17 Hypothesis testing

We often gather sample data in order to assess how much evidence there is against a specific hypothesis about the population. We use a process known as **hypothesis testing** (or **significance testing**) to quantify our belief against a particular hypothesis.

This chapter describes the format of hypothesis testing in general (Box 17.1); details of specific hypothesis tests are given in subsequent chapters. For easy reference, each hypothesis test is contained in a similarly formatted box.

Box 17.1 Hypothesis testing – a general overview

We define five stages when carrying out a hypothesis test:

1 Define the *null* and *alternative hypotheses* under study

2 Collect relevant data from a sample of individuals

3 Calculate the value of the *test statistic* specific to the null hypothesis

4 Compare the value of the test statistic to values from a known probability distribution

5 Interpret the *P-value* and results

Defining the null and alternative hypotheses

We usually test the **null hypothesis** (**H_0**) which assumes *no effect* (e.g. the difference in means equals zero) in the *population*. For example, if we are interested in comparing smoking rates in men and women in the population, the null hypothesis would be:

H_0: smoking rates are the same in men and women in the population

We then define the **alternative hypothesis** (**H_1**) which holds if the null hypothesis is not true. The alternative hypothesis relates more directly to the theory we wish to investigate. So, in the example, we might have:

H_1: the smoking rates are different in men and women in the population.

We have not specified any direction for the difference in smoking rates, i.e. we have not stated whether men have higher or lower rates than women in the population. This leads to what is known as a **two-tailed test** because we allow for either eventuality, and is recommended as we are rarely certain, *in advance*, of the direction of any difference, if one exists. In some, very rare, circumstances, we may carry out a **one-tailed test** in which a direction of effect is specified in H_1. This might apply if we are considering a disease from which all untreated individuals die (a new drug cannot make things worse) or if we are conducting a trial of equivalence or non-inferiority (see last section in this chapter).

Obtaining the test statistic

After collecting the data, we substitute values from our sample into a formula, specific to the test we are using, to determine a value for the **test statistic**. This reflects the amount of evidence in the data *against* the null hypothesis —usually, the larger the value, ignoring its sign, the greater the evidence.

Obtaining the *P*-value

All test statistics follow known theoretical probability distributions (Chapters 7 and 8). We relate the value of the test statistic obtained from the sample to the known distribution to obtain the ***P*-value**, the area in both (or occasionally one) tails of the probability distribution. Most computer packages provide the two-tailed *P*-value automatically. **The *P*-value is the probability of obtaining our results, or something more extreme, if the null hypothesis is true.** The null hypothesis relates to the population of interest, rather than the sample. Therefore, the null hypothesis is either true or false and we *cannot* interpret the *P*-value as the probability that the null hypothesis is true.

Using the *P*-value

We must make a decision about how much evidence we require to enable us to decide to reject the null hypothesis in favour of the alternative. The smaller the *P*-value, the greater the evidence against the null hypothesis.

- Conventionally, we consider that if the *P*-value is less than 0.05, there is sufficient evidence to reject the null hypothesis, as there is only a small chance of the results occurring if the null hypothesis were true. We then *reject* the null hypothesis and say that the results are **significant** at the 5% level (Fig. 17.1).
- In contrast, if the *P*-value is equal to or greater than 0.05, we usually conclude that there is insufficient evidence to reject the null hypothesis. We *do not reject* the null hypothesis, and we say that the results are **not significant** at the 5% level (Fig. 17.1). This does not mean that the null hypothesis is true; simply that we do not have enough evidence to reject it.

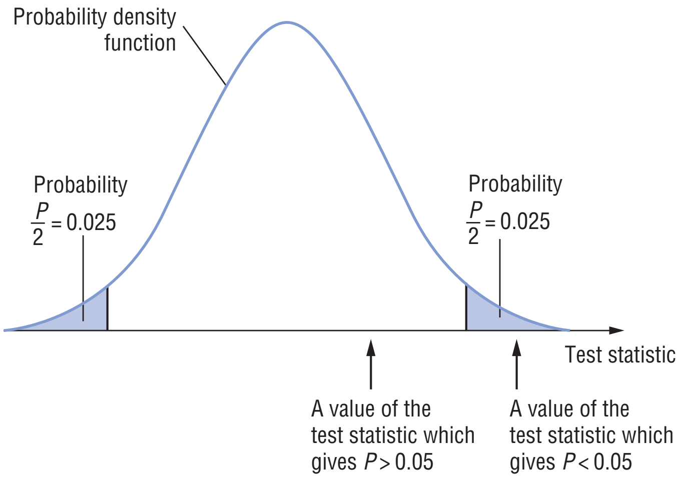

Figure 17.1 Probability distribution of the test statistic showing a two-tailed probability, $P = 0.05$.

The choice of 5% is arbitrary. On 5% of occasions we will incorrectly reject the null hypothesis when it is true. In situations in which the clinical implications of incorrectly rejecting the null hypothesis are severe, we may require stronger evidence before rejecting the null hypothesis (e.g. we may decide to reject the null hypothesis if the *P*-value is less than 0.01 or 0.001). The chosen cut-off for the *P*-value (e.g. 0.05 or 0.01) is called the **significance level** of the test; it must be chosen before the data are collected.

Quoting a result only as significant at a certain cut-off level (e.g. stating only that $P < 0.05$) can be misleading. For example, if $P = 0.04$ we would reject H_0; however, if $P = 0.06$ we would not reject it. Are these really different? Therefore, we recommend quoting the exact *P*-value, often obtained from the computer output.

Non-parametric tests

Hypothesis tests which are based on knowledge of the probability distributions that the data follow are known as **parametric tests.** Often data do not conform to the assumptions that underly these methods (Chapter 35). In these instances we can use **non-parametric tests** (sometimes referred to as **distribution-free** tests, or **rank methods**). These tests generally replace the data with their ranks (i.e. the numbers 1, 2, 3 etc., describing their position in the ordered data set) and make no assumptions about the probability distribution that the data follow.

Non-parametric tests are particularly useful when the sample size is small (so that it is impossible to assess the distribution of the data), and/or when the data are measured on a categorical scale. However, non-parametric tests are generally wasteful of information; consequently they have less power (Chapter 18) to detect a real effect than the equivalent parametric test if all the assumptions underlying the parametric test are satisfied. Furthermore, they are primarily significance tests that often do not provide estimates of the effects of interest; they lead to decisions rather than an appreciation or understanding of the data.

Which test?

Deciding which statistical test to use depends on the design of the study, the type of variable and the distribution that the data being studied follow. The flow chart on the inside front cover will aid your decision.

Hypothesis tests versus confidence intervals

Confidence intervals (Chapter 11) and hypothesis tests are closely linked. The primary aim of a hypothesis test is to make a decision and provide an exact *P*-value. A confidence interval quantifies the effect of interest (e.g. the difference in means), and enables us to assess the clinical implications of the results. However, because it provides a range of plausible values for the true effect, it can also be used to make a decision although an exact *P*-value is not provided. For example, if the hypothesized value for the effect (e.g. zero) lies outside the 95% confidence interval then we believe the hypothesized value is implausible and would reject H_0. In this instance we know that the *P*-value is less than 0.05 but do not know its exact value.

Equivalence and non-inferiority trials

In most randomized controlled trials (Chapter 14) of two or more different treatment strategies, we are usually interested in demonstrating the **superiority** of at least one treatment over the other(s). However, in some situations we may believe that a new treatment (e.g. drug) may be no more effective clinically than an existing treatment but will have other important benefits, perhaps in terms of reduced side effects, pill burden or costs. Then, we may wish to show simply that the efficacy of the new treatment is similar (in an **equivalence** trial) or not *substantially* worse (in a **non-inferiority** trial) than that of the existing treatment.

When carrying out an equivalence or non-inferiority trial, the hypothesis testing procedure used in the usual superiority trial which tests the null hypothesis that the two treatments are the same is irrelevant. This is because (1) a non-significant result does not imply non-inferiority/equivalence, and (2) even if a statistically significant effect is detected, it may be clinically unimportant. Instead, we essentially reverse the null and alternative hypotheses in an equivalence trial, so that the null hypothesis expresses a difference and the alternative hypothesis expresses equivalence.

Rather than calculating test statistics, we generally approach the problem of assessing equivalence and non-inferiority[1] by determining whether the confidence interval for the effect of interest (e.g. the difference in means between two treatment groups) lies wholly or partly within a predefined **equivalence range** (i.e. the range of values, determined by clinical experts, that corresponds to an effect of no clinical importance). If the whole of the confidence interval for the effect of interest lies within the equivalence range, then we conclude that the two treatments are equivalent; in this situation, even if the upper and lower limits of the confidence interval suggest there is benefit of one treatment over the other, it is unlikely to have any clinical importance. In a non-inferiority trial, we want to show that the new treatment is not substantially worse than the standard one (if the new treatment turns out to be better than the standard, this would be an added bonus!). In this situation, if the lower limit of the appropriate confidence interval does not fall below the lower limit of the equivalence range, then we conclude that the new treatment is not inferior.

Unless otherwise specified, the hypothesis tests in subsequent chapters are tests of superiority. Note that the methods for determining sample size described in Chapter 36 do not apply to equivalence or non-inferiority trials. The sample size required for an equivalence or non-inferiority trial[2] is generally greater than that of the comparable superiority trial if all factors that affect sample size (e.g. significance level, power) are the same.

[1] John, B., Jarvis, P., Lewis, J.A. and Ebbutt, A.F. (1996). Trials to assess equivalence: the importance of rigorous methods. *British Medical Journal* **313**; 36–39.

[2] Julious, S.A. (2004). Tutorial in Biostatistics: Sample sizes for clinical trials with Normal data. *Statistics in Medicine* **23**: 1921–1986.

18 Errors in hypothesis testing

Making a decision

Most hypothesis tests in medical statistics compare groups of people who are exposed to a variety of experiences. We may, for example, be interested in comparing the effectiveness of two forms of treatment for reducing 5 year mortality from breast cancer. For a given outcome (e.g. death), we call the *comparison of interest* (e.g. the difference in 5 year mortality rates) the **effect** of interest or, if relevant, the **treatment effect**. We express the null hypothesis in terms of no effect (e.g. the 5 year mortality from breast cancer is the same in two treatment groups); the two-sided alternative hypothesis is that the effect is not zero. We perform a hypothesis test that enables us to decide whether we have enough evidence to reject the null hypothesis (Chapter 17). We can make one of two decisions; either we reject the null hypothesis, or we do not reject it.

Making the wrong decision

Although we hope we will draw the correct conclusion about the null hypothesis, we have to recognize that, because we only have a sample of information, we may make the wrong decision when we reject/do not reject the null hypothesis. The possible mistakes we can make are shown in Table 18.1.

- **Type I error**: *we reject the null hypothesis when it is true*, and conclude that there is an effect when, in reality, there is none. The maximum chance (probability) of making a Type I error is denoted by α (alpha). This is the significance level of the test (Chapter 17); we reject the null hypothesis if our P-value is less than the significance level, i.e. if $P < \alpha$.

We must decide on the value of α before we collect our data; we usually assign a conventional value of 0.05 to it, although we might choose a more restrictive value such as 0.01 or a less restrictive value such as 0.10. Our chance of making a Type I error will never exceed our chosen significance level, say $\alpha = 0.05$, because we will only reject the null hypothesis if $P < 0.05$. If we find that $P > 0.05$, we will not reject the null hypothesis, and, consequently, do not make a Type I error.

- **Type II error**: *we do not reject the null hypothesis when it is false*, and conclude that there is no effect when one really exists. The chance of making a Type II error is denoted by β (beta); its compliment, $(1 - \beta)$, is the **power** of the test. The power, therefore, is the *probability of rejecting the null hypothesis when it is false*; i.e. it is the chance (usually expressed as a percentage) of detecting, as statistically significant, a real treatment effect of a given size.

Ideally, we should like the power of our test to be 100%; we must recognize, however, that this is impossible because there is always a chance, albeit slim, that we could make a Type II error. Fortunately, however, we know which factors affect power, and thus we can control the power of a test by giving consideration to them.

Power and related factors

It is essential that we know the power of a proposed test at the planning stage of our investigation. Clearly, we should only embark on a study if we believe that it has a 'good' chance of detecting a clinically relevant effect, if one exists (by 'good' we mean that the power should be at least 80%). It is ethically irresponsible, and wasteful of time and resources, to undertake a clinical trial that has, say, only a 40% chance of detecting a real treatment effect.

A number of factors have a direct bearing on power for a given test.

- The **sample size**: power increases with increasing sample size. This means that a large sample has a greater ability than a small sample to detect a clinically important effect if it exists. When the sample size is very small, the test may have an inadequate power to detect a particular effect. We explain how to choose sample size, with power considerations, in Chapter 36. The methods can also be used to evaluate the power of the test for a specified sample size.
- The **variability of the observations**: power increases as the variability of the observations decreases (Fig. 18.1).
- The **effect of interest**: the power of the test is greater for larger effects. A hypothesis test thus has a greater chance of detecting a large real effect than a small one.
- The **significance level**: the power is greater if the significance level is larger (this is equivalent to the probability of the Type I error (α) increasing as the probability of the Type II error (β) decreases). So, we are more likely to detect a real effect if we decide at the planning stage that we will regard our P-value as significant if it is less than 0.05 rather than less than 0.01. We can see this relationship between power and the significance level in Fig. 18.2.

Note that an inspection of the confidence interval (Chapter 11) for the effect of interest gives an indication of whether the power of the test was adequate. A wide confidence interval results from a small sample and/or data with substantial variability, and is a suggestion of low power.

Multiple hypothesis testing

Often, we want to carry out a number of significance tests on a data set, e.g. when it comprises many variables or there are more than two treatments. The Type I error rate increases dramatically as the number of comparisons increases, leading to spurious conclusions. Therefore, we should only perform a small number of tests, chosen to relate to the primary aims of the study and specified *a priori*. It is possible to use some form of *post-hoc* adjustment to the P-value to take account of the number of tests performed (Chapter 22). For example, the **Bonferroni** approach (often regarded as rather conservative) multiplies each P-value by the number of tests carried out; any decisions about significance are then based on this adjusted P-value.

Table 18.1 The consequences of hypothesis testing.

	Reject H_0	Do not reject H_0
H_0 true	Type I error	No error
H_0 false	No error	Type II error

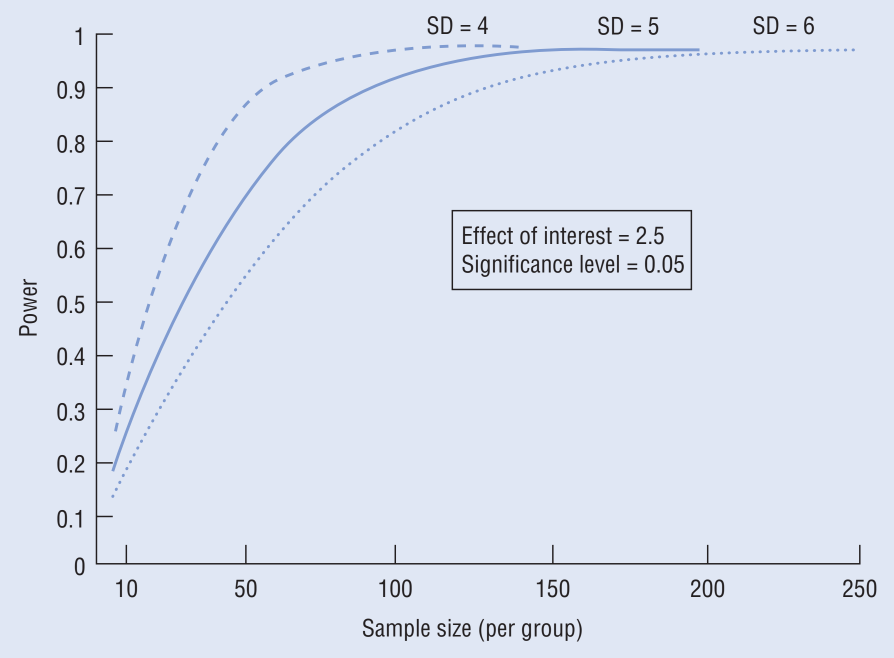

Figure 18.1 Power curves showing the relationship between power and the sample size in each of two groups for the comparison of two means using the unpaired *t*-test (Chapter 21). Each power curve relates to a two-sided test for which the significance level is 0.05, and the effect of interest (e.g. the difference between the treatment means) is 2.5. The assumed equal standard deviation of the measurements in the two groups is different for each power curve (see Example, Chapter 36).

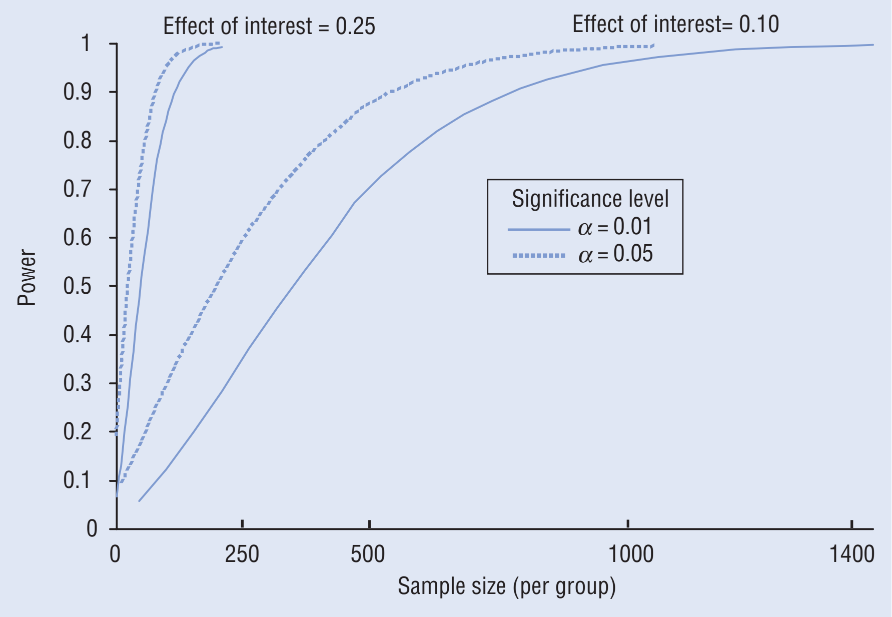

Figure 18.2 Power curves showing the relationship between power and the sample size in each of two groups for the comparison of two proportions using the Chi-squared test (Chapter 24). Curves are drawn when the effect of interest (e.g. the difference in the proportions with the characteristic of interest in the two treatment groups) is either 0.25 (i.e. 0.65 – 0.40) or 0.10 (i.e. 0.50 – 0.40); the significance level of the two-sided test is either 0.05 or 0.01 (see Example, Chapter 36).

19 Numerical data: a single group

The problem

We have a sample from a single group of individuals and one numerical or ordinal variable of interest. We are interested in whether the average of this variable takes a particular value. For example, we may have a sample of patients with a specific medical condition. We have been monitoring triglyceride levels in the blood of healthy individuals and know that they have a geometric mean of 1.74 mmol/L. We wish to know whether the average level in our patients is the same as this value.

The one-sample *t*-test

Assumptions

In the population, the variable is Normally distributed with a given (usually unknown) variance. In addition, we have taken a reasonable sample size so that we can check the assumption of Normality (Chapter 35).

Rationale

We are interested in whether the mean, μ, of the variable in the population of interest differs from some hypothesized value, μ_1. We use a test statistic that is based on the difference between the sample mean, $\bar{x}$, and μ_1. Assuming that we do not know the population variance, then this test statistic, often referred to as t, follows the t-distribution. If we do know the population variance, or the sample size is very large, then an alternative test (often called a z-test), based on the Normal distribution, may be used. However, in these situations, results from both tests are virtually identical.

Additional notation

Our sample is of size n and the estimated standard deviation is s.

1 Define the null and alternative hypotheses under study

H_0: the mean in the population, μ, equals μ_1

H_1: the mean in the population does not equal μ_1.

2 Collect relevant data from a sample of individuals

continued

3 Calculate the value of the test statistic specific to H_0

$$t = \frac{(\bar{x} - \mu_1)}{s/\sqrt{n}}$$

which follows the t-distribution with $(n - 1)$ degrees of freedom.

4 Compare the value of the test statistic to values from a known probability distribution

Refer t to Appendix A2.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the true mean in the population (Chapter 11).

The 95% confidence interval is given by:

$$\bar{x} \pm t_{0.05} \times (s/\sqrt{n})$$

where $t_{0.05}$ is the percentage point of the t-distribution with $(n - 1)$ degrees of freedom which gives a two-tailed probability of 0.05.

Interpretation of the confidence interval

The 95% confidence interval provides a range of values in which we are 95% certain that the true population mean lies. If the 95% confidence interval does not include the hypothesized value for the mean, μ_1, we reject the null hypothesis at the 5% level. If, however, the confidence interval includes μ_1, then we fail to reject the null hypothesis at that level.

If the assumptions are not satisfied

We may be concerned that the variable does not follow a Normal distribution in the population. Whereas the t-test is relatively **robust** (Chapter 35) to some degree of non-Normality, extreme skewness may be a concern. We can either transform the data, so that the variable is Normally distributed (Chapter 9), or use a non-parametric test such as the sign test or Wilcoxon signed ranks test (Chapter 20).

The sign test

Rationale

The sign test is a simple test based on the median of the distribution. We have some hypothesized value, λ, for the median in the population. If our sample comes from this population, then approximately half of the values in our sample should be greater than λ and half should be less than λ (after excluding any values which equal λ). The sign test considers the number of values in our sample that are greater (or less) than λ.

The sign test is a simple test; we can use a more powerful test, the Wilcoxon signed ranks test (Chapter 20), which takes into account the ranks of the data as well as their signs when carrying out such an analysis.

1 Define the null and alternative hypotheses under study

H_0: the median in the population equals λ

H_1: the median in the population does not equal λ.

2 Collect relevant data from a sample of individuals

3 Calculate the value of the test statistic specific to H_0

Ignore all values that are equal to λ, leaving n' values. Count the values that are greater than λ. Similarly, count the values that are less than λ. (In practice this will often involve calculating the difference between each value in the sample and λ, and noting its sign.) Consider r, the smaller of these two counts.

- If $n' \leq 10$, the test statistic is r
- If $n' > 10$, calculate $z = \dfrac{\left|r - \dfrac{n'}{2}\right| - \dfrac{1}{2}}{\dfrac{\sqrt{n'}}{2}}$

where $n'/2$ is the number of values above (or below) the median that we would expect if the null hypothesis were true. The vertical bars indicate that we take the absolute (i.e. the positive) value of the number inside the bars. The distribution of z is approximately Normal. The subtraction of $\frac{1}{2}$ in the formula for z is a **continuity correction**, which we have to include to allow for the fact that we are relating a discrete value (r) to a continuous distribution (the Normal distribution).

4 Compare the value of the test statistic to values from a known probability distribution

- If $n' \leq 10$, refer r to Appendix A6
- If $n' > 10$, refer z to Appendix A1.

5 Interpret the *P*-value and results

Interpret the P-value and calculate a confidence interval for the median—some statistical packages provide this automatically; if not, we can rank the values in order of size and refer to Appendix A7 to identify the ranks of the values that are to be used to define the limits of the confidence interval. In general, confidence intervals for the median will be wider than those for the mean.

Example

There is some evidence that high blood triglyceride levels are associated with heart disease. As part of a large cohort study on heart disease, triglyceride levels were available in 232 men who developed heart disease over the 5 years after recruitment. We are interested in whether the average triglyceride level in the population of men from which this sample is chosen is the same as that in the general population. A **one-sample *t*-test** was performed to investigate this. Triglyceride levels are skewed to the right (Fig. 8.3a); log triglyceride levels are approximately Normally distributed (Fig. 8.3b), so we performed our analysis on the log values. In the men in the general population, the mean of the log values equals $0.24 \log_{10}$ (mmol/L) equivalent to a geometric mean of 1.74 mmol/L.

1 H_0: the mean $\log_{10}$ (triglyceride level) in the population of men who develop heart disease equals 0.24 log (mmol/L)

H_1: the mean $\log_{10}$ (triglyceride level) in the population of men who develop heart disease does not equal 0.24 log (mmol/L).

2 Sample size, $n = 232$

Mean of log values, $\bar{x} = 0.31$ log (mmol/L)

Standard deviation of log values, $s = 0.23$ log (mmol/L).

3 Test statistic, $t = \dfrac{0.31 - 0.24}{0.23/\sqrt{232}} = 4.64$

4 We refer t to Appendix A2 with 231 degrees of freedom: $P < 0.001$

5 There is strong evidence to reject the null hypothesis that the geometric mean triglyceride level in the population of men who develop heart disease equals 1.74 mmol/L. The geometric mean triglyceride level in the population of men who develop heart disease is estimated as antilog (0.31) = $10^{0.31}$, which equals 2.04 mmol/L. The 95% confidence interval for the geometric mean triglyceride level ranges from 1.90 to 2.19 mmol/L (i.e. antilog $[0.31 \pm 1.96 \times 0.23/\sqrt{232}\,]$). Therefore, in this population of patients, the geometric mean triglyceride level is significantly higher than that in the general population.

continued

We can use the **sign test** to carry out a similar analysis on the untransformed triglyceride levels as this does not make any distributional assumptions. We assume that the median and geometric mean triglyceride level in the male population are similar.

1 H_0: the median triglyceride level in the population of men who develop heart disease equals 1.74 mmol/L.

H_1: the median triglyceride level in the population of men who develop heart disease does not equal 1.74 mmol/L.

2 In this data set, the median value equals 1.94 mmol/L.

3 We investigate the differences between each value and 1.74. There are 231 non-zero differences, of which 135 are positive and 96 are negative. Therefore, $r = 96$. As the number of non-zero differences is greater than 10, we calculate:

$$z = \frac{\left|96 - \frac{231}{2}\right| - \frac{1}{2}}{\frac{\sqrt{231}}{2}} = 2.50$$

4 We refer z to Appendix A1: $P = 0.012$.

5 There is evidence to reject the null hypothesis that the median triglyceride level in the population of men who develop heart disease equals 1.74 mmol/L. The formula in Appendix A7 indicates that the 95% confidence interval for the population median is given by the 101st and 132nd ranked values; these are 1.77 and 2.16 mmol/L. Therefore, in this population of patients, the median triglyceride level is significantly higher than that in the general population.

Data kindly provided by Dr F.C. Lampe, Ms M. Walker and Dr P. Whincup, Department of Primary Care and Population Sciences, Royal Free and University College Medical School, London, UK.