Hypothesis testing

A Sheikh and A Cook

Aziz Sheikh
NHS R&D National Primary Care Training Fellow

Adrian Cook
Statistician

Department of Primary Health Care and General Practice, Imperial College School of Medicine, Norfolk Place, London W2 1PG, UK

Correspondence to:
Dr A Sheikh
aziz.sheikh@ic.ac.uk

Date submitted: 14/03/00
Date accepted 03/05/00

Prim. Care Respir. J.
2000:9(1);16–17

INTRODUCTION

Previous papers in this series have discussed data description and the summary of results.[1,2] We now progress to statistical inference, showing how real effects can be discerned in the presence of random fluctuation. For example, clinical trials typically compare a new treatment with an existing treatment or placebo. If more patients recover in the new treatment group is this because the new intervention represents a real improvement or is the observed difference simply a chance finding? It is to questions such as this that significance tests provide valuable insight.

SIGNIFICANCE TESTS

Many different significance tests exist, each one developed for a particular type of data. Figure 1 shows the most commonly used tests and illustrates a suggested process of selection. The first step is to determine if one is dealing with continuous data or categorical, and then, whether the data is paired or unpaired. Paired data most typically arises when a single group of subjects is measured before and after some event, each observation in the 'before' group thus being paired with an observation in the 'after' group; this pairing needs to be recognised in the analysis. For continuous data, it is also necessary to examine variable distributions in order to select the most appropriate test. Calculations are usually performed using a computer statistical package, but may be performed manually (see Recommended Reading).

Continuous data

In a Tasmanian study, 753 non-asthmatic seven year olds had their forced expiratory volume in one second (FEV_1) measured.[3] Members of the cohort were then contacted at age 30 years and asked whether they were asthmatic. Table 1 shows that 81/753 classified themselves as asthmatic. Those reporting asthma had an average percent-predicted FEV_1 of 98 at age seven; the mean FEV_1 at age seven in those not reporting asthma was slightly higher (Table 1). Does the observed difference indicate an association between childhood lung function and adult onset asthma, or is this a chance finding?

Table 1: Lung function at age seven (percent predicted FEV_1), by adult asthmatic status

Adult asthma	n	mean (sd)
Yes	81	98.0 (13.5)
No	672	100.8 (13.0)

Table 2: Prevalence of hay fever, by type of heating

Heating	Hay fever, % (n)		
	Yes	No	Total
Central heating	7.8 (48)	92.2 (569)	617
Coal or wood	4.2 (28)	95.8 (634)	662
Total	76	1203	1279

We start by making the assumption that the observed difference between asthmatics and non-asthmatics occurred by chance; this is known as the 'null hypothesis'. The probability that the observed difference did indeed arise by chance is then calculated, three factors influencing this probability:

- A large difference is less likely to arise by chance than a small one
- A large difference is even less likely to arise by chance if the data do not vary much (low standard deviation)
- A large difference is more likely to arise by chance with a small number of observations, since even one spurious observation may have a large effect.

The resulting probability, known as the p-value, indicates the likelihood of the null hypothesis being true. A small value suggests the null hypothesis is not true and that a real difference exists. A p-value of 0.05 or less is widely regarded as strong evidence of a true difference.

In our example, the null hypothesis is of no association between childhood lung function and adult onset asthma. The data are continuous, observations are not paired and the data in each group are normally distributed, so from Figure 1 an unpaired t-test is most appropriate. The unpaired t-test uses the three factors noted above: size of difference, variability of data and sample size. The result is $p = 0.08$, an eight per cent likelihood of the observed difference being due to chance. We would therefore conclude that only weak evidence exists against the null hypothesis; the possibility that the observed difference arose by chance should not be excluded.

For paired continuous data, the first step is to calculate the difference between each pair of observations. Analysis then concentrates on these differences. Considering data collected before and after a particular intervention, if the intervention had no effect we would expect the average difference to be zero. The null hypothesis is therefore that any observed difference is due to chance. The most appropriate significance test is determined by the distribution of the differences (Figure 1).

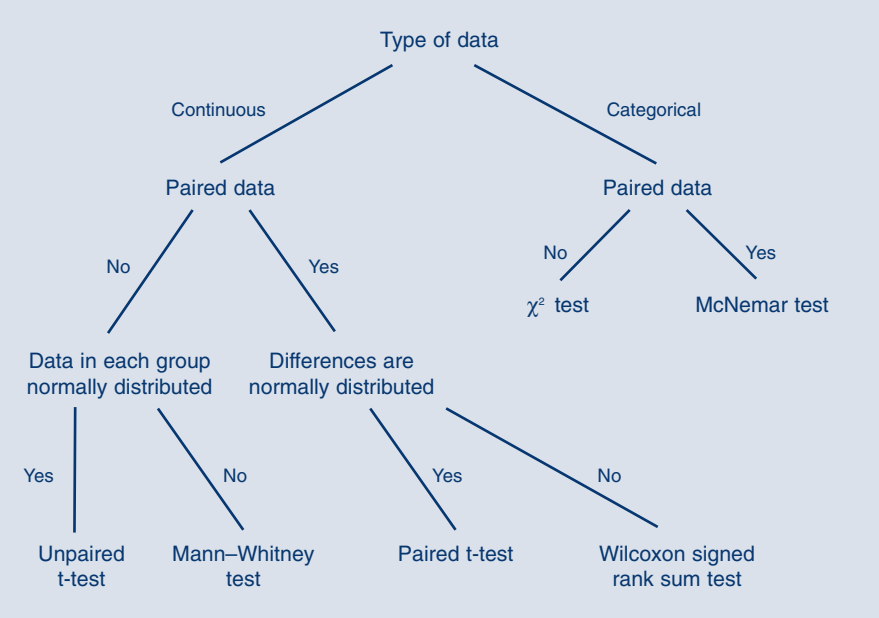
Categorical data

Associations between categorical variables can similarly be tested for statistical significance. The association between indoor heating and atopic disease among children was investigated in a Bavarian study.[4] Table 2 shows that in centrally heated homes almost eight percent of children suffered from hay fever compared with just over four percent in homes heated by coal or wood. Again, we must ask whether this indicates a real association between type of heating and hay fever, the null hypothesis being that no association exists.

From the total numbers in each row and each column it is possible to estimate how many children would be expected in each cell of the table if indoor heating and hay fever were unrelated, i.e. if the null hypothesis were true. The χ^2 test does this, and then compares the observed numbers with those expected. Large differences between observed and expected values suggest that the null hypothesis is not true and will result in small p-values. In our example, $p = 0.007$ indicating a 0.7% chance of the null hypothesis being true and thus providing strong evidence of an

association between the type of indoor heating and hay fever. While this test may appear very different to the significance tests for continuous data it is based on assessing the same three factors and is mathematically equivalent.

Figure 1: Selection of significance tests



ONE-SIDED OR TWO-SIDED *P*-VALUES

The *p*-value is the probability of observing the data if the null hypothesis is true. If a difference does exist, then the difference could be in either direction; in other words a new treatment may be either significantly better or significantly worse. To allow for either situation we use two-sided *p*-values. It is occasionally appropriate to use a one-sided test, for instance if we know a new treatment has better health outcomes and are evaluating its cost-effectiveness we may not be interested in whether or not it is cheaper, only in whether it is significantly more expensive. One-sided tests are not commonly used and most statistical software gives two-sided *p*-values as the default.

TYPE I AND TYPE II ERRORS

While observed data may provide very strong evidence of an effect, the possibility of the data arising by chance is never fully excluded. Consequently, a risk always exists that a treatment may be deemed beneficial, or an exposure harmful, when in reality it is not. Such a conclusion is known as a Type I error.

Significance tests can also result in Type II errors, these being a conclusion of no treatment benefit or of no harm from exposure, when in fact the treatment is beneficial or the exposure harmful. Small studies are particularly prone to Type II errors, a difference may be observed but in a small group it is hard to exclude the possibility that the difference arose by chance. Such studies are said to have low power to detect a difference.

The possibility of Type I and Type II errors should always be remembered, both by researchers and by those seeking to implement research results.

CONCLUSIONS

In this paper we have shown how statistical significance tests can be used to decide whether an observed difference between two groups is due to chance, or whether it signals the existence of a true difference. The different tests share a common approach in answering this question, looking at the magnitude of the difference together with the number and variability of observations. Significance tests attempt to answer the question of whether two groups differ. By themselves, they give little indication of how large or small any real difference might be, a question that we address in the next paper in this series. ■

Recommended reading

- Campbell MJ, Machin D. *Medical statistics: A common sense approach.* New York: John Wiley & Sons; 1990. p. 132–6

References

1. Sheikh A, Cook A. Descriptive statistics (Part I). *Asthma Gen Pract* 1999;7(3):32–4
2. Cook A, Sheikh A. Descriptive statistics (Part II): Interpreting study results. *Asthma Gen Pract* 2000;8(1):16–7
3. Jenkins MA, Hopper JL, Bowes G *et al*. Factors in childhood as predictors of asthma in adult life. *BMJ* 1994;309:90–3
4. von Mutius E, Illi S, Nicolai T *et al*. Relation of indoor heating with asthma, allergic sensitisation, and bronchial responsiveness: survey of children in South Bavaria. *BMJ* 1996;312:1448–50