non-smokers?

The first step in looking at the question is to summarise the age at diagnosis in the two groups (Table 1). Smokers had a younger mean age at diagnosis than non-smokers at recruitment, the difference being 1.5 years. A histogram

indicated that age at diagnosis was Normally distributed. Therefore the unpaired *t*-test (also referred to as Student's *t*-test) was applied (Figure 1). Had the data not been normally distributed we could have attempted to transform the data (see previous article)[1] or use a non-parametric test, such as the Mann-Whitney U test.[5] In our example, the p value from the unpaired *t*-test was p = 0.39. Thus, it is likely that our observed difference of 1.5 years occurred purely by chance and there is no evidence to reject the null hypothesis. In clinical terms, our study results do not suggest that there is an association between age at diagnosis and smoking status at recruitment.

Table 1 *Mean age at diagnosis of coronary heart disease by smoking status at recruitment*

	Smoker (n = 67)	Non-smoker (n = 36)	Statistical significance
Mean age at diagnosis (SD)	51.8 (8.10)	53.3 (8.95)	p = 0.39

Difference in mean age (95% confidence interval), 1.5 years (–1.9 to 5.0).
SD, Standard deviation.

Confidence intervals can be computed for the difference in mean age at diagnosis. Although our sample data indicated that smokers tended to be diagnosed with coronary heart disease 1.5 years younger than non-smokers, the 95% confidence interval for the difference in age at diagnosis was –1.9 to +5.0 years. Hence, the true difference in age at diagnosis in the population is likely to be between –1.9 years (i.e. 1.9 years greater in smokers than non-smokers) and 5.0 years (i.e. 5 years younger in smokers than non-smokers). This confidence interval includes zero and so it is plausible that in the population there is no difference in age at diagnosis between smokers and non-smokers. Both the unpaired *t*-test and the 95% confidence interval suggest that the observed difference could have arisen by chance (sometimes expressed as a 'statistically non-significant difference').

Example 3: Comparison of two means (paired)

Among women who have continued to smoke during the study, has there been a change in the quantity of cigarettes smoked?

This question looks at data for each woman at different points in time. Since the difference in the quantity of cigarettes smoked was normally distributed a paired *t*-test is appropriate (Figure 1). The null hypothesis is that there has been no change over time in the number of cigarettes smoked.

Table 2 shows descriptive data for the 1851 smokers in our study who had their smoking consumption recorded at recruitment and who provided details of current smoking consumption in the 1994–1995 survey. The paired *t*-test shows that the probability that the observed average increase of 2.3 cigarettes smoked per day occurred purely by chance was highly unlikely (p < 0.001). Therefore, there is sufficient evidence to reject the null hypothesis. In other words, our study suggests that among women who continue to smoke, there is a statistically significant increase in the number of cigarettes smoked per day.

It is possible to compute a confidence interval to quantify the uncertainty surrounding our sample estimate. The 95% confidence interval for the mean difference in the number of cigarettes is 2.0 to 2.6. Since the interval does not include zero, we can interpret this finding as being statistically significant.

Table 2 *Mean number of cigarettes smoked per day at recruitment and follow-up*

	Recruitment (n = 1851)	Follow-up (n = 1851)	Statistical significance
Mean number of cigarettes per day (SD)	13.3 (6.6)	15.6 (6.8)	p < 0.001

Difference in mean number of cigarettes (95% confidence interval), 2.3 (2.0 to 2.6).
SD, Standard deviation.

Example 4: Comparison of two proportions (unpaired)

Is there an association between smoking status and social class at recruitment?

Table 3 shows these categorical data presented as a cross-tabulation. Our sample data suggest that there is an association between smoking and social class (44.6% of women with a 'manual' social class were smokers at recruitment compared with 33.1% of 'non-manual' women). The data are unpaired and the variables are categorical, therefore a Chi-squared test is appropriate (Figure 1). The Chi-squared test starts with the assumption that the two variables are not associated (i.e. the null hypothesis exists). First, it calculates the expected number of women in each of the four groups, based on the relative size of social class and smoking groups in the sample. For example, assuming there is no association between smoking and social class, 2756.7 women of 'manual' social class would have been expected to be smokers. In fact, we found that there were 3008 in this group. The Chi-squared test then compares the observed and expected results. The greater the difference between the observed and expected values, the less likely it is that the association occurred purely by chance. In our example, the chances that the observed data came from a population in which there is no association between smoking and social class are very low since the p value was less than 0.001. The p value indicates that there is sufficient evidence to reject the null hypothesis and suggests that there is an association between smoking and social class.

Table 3 *Smoking status and social class at recruitment*

	Social class		Statistical significance
	Manual (n = 6748)	Non-manual (n = 3222)	
Smoker	3008 (44.6%)	1065 (33.1%)	p < 0.001
Non-smoker	3740 (55.4%)	2157 (66.9%)	

Difference in proportion of smokers (95% confidence interval), 11.5% (9.5% to 13.5%).

Using our alternative approach, we can calculate the 95% confidence interval around the difference in the proportion of smokers in manual and non-manual groups or, if preferred, the difference in proportion of non-smokers in the two social class groups. Looking at the difference in smokers (11.5%), the corresponding 95% confidence interval of 9.5% to 13.5% excludes zero, indicating a statistically significant result.

Example 5: Comparison of two proportions (paired)

Have women in our study changed their smoking habits over time?

Table 4 details the smoking status of women in our survey at two time points. To assess whether there has been a

change in smoking status, it is necessary to compare the proportion of women who have stopped smoking since recruitment (1703, 38.8%) with the proportion of women who have started smoking since recruitment (183, 4.2%). This corresponds to an overall reduction in the proportion of women smoking of 34.6%. The McNemar test examines whether this observed difference could have occurred by chance, again testing against the null hypothesis that in the population there has been no change in smoking habits. In our example the p value is less than 0.001, indicating that we have strong evidence to reject the null hypothesis. We can conclude that there has been a statistically significant reduction in smoking among women in our survey.

Table 4 *Smoking status at recruitment and follow-up*

	Smoker at follow-up	Non-smoker at follow-up	Statistical significance
Smoker at recruitment	1880 (42.9%)	1703 (38.8%)	$p < 0.001$
Non-smoker at recruitment	183 (4.2%)	618 (14.1%)	

Overall reduction in proportion of smokers over time (95% confidence interval), 34.6% (33.0% to 36.3%).

The 95% confidence interval for the observed reduction of 34.6% in the proportion of smokers is 33.0% to 36.3%, again indicating a statistically significant result. The width of the confidence interval is narrow due to the large sample size, increasing our certainty about the sample estimate.

General issues when presenting and interpreting the results of hypothesis testing

When reporting p values, the actual p value should always be presented. If the p value is greater than 0.05, indicating a statistically non-significant result, it is inappropriate to simply use 'NS' to reflect non-significant. Results of statistical tests reporting p values close to 0.05 (both above and below) are of borderline significance and should be interpreted cautiously.

Concluding comments

This article has described two approaches that are closely related to one another: estimation and hypothesis testing. Both are valid. However estimation (calculating confidence intervals) conveys more useful information than hypothesis testing (calculating p values). It is recommended that both p values and confidence intervals are presented when reporting results.[5]

Finally, it is essential to remember that statistical associations by themselves do not imply causation. There may be other explanations for an association, such as bias or confounding. It is also important to assess whether statistically significant associations and differences are large enough to be clinically important. The importance of clinical and statistical significance, and their relevance to sample size calculations, will be discussed in a future article.

Statements on funding and competing interests

Funding. None declared.

Competing interests. None declared.

References

1 Simpson JA, Mollison J, Hannaford PC. Summarising data. *J Fam Plann Reprod Health Care* 2001; **27:** 234–236.

2 Swinscow TDW. *Statistics at square one*.. 9th edition revised by MJ Campbell. London: British Medical Association, 1996.

3 Altman DG, Machin D, Bryant TN, Gardner MJ (eds). *Statistics with confidence*. London: British Medical Association, 2000.

4 Owen-Smith V, Hannaford PC, Warskyj M, et al. Effects of changes in smoking status on risk estimates for myocardial infarction among women recruited for the Royal College of General Practitioners' Oral Contraception Study in the UK. *J Epidemiol Community Health* 1998; **52:** 420–424.

5 Altman DG. *Practical statistics for medical research*. London: Chapman and Hall, 1991.

Glossary of terms

Population	- entire set of items of interest
Sample	- subset of a population
Standard error	- the uncertainty in the sample statistic
Confidence interval	- a range of values within which the population value is likely to lie