

Cross-Cultural Tourism Behaviour: Concepts and Analysis

Yvette Reisinger, PhD and Lindsay W. Turner, PhD

CONTENTS

Hypothesis testing for cross-cultural comparison

- 1 Introduction to parametric and non-parametric hypothesis testing
- 2 The hypothesis test
- 3 Parametric hypothesis test
 - 3.1 When to use z - and t -tests
 - 3.2 One- and two-tailed tests
 - 3.3 One-sample means test example
 - 3.4 Type I and Type II errors
 - 3.5 Two-sample means test
 - 3.6 Unpaired test
 - 3.7 Paired sample test
 - 3.8 Hypothesis interpretation
- 4 Introduction to non-parametric hypothesis testing
 - 4.1 One-sample non-parametric test
 - 4.2 Paired two-sample non-parametric test
 - 4.3 Unpaired two-sample non-parametric test
 - 4.4 Multiple paired sample test
 - 4.5 Multiple unpaired sample test
- 5 Cross-cultural behaviour: example analysis
 - Summary
 - Discussion points and questions
 - Further reading

Hypothesis testing for cross-cultural comparison

OBJECTIVES: *After completing this text the reader should be able to:*

1. Understand the difference between parametric and non-parametric hypotheses tests
2. Conduct one- and two-sample parametric hypotheses tests
3. Conduct a range of non-parametric hypotheses tests
4. Understand the application of hypothesis testing to cultural tourism analysis

1. Introduction to parametric and non-parametric hypothesis testing

Statistical inference procedures enable researchers to determine, in terms of probability, whether the observed differences between sample data could easily occur by chance or not. Whenever random data is collected – usually in tourism this is by some type of survey – there are likely to be some differences between the survey data and the general population at large or between different samples. For example, a comparison of the average age (the *sample statistic*) between Japanese tourists to Hawaii and the average age of the resident population of Hawaii is likely to be different. The questions these methods answer is whether the difference is simply due to chance because a sample of Japanese tourists was surveyed, or whether it is highly probable that the difference is real. Of course, if the Japanese tourists were not surveyed because we already knew, perhaps from official immigration records, the real average age of the population of Japanese tourists (*population parameter*), as well as the real average age of the resident population, then the issue of chance variation from sampling does not arise. In such a case, the comparison is direct and accurate and the degree of difference is the known non-probable degree of difference. As such, we no longer need to test for the probability of a difference and we no longer need to construct an hypothesis test to measure the probability of there being a likely difference.

Since social scientists are commonly dealing with surveyed data, the need for constructing hypothesis tests to test the difference between means is common. Many statistical texts deal with this material, along with the concepts of probability and probability distributions (an understanding of which is needed to use the following material). Here we use this

opportunity to place the testing of hypotheses into a tourism example framework, describe the major issues of hypothesis testing facing the tourism cultural researcher, and provide an example of hypothesis testing in tourism culture research. Further research and a more complex methodology can be found in our book *Cross-Cultural Tourism Behaviour: Concepts and Analysis* (2003) published by Elsevier Science Ltd.

Before beginning this discussion it is both interesting and important to discuss the difference between parametric and non-parametric analysis because this leads to a major decision choice for the researcher – whether to use a parametric or non-parametric hypothesis test. In the development of modern statistics the first methods developed made a lot of assumptions about the characteristics of the population from which the samples were drawn. That is, they made assumptions about the statistical values of the population (called *parameters*), which became referred to as parametric tests. The most obvious assumption is that the scores in the survey were randomly drawn from a normally distributed population. Another less well-known assumption is that the scores are randomly drawn from populations having the same variance (standard deviation squared), or spread of scores. These assumptions make the general overriding assumption that the probability distribution of the population (from which the sample was drawn) is known in advance. The most common distribution assumed is the normal distribution.

More recently, distribution free or non-parametric tests have been developed and subsequently commonly used. These tests have fewer qualifications and in particular do not have the overriding assumption of a normally distributed population base.

In quantitative terms the difference rests upon the way in which the scores are manipulated. In parametric tests the scores are added, divided and multiplied and these processes introduce distortions to the scores so that tests upon the data must use methods assuming a truly numeric distribution. On the other hand, many non-parametric tests manipulate the data by ranking and thus avoid the numeric value of the scores themselves. Such tests then summarize the scores by creating summary statistics (statistics come from samples) that are

derived physically such as the mode, or the median (where the data is ranked), instead of the mean (which involves addition and division).

In this text we will look at both types of hypothesis tests (parametric and non-parametric) and describe their calculation and use. Particular attention will be given to the often under-used non-parametric tests because data that is culture-based is quite likely to not have a normal distribution on which to base parametric data manipulations.

2. The hypothesis test

The hypothesis test comprises two mutually exclusive statements, the alternative and the null hypotheses. The null hypothesis states the negative case, that 'it is not true or there is no difference', and the alternative hypothesis states that 'it is true or there is a difference'. The procedure involved is a scientific one that is founded in simple logic for the purpose of being both open and potentially repetitive (can be replicated by others).

The following steps outline the hypothesis testing procedure:

1. State the null (H_0) and alternative (H_1) hypotheses.
2. Choose a statistical test to test H_0 . Decide whether parametric or non-parametric.
3. Specify a significance level (α) or probability level for rejection of H_0 .
4. Determine the sample size (N).
5. Assume (or find) the sampling distribution of the statistical test in 2.
6. On the basis of 2, 3, 4 and 5 above, define the region of rejection of H_0 .
7. Compute the value of the statistical test using the sample data.
8. If the resultant value of the test is in the rejection area, reject H_0 .
9. If the resultant value of the test is outside the rejection area, H_0 is not rejected at the level of α .

Note that not rejecting H_0 does not lead unequivocally to the acceptance of H_1 . This is because the test did not test H_1 , but H_0 . Note also that the test can be directional, that is,

one mean is greater than or lesser than the other, and these types of tests are discussed later under the heading One- and two-tailed tests (see Section 3.2).

3. Parametric hypothesis test

The most common parametric hypothesis tests are the z -, t - and F-tests. The z - and t -tests are commonly used for the testing of means and are the focus of the following discussion. F-tests are most commonly used in multivariate analysis and assume the F probability distribution. The F-tests are slightly more rigorous in their assumptions than z - and t -tests (discussed below).

3.1 *When to use z - and t -tests*

In order to test the null hypothesis it is necessary to determine first whether or not we know what the standard deviation of the population is. If the population standard deviation (σ) is known, the hypothesis testing can be done using the z -test. If the population standard deviation is unknown we should use the t -test. Both tests assume the distribution to be symmetric but the tails of the distribution are higher for the t than the z -test and there is an individual difference in the heights of the tails for each sample size (N). This occurs because the sample standard deviation (s) must be substituted for the population standard deviation (σ) so that there is more variability resulting from the independence of s and the sample mean (\bar{x}). When \bar{x} is very large, s may be very small and vice versa. This variability does not occur in the normal distribution as the only random variation occurs with the population mean (μ) as the other quantities population size (N) and σ are non-random.

The following conditions are required for the use of parametric tests:

1. The observations must be independent of each other. The selection of one person for the survey should not influence the choice of others, and the answer to one question by a respondent should not bias the answers to other questions.
2. The observations (scores) must be drawn from normally distributed populations.

3. The populations must have the same variance (be homoscedastic).
4. The variables must be measured in at least an interval scale (known gaps between each unit of measurement – usually the gap is one: 1,2,3,4,5) so that meaningful addition, multiplication, subtraction and division can occur.
5. For the F-test the means of these normal and homoscedastic populations must be linear combinations of effects due to columns and/or rows. The effects must be additive.

If the conditions above are met then the choice of test should be parametric because these tests are more powerful (more confidence can be placed upon the result) than non-parametric equivalents. When these conditions are not met and the analysis is still used there can be no confidence in the results, meaning they are powerless.

3.2 One- and two-tailed tests

The difference between one- and two-tailed tests relates back to point 6 in the steps of hypothesis testing (see earlier in Section 2), the region of rejection. The region of rejection in the probability distribution is in an extreme tail of the distribution of values that are very high or low relative to the mean (which is toward the middle). These values are so extreme that when H_0 is true the probability is very small (less or equal to alpha) that the sample we actually observe will yield a mean value among them.

The location of the region of rejection is determined by whether the test is one-tail positive (one mean is larger than another and the region of rejection is upper values above the mean) or one-tail negative (one mean is less than another and the region of rejection is lower values below the mean). If the test is two-tail, there is either a positive or negative difference (the comparative time taken to register guests in one hotel is faster or slower than another) and the area of rejection can be in either tail of the distribution.

The size of the region of rejection is expressed as alpha, the level of significance. If $\alpha=0.05$, then the size of the rejection region is 5 per cent of the entire space under the curve of the probability distribution. Refer to Figure 1.

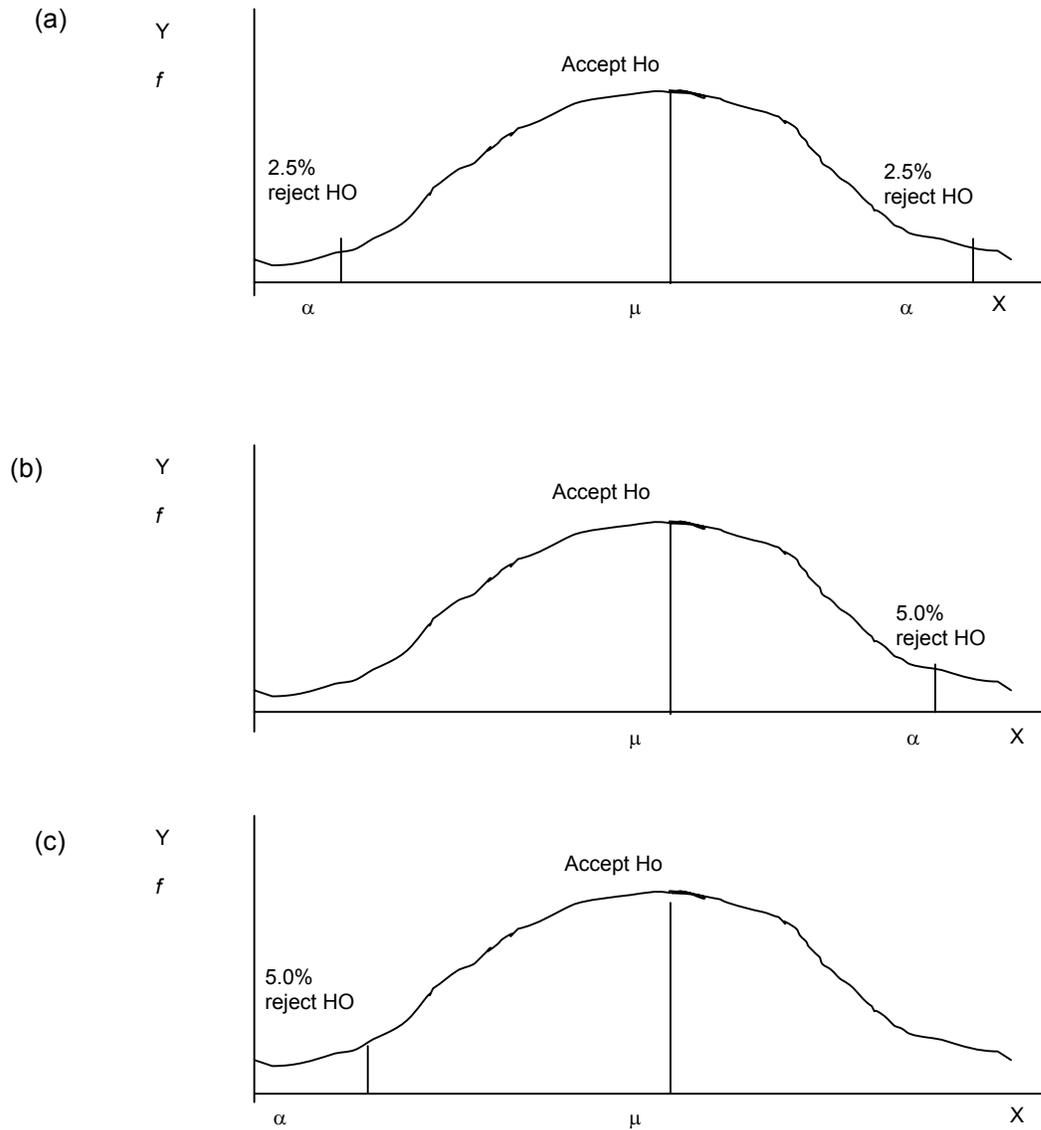


Figure 1. One- and two-tailed tests. (a) two-tailed test; (b) one-tail positive test; (c) one-tail negative test.

Notice that the probability distribution is represented in Figure 1 by a symmetrical bell-shaped curve, with μ representing the population mean (the most commonly occurring value at the centre of the curve).

To reject the null hypothesis the calculated value of z or t from the analysis (t or z obtained) must be a greater or lesser value than $\alpha = \alpha$ in the appropriate test (positive or negative or two-tailed).

In this comparison it can be seen from Figure 1 that the critical value of α is closer to the mean (μ) in the one-tail test (all other factors being equal). That is the area of rejection in the one-tail test is smaller in each tail than the two-tail test. Hence, the chance of rejection occurring is higher for a one-tail test than the equivalent two-tail test. This is because the additional knowledge of knowing the direction of rejection allows for a less rigorous test.

3.3 One-sample means test example

The one-sample means test compares a single sample mean against a known population mean. Later on, the cases of using two sample means are analysed separately.

In the study of tourism culture there are not many known population means because culture is not the focus of most official tourism or census database collections. However, there are known population demographics such as average age and income obtainable from government census data. The following example is from preliminary analysis of a collected sample used to assess the representativeness of the data to the base population.

A survey of 250 Australian host workers on the Gold Coast (Australia) was conducted in 1996. The average age of the workers who were employed in various contact positions with international tourists was 33 years of age with a standard deviation of 4.85 years of age. From the 1996 Australian census the average age of the workforce in the area was 38.5 years of age. However, the standard deviation for the population workforce was not available. It is not uncommon that it is difficult to get variance measures from official databases and in consequence it is not possible to conduct a z -test. It is reasonable to assume the population of ages is normally distributed. Nevertheless a test was made for skew in the survey data and the skew was a small 0.045 measure on a Pearson Skewness test where zero indicates absolutely no skew and measures above 1.5 are clearly skewed. Consequently, from the earlier discussion a t -test is required.

The *null* hypothesis in this case is: the average age of the hosts is equal to or greater than the workforce in general.

The *alternative* hypothesis is: the average age of the hosts is lower than the average age of the workforce in general.

Since the rejection region is determined by the null hypothesis, to reject H_0 a one-tail negative test is required.

The value of alpha can be determined for the level of significance $\alpha=0.05$. For a *t*-test the sample size is needed to determine alpha and at a very large sample size of 250. However, the *t* critical values are approximate to values of the normal distribution beyond about 120° of freedom (measured as $N-1$), so for 249° of freedom the critical value becomes 1.96 (two-tail) and 1.645 (one-tail), the same as the normal distribution.

This example draws out the difference between the *z*- and *t*-test and how sample size can be used as another rule for choosing between the *z*- and *t*-test procedures. The standard deviation initially used to determine a *z*-test could not be done because the standard deviation of the population was unknown. It has been found in early statistical research that the normal distribution varies with small sample sizes (the tails get higher). So there is a different distribution for *t* from $N=1$ to 100 or maybe 120 depending on the number of decimal places. So if the sample size is small the *t*-test is definitely needed, but if the sample is 100 or more, the *t*-test is no longer needed regardless of the known or unknown values of the standard deviation. This leads to the question of how small does the sample size need to be before the difference between the values of *t* and *z* become so great as to be worth worrying about. There is no clear answer – some researchers say as low as 30 or 40; some say as low as 60.

For our example, a sample size of 250 is way beyond 120 and a *z*-test can be used and the sample standard deviation can be substituted for the population standard deviation (that remains unknown).

Therefore, a simple rule can be stated for the choice between *z*- and *t*-tests.

If you don't know the standard deviation of the population and the sample size is less than 100 (maybe 60) use the normal distribution z -test. In all other cases use the t distribution and the t -test.

To conduct the analysis the following information is required:

the sample mean = \bar{x}

the population standard deviation = σ or sample standard deviation = s

the sample size = N

the alpha significance value = α

the population mean = μ

In our example the information available is:

\bar{x} = 33 years of age (sample mean – statistic)

s = 4.85 years (sample standard deviation) population value unknown

N = 250 respondents to the survey

α = 1.645 (t critical one-tail negative at 0.05 significance)

μ = 38.5 years of age (population mean – parameter)

We can now rephrase our scientific model in simple terms to be: is the value of 33 when drawn from a sample of size $N=250$, 95 per cent certain to be less than a value of 38.5 given a standard deviation of 4.85?

We know that sample means are approximately normally distributed when the sample size is 100 or larger. So if the survey were conducted numerous times, with each sample size larger than 100, we would expect a normal bell shaped curve.

The standard error is calculated as:

$$\begin{aligned} & \delta \div \sqrt{N} \\ & = 4.85 \div \sqrt{250} \\ & = 0.3067 \end{aligned}$$

The standard error can best be understood as the ‘sampling error’. It is a measure of the inaccuracy of using a sample from a population and weights out the degree of variation in the data (measured by the standard deviation) by the size of the sample N . The smaller the degree to which using a sample causes error, the better. So if the variation is high then a large sample will be needed to reduce error (because the sample size is divided into the variation). This calculation is a good example of the reason why it is always a good idea to have a larger rather than smaller sample.

A sampling distribution may have a mean value of 38.5 (represented by zero as the mean of the normal distribution), occasionally a mean as low as say 30 might occur in any given sample or markedly higher over 40. The question here is whether the mean found in the sample is merely lower than 38.5 because it comes from one of many possible sample variants around a true mean of 38.5, or whether it is so low that it will likely occur five or fewer times if 100 survey samples were repeatedly taken. If that is the case, the mean of 33 probably comes from another population that has a lower mean than 38.5. So when 33 are converted to the normal distribution equivalent (where 38.5 is equal to zero) is the value of 33 less than -1.645 ?

Put in the terminology of hypothesis testing:

If (Z obtained) exceeds -1.645 reject H_0 .

Note here the important use of the word ‘exceeds’. In this context the use of ‘exceeds’ means either greater in a positive direction (in our example greater than 1.645) or greater in the negative direction (in our example a greater negative than -1.645). Therefore, it is important not to substitute the words ‘greater than’ for the word ‘exceeds’ because while this would be correct in the positive direction, it would be incorrect in the negative direction.

To derive z obtained the formula is:

$$Z = \frac{\mu - \bar{x}}{s/\sqrt{N}} = \frac{33 - 38.5}{0.3067} = -17.93$$

Since Z obtained at -17.93 exceeds -1.645 as a greater negative value, the null hypothesis H_0 is rejected and the alternative hypothesis H_1 is accepted.

Therefore, it has been decided that the difference between the sample mean and the population mean is due to something other than random variation. It is 95 per cent certain that the mean age of the hosts is less than the mean age of the general workforce. In the example case the Z obtained value is so large at 17.93 that the significance level of 95 per cent is well and truly exceeded. In fact the significance exceeds 99 per cent (z critical 2.35).

3.4 Type I and Type II errors

Regardless of the results of the computation there is the possibility that an error has been made. In hypothesis testing the error can take two forms, a Type I error and a Type II error.

A *Type I error* occurs when the null hypothesis is rejected when in fact it is true.

A *Type II error* occurs when the null hypothesis is accepted when in fact it is false.

The probability of making a Type I error is determined by the level of significance in the test. If 95 per cent significance is chosen, then the chance of the null hypothesis being rejected, when in fact it is true, must be 5 per cent. Consequently, the higher the level of significance, the less the chance of a Type I error.

However, what is the probability of committing a Type II error? This probability is less obvious than the probability of a Type I error and requires separate and less specific calculation. Before considering this calculation it is best to see the relationship between each error Type as depicted in Figure 2.

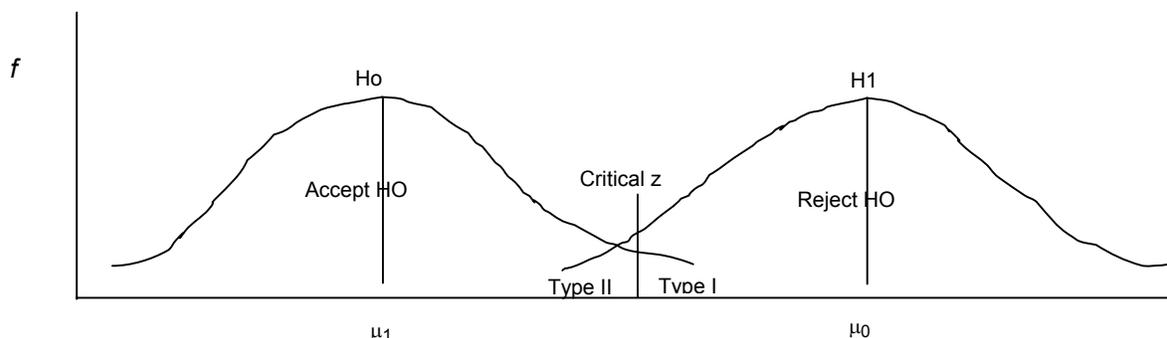


Figure 2. The probabilities of a Type I and Type II error one-tail positive test.

The probability of a Type I error is the area of rejection of the null hypothesis (H_0) to the right of the critical z -line. The probability of Type II error is the area of the tail of the alternative hypothesis (H_1) that extends into the region of acceptance of H_0 to the left of the vertical line critical z ; this has already been defined as the probability that the sample mean will fall into the acceptance region of H_0 , when in fact H_1 is correct.

The problem of calculating the area of a Type II error under the normal curve now becomes apparent in that the calculation requires the knowledge of μ_1 and an estimate of σ_1 that is not known unless the alternative hypothesis structure is a simple equality. In such a case μ_1 is a known value and σ_1 can be estimated by σ and calculation can be straightforward. However, in most social science situations that is not the case, and the alternative hypothesis is complex; μ_1 can take on a range of values above μ_0 and for a given level of significance the probability of a Type II error is specified as a function and not a single value.

As can be seen from Figure 2 if there is a reduction in the Type I error (the critical z -value line moves to the right) so there is an increase in the Type II error and vice versa. Consequently, in selecting the level of significance the researcher must determine the importance of each type of error.

The consequences for the hosts survey on the Gold Coast of a Type II error are not really high: the average age of the hosts would not be recognized to be less than the general workforce average age. On the other hand, a Type I error would cause the claim to be made that the average age of the host workforce is younger when in fact it was not. Therefore, the higher the significance level the better.

There are situations where the Type II error can be important and they are common in both medical and legal research. Consider the case of a corporate lawyer making a decision concerning whether or not a franchise contract for a travel agent can be broken or not. According to the results of a survey of lawyers, the corporate lawyer may decide that his client's case differs enough from an adverse precedent to continue with the case, despite

costs. If in fact the precedent covers the case, the lawyer will likely lose. A Type II error would commit the client to legal costs and subsequent loss of the case and potentially an adverse continuing financial position for the travel agent. A Type I error would mean the precedent was judged important when in fact it was not and clients not proceeding with the case when in fact they would likely win. In such a case the lawyer may use a lower significance level to decrease the chance of incurring unnecessary costs (a Type II error).

Fortunately, there is a more reliable way to reduce the probabilities of both error types. That is by increasing the sample size N . Since the spread in the sampling distribution is equal to σ/\sqrt{N} (which is estimated by s/\sqrt{N} where the population standard deviation is unknown) a larger value of N will produce a smaller value of the deviation among the means of the sampling distribution. When sample size increases from N to some larger size N the distributions condense around the true means and both error areas in the distribution curve tails decrease in size. Another reason why using a large sample is important for parametric analysis.

3.5 Two-sample means test

In the case of a two-sample means tests it is possible to determine the likelihood of two-sample means coming from the same population. For example, are the average responses to the question ‘I want to see the historic buildings and architecture’ the same for Chinese and French tourists? Literature suggests (Guo *et al*, 2002) that the Chinese still have a historically based interest in buildings and architecture that is less common among western tourists.

In terms of the statistical analysis three confounding complexities enter into the study of the difference between two means. One of the complexities has already been encountered in the previous section; on one-sample tests a difference in statistical manipulations occurs when the population standard deviation is known and unknown leading to the question of whether to use a z - or a t -test. The second complexity requires a different calculation when the samples selected are independent (unpaired) samples as opposed to related (paired) samples, and the third complexity is that the two-sample test cannot be conducted unless

the populations from which the two samples have been drawn have homogeneous variances (homoscedasticity). This assumption is often testable in the two-sample case, unlike the one-sample case.

First of all we shall examine the problems of homogeneous variances and then discuss the unpaired design analysis (assuming σ unknown) followed by the paired design analysis (with σ unknown).

Although the ultimate interest is the difference between means, any test first assumes homogeneous variances of the populations. Consequently, a pre-test should be conducted to test for a difference between the two standard deviations.

The hypotheses for this homoscedastic test are:

The null hypothesis (H_0) states that the observed difference between s^2 (1) and s^2 (2) (the variance of the first and second samples) is due simply to random variability, because the population variances are equal.

The alternative hypothesis states that the sample variances are different because the population variances differ.

The test is two-tail as no direction for the difference is specified.

The unbiased sample variances are used to calculate a statistic known as the F-ratio, which is calculated by dividing the larger of the two variances by the smaller (thus always yielding a positive result):

$$F - \text{obtained} = \frac{S^2 \text{ larger}}{S^2 \text{ smaller}}$$

There is a different F -distribution for different degrees of freedom in the denominator and numerator of the obtained F -ratio. Critical values of the F -distribution can be read off the relevant tables, with degrees of freedom at $N-1$ for both the denominator and numerator, where N is the sample size.

Values of F smaller than 1 cannot occur in this test, because the larger variance is always placed over the smaller variance. The conclusion upon the analysis is obtained by the test:

if (F -obtained) positively exceeds F -critical, reject H_0 .

In order to continue with the analysis it is necessary that the statistician *does not* reject the null hypothesis (H_0), and thus concludes that the samples are likely to be drawn from populations with equal variances. It is reasonable, and conventional to accept or reject the null hypothesis at the 95 per cent level of significance. However, acceptance of the null hypothesis does not mean the statistician is 95 per cent sure that the samples have been drawn from populations with equal variances, it is merely 'likely to be so' because of the way the hypotheses have been conventionally structured to test the null hypothesis that there is no difference.

The difference between two sample means can be assumed to create a sampling distribution, in the same way that the means of samples can also be used to create a distribution. That is, if numerous means were calculated from numerous samples, or, if sampling numerous differences between two means from numerous sets of samples were calculated, they could be used to plot a sampling distribution.

The parameters would be: $\mu_{(\bar{x}_1 - \bar{x}_2)} = \mu_1 - \mu_2$

$$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

where

$\mu_{(\bar{x}_1 - \bar{x}_2)}$ = mean of the sampling distribution of the difference between two means.

$\sigma_{(\bar{x}_1 - \bar{x}_2)}$ = standard error of the sampling distribution.

The distribution, once plotted, would be normal in shape if large enough sample sizes are used, and the standard deviations of the populations are known. Unfortunately, the population standard deviations are rarely known and consequently it is necessary to use the t -distribution rather than the Z -distribution.

Now let us consider the analysis of the difference between two sample means, first in the case of an unpaired analysis and secondly in the case of a paired analysis.

In an *unpaired* sample design, the individuals making up the samples would be selected randomly from the relevant population with each member of the population having an equal chance of being chosen. In a *paired* sample design the individuals would be chosen as a matched sample (characteristics of each sample matched as closely as possible to reduce variability of characteristics other than that being tested) or a related sample (same individuals used for both samples) and the sample sizes would be the same.

It could be expected that a greater degree of variability between sample observations occurs with unpaired designs so that extra attempts are made to maintain a control upon random error.

3.6 Unpaired test

When σ is unknown, s must be used to estimate σ ($\bar{x}_1 - \bar{x}_2$) and the sampling distribution is shaped as a t distribution, with degrees of freedom equal to $N1+N2-2$.

Let us use the example discussed previously of the difference between the Chinese and French in regard to the statement ‘I want to see the historic buildings and architecture’, measured on a seven-point scale (0=no importance at all to 7=extremely important). Let us assume the following statistics:

| | | |
|-------------------|-----------|-------------------|
| <i>Chinese</i> | | <i>French</i> |
| $\bar{x}_1 = 6.1$ | $N1 = 40$ | $\bar{x}_2 = 5.2$ |
| $s_1^2 = 1.2$ | $N2 = 35$ | $s_2^2 = 0.9$ |

The variances when tested are unlikely to be non-homoscedastic.

The null hypothesis states that the two samples come from the same population, that is, the means are equal to each other or the Chinese mean is lower than the French mean. The alternative hypothesis states that the Chinese mean is higher than the French mean. Consequently, it is a one-tail positive test.

If (t -obtained) exceeds t -critical, reject H_0 .

To find t -obtained:

$$t\text{-obtained} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\text{Estimated } \sigma_{\bar{x}_1 - \bar{x}_2}}$$

where: $(\bar{x}_1 - \bar{x}_2)$ is the calculated statistic. $(\mu_1 - \mu_2)$ is the mean of the sampling distribution specified by H_0 .

$$\text{Thus: } t\text{-obtained} = \frac{(6.1 - 5.2) - (0 - 0)}{\text{Estimated } \sigma_{\bar{x}_1 - \bar{x}_2}}$$

In order to find the estimated standard error in the denominator it is necessary to compute the pooled variance (Sp^2). It is feasible to pool the simple variances because homogeneity of variances has been tested for (refer to the F -test above):

$$\text{Pooled variance}(Sp^2) = \frac{(N1-1)(s_1^2) + (N2-1)(s_2^2)}{N1 + N2 - 2}$$

$$\text{Thus: } Sp^2 = \frac{(40-1)(1.2) + (35-1)(0.9)}{40 + 35 - 2} = \frac{46.8 + 30.6}{73} = \frac{77.4}{73} = 1.06$$

The pooled variance will always fall in between the values of the two individual variances.

The pooled variance is then used to compute the estimated $\sigma_{\bar{x}_1 - \bar{x}_2}$.

$$\text{Estimated } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{1.06}{40} + \frac{1.06}{35}} = \sqrt{0.0265 + 0.0303} = 0.2383.$$

Now the value of t -obtained can be calculated:

$$t - \text{obtained} = \frac{(6.1 - 5.2) - (0 - 0)}{0.2383} = \frac{0.9 - 0}{0.2383} = 3.78$$

Critical t is derived from the t -distribution at $N1+N-2$ degrees of freedom ($40+35-2$), or 73 degrees of freedom, to be 1.66 at 95 per cent significance for a one-tail test.

Since 15.85 is greater than 1.66 (critical t) the null hypothesis is rejected and it is concluded at 95 per cent significance that the Chinese rate buildings and architecture more importantly as a destination characteristic than the French. The high t -obtained indicates that rejection is also possible at 99 per cent statistical significance.

3.7 Paired sample test

The use of paired samples eliminates some random variability. Variability is now concentrated to *within* sample variability, or variations among the sample individuals. The *between* sample variability has been reduced in the paired sample test through the use of particular sampling methodology such as matched sampling.

The consequence in statistical terms of reduced variability is a smaller estimated standard error of the sampling distribution of mean differences ($\sigma(\bar{x}_1 - \bar{x}_2)$) when the correlation between the samples is positive, which in turn produces a larger t -obtained. The reverse occurs when the correlation is negative.

Calculation of the paired sample test can be computed via a technique known as the direct-differences method.

A hypothetical example is used to illustrate the analysis. Management of a particular airline has agreed that the time taken for some service teams to turn around a large international aircraft depends upon the length of the incoming flight duration. However, management has argued with the cleaning union that the time taken to clean the plane is

the same regardless of flight times. In order to test this argument, two eight-person cleaning teams are compared in the times taken to complete each of their assigned tasks, team one after a 12-hour flight and team two after a four-hour flight. The teams are matched in most relevant characteristics including training, job specification and experience. Table 1 lists the data collected from the test and the analysis procedure. The data is collected as a percentage efficiency rating for each team member matched by his or her specific job, and includes both time taken and quality of the finished job measured as a percentage.

Table 1. Direct difference method for the paired-sample test

| <i>Pair</i> | <i>Sample 1</i> | <i>Sample 2</i> | <i>Difference (D)</i> | $(D - \bar{D})$ | <i>Difference²</i> |
|-------------|-----------------|-----------------|-----------------------|-----------------|-------------------------------|
| 1 | 70 | 68 | -2 | 0.5 | 0.25 |
| 2 | 65 | 69 | 4 | -5.5 | 30.25 |
| 3 | 56 | 54 | -2 | 0.5 | 0.25 |
| 4 | 73 | 69 | -4 | 2.5 | 6.25 |
| 5 | 62 | 63 | 1 | -2.5 | 6.25 |
| 6 | 60 | 57 | -3 | 1.5 | 2.25 |
| 7 | 78 | 70 | -8 | 6.5 | 42.25 |
| 8 | 65 | 67 | 2 | -3.5 | 12.25 |
| | $\Sigma = 529$ | $\Sigma = 517$ | $\Sigma = -12$ | | $\Sigma = 100.00$ |

$$\bar{x}_1 = 66.125 \quad \text{Difference } \bar{x}_1 - \bar{x}_2 = -1.5 \quad \bar{D} = \frac{-12}{8} = -1.5$$

$$\bar{x}_2 = 64.625$$

$$SD = \sqrt{\frac{\sum d^2}{N-1}} = \text{standard deviation of the } D \text{ array}$$

$$SD = \sqrt{\frac{100}{7}} = \sqrt{14.285714} = 3.7796$$

$$\text{Standard error of the } \bar{D} \text{ statistic } S\bar{D} = \frac{SD}{\sqrt{N}} = \frac{3.7796}{\sqrt{8}} = \frac{3.7796}{2.8284} = 1.3363$$

$$t \text{ obtained} = \frac{\bar{D}}{S\bar{D}} = \frac{-1.5}{1.3363} = -1.1225$$

The direct-differences method effectively converts to a single sample means test where μ_D equals the mean difference between members of each pair and becomes a single mean.

The similarity can be seen in Table 1 by the calculations of \bar{D} and $\bar{x}_1 - \bar{x}_2$ as both equal -1.5 . Notice that sample two is always deducted from sample one to derive \bar{D} .

The statistic \bar{D} is assumed to come from a sampling distribution that is shaped as a t -distribution with $N-1$ degree of freedom. In the example $N=8$, so d.f.=7. The critical t -value at 7 d.f. and 95 per cent significance is 2.365 for a two-tail test.

If (t -obtained) exceeds ± 2.365 , reject H_0 .

Since t -obtained at -1.1225 does not equal or exceed t critical at -2.365 , the null hypothesis is not rejected and the conclusion is that the two samples come from populations with the same mean, so that there is likely to be no difference between the two teams in terms of cleaning efficiency. This test could be repeated for numerous team comparisons to see if the findings (for various different flight times) result in a consistent conclusion.

3.8 Hypothesis interpretation

Finally, in interpreting the results of an analysis, the researcher must be very careful not to, on the one hand, break any of the statistical rules, and on the other hand make a post-hoc error.

The statistical rules that are most important concern the assumption of data distributions, most commonly the normal distribution. Both the z -test and the t -test assume data is distributed normally.

If the data does not come from a population that is normally distributed and the sample is small in number, (may not be normally distributed) all analysis is circumspect. Second, the assumption of homogeneous variance is equally important, and failure to test for this is a major weakness that may render the analysis useless.

Post-hoc errors are often made in social science analysis where the data is less likely to take the form of an experiment, and this point can best be described by example. If any factor (independent variable) remains unaccounted for in the analysis it may be a hidden cause of

the effect under study. Random selection of individuals can often be used incorrectly, and thus not control for extraneous independent variables. For example, it is often stated that young drivers are more dangerous on the roads because they are less experienced and have more accidents. While that may be correct, a test that did not allow for the fact that young drivers choose more dangerous vehicles (faster cars, or older or cheaper cars) may be committing a post-hoc error.

4. Introduction to non-parametric hypothesis testing

Non-parametric statistical techniques place few limitations upon the level of measurement for data because they do not require the use of statistical parameters such as the mean and standard deviation. They make few assumptions that data is distributed in a particular manner, and no assumption of the normal distribution is necessary except in cases where an approximation is made because of large sample sizes. They can do this because their manipulations are based upon nominal or ordinal measurement rather than ratio measures. Consequently, where the mean (requiring ratio data) may be used in a parametric analysis, the median (requiring ordinal data) may be used in a non-parametric calculation.

To facilitate the use of non-parametric techniques, this text discusses several of the most common tests for use in tourism analysis. These are divided into groups on the basis of the type of sample required. First, one-sample tests are discussed. Second, tests requiring two related (matched or paired) samples. Third, tests requiring two independent (unrelated or unpaired) samples. Fourth, tests requiring more than two (multiple) related samples and finally, tests requiring multiple independent samples.

Some non-parametric tests are not covered here, and for wider reference, Brownlee (1965) is a recommended reading.

4.1 *One-sample non-parametric test*

The most common non-parametric one-sample test is the Chi-square test and that is what is covered here. The other reasonably well-known test is the Kolmogorov-Smirnov test.

The one-sample test procedure usually involves the question of goodness-of-fit, where a single randomly selected sample is compared against the known parameters of a given distribution. The most common parametric equivalent of this analytic technique is the t -test used to test the difference between the observed sample mean and an expected population mean. The Chi-square test can be used when the observed scores fall into two or more categories (nominal data) and can determine whether there is a significant difference between the observed frequencies and those expected to occur. The Kolmogorov-Smirnov test can be used with ordinal data to test for the degree of goodness-of-fit between an observed set of ranked scores and some theoretical distribution.

The number of categories is undetermined, but there must be at least two. Categorization of a variable is extremely common and the Chi-square test is often used. Some examples of variable categorization are:

1. Sex, divided into males and females
2. Income, divided into low, medium and high
3. Opinion when divided into yes, no, undecided.

The null hypothesis states what proportion of scores could be expected in each category of the population. Therefore, it is the null hypothesis that determines the expected frequency distribution.

The steps for the calculation of the one-sample Chi-square are as follows:

1. Determine the frequency for each category of the observed data. The sum of frequencies= N .
2. Determine the expected frequency for each category according to the null hypothesis.
3. Calculate the value Chi-square (χ^2); where:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

O =observed frequency.

E =expected frequency.

4. Calculate the number of degrees of freedom; where: d.f.= $k-1$.
5. Compare the calculated χ^2 value with the critical values given on the Chi-square critical value table. If the obtained χ^2 value is greater than the critical table value, the null hypothesis is rejected.

Where more than 20 per cent of the expected values are smaller than 5, it is necessary to combine categories in order to increase the expected count (by decreasing the value of k). Where only two categories exist ($k=2$), both expected values must be equal to, or greater than 5.

An *example* using the one-sample Chi-square test: a car rental company has five airport rental desks at five major international airports. The marketing manager considers one of the desks (No. 3) will outsell the other four by twice the number of sales, and has advised the company to increase staff and car numbers accordingly. The publisher decides to sample the market for the first 200 car rentals at each desk over a period of three days. The following ninety rentals are made:

| Desk | 1 | 2 | 3 | 4 | 5 | Total |
|---|----|----|----|----|----|-------|
| Rentals | 12 | 15 | 28 | 18 | 17 | 90 |
| The expected rentals as defined by the marketing manager are: | | | | | | |
| | 15 | 15 | 30 | 15 | 15 | |

The calculation of χ^2 :

$$\chi^2 = \frac{(12-15)^2}{15} + \frac{(15-15)^2}{15} + \frac{(28-30)^2}{30} + \frac{(18-15)^2}{15} + \frac{(17-15)^2}{15}$$

$$= 0.6 + 0.1333 + 0.26667 = 1.6 \quad \text{The degrees of freedom } (k-1) = 4.$$

The table of critical values of Chi-square for 4 degree of freedom at 95 per cent significance give a critical Chi-square=9.49. Since the obtained Chi-square value of 1.6 is less than 9.49 the null hypothesis is not rejected. Therefore, the company may conclude that the marketing manager correctly assessed the market for the five desks.

Some of the problems associated with Chi-square include, notably, the requirement to have at least a 5 expected frequency per category 80 per cent of the time, to have an overall sample exceeding 25 (small samples are not applicable to Chi-square) and no reliance on order (rank) in the null hypothesis. These problems can often be overcome by using the Kolmogorov-Smirnov technique provided the data could be ranked.

4.2 Paired two-sample non-parametric test

In the case of two samples, tests are made to determine whether the samples are drawn from the same, or from different populations. Most commonly one sample has been exposed to an independent variable (drug, travel, training, changed income, experiment, and so on) and the other sample forms a control group that has not been so exposed.

The two related samples are either matched (paired) or comprise the same objects (individuals). The parametric technique for analysing two related samples is the paired t -test discussed above. Apart from being distribution free, the non-parametric test equivalents have an additional advantage in that they do not require that all pairs of observations be drawn from the same population.

There are three common tests: the Sign test, the Wilcoxon Matched-pairs Signed Ranks test and the Mc Nemar test for the significance of changes. The Sign test is used to establish that two conditions are different, making only the one assumption that the independent variable under study has a continuous distribution. Where the magnitude as well as the difference between pairs is known, the Wilcoxon Signed Ranks test is a more powerful technique than the Sign test. The Mc Nemar test for the significance of changes is particularly useful in the 'before and after' research design where the individuals (objects) in the samples are both the same. For example, the Mc Nemar test may be used to examine the effect of a training session on a particular type of skill or the influence of an advertising promotion upon sales.

The test given here is the Wilcoxon Matched-pairs Signed Ranks test because of its relatively strong power. The Wilcoxon Signed Ranks test provides a more powerful

analysis than the Sign test, considering not only the direction of ranked differences between two related samples, but also the relative magnitude of the differences. More weight is given to a pair with a large difference than a pair with a small difference.

This analysis requires slightly more information than the ordinal scale of measurement because the data must be ranked, and the differences between pairs known. However, interval scale measurement is not required for each sample because the interval between observations is not needed. This scale falls in between the ordinal and interval scales and is termed the ordered metric scale.

The same example given in the paired sample test (section 3.7) can be used here. The two teams of aircraft cleaners are a matched paired sample set. The level of measurement is higher than required for this test but there is no longer an assumption of the normal distribution, which may well be more accurate. The steps for computation are as shown below.

1. For each related pair calculate the difference between the two scores and note the sign.
2. Rank the differences calculated in Step 1 regardless of their sign.
3. Disregard those differences equal to zero.
4. If two or more difference values are tied, assign the rank that is the average of the ranks that would have been assigned if the values had been slightly apart. That is, assign the average of the tied ranks. For example, $-2, -2, +2$ are tied differences. In the ranked array each pair would be assigned say the ranks 3, 4 and 5. Thus $3+4+5/3=4$. Ranks 1 and 2 would have been assigned to values below -2 and the next value would be ranked 6, because ranks 3, 4 and 5 have already been used and listed as the average 4.
5. Compute T , the Wilcoxon statistic as equal to the smaller sum of like-signed ranks.

For small samples, where the value of $N=(N1+N2)$ is less than 26, the T value is read off against the Wilcoxon T table. The null hypothesis states that the two samples are not different, that there is no difference in cleaning efficiency between the two flights. If the obtained T value is equal to or less than the critical T value in the Wilcoxon T table the null hypothesis can be rejected at the given level of significance.

For large samples, where the value of N is greater than 25, the distribution under T is approximately normally distributed with:

$$\text{mean equal to: } \mu = \frac{N(N+1)}{4}$$

$$\text{and standard deviation equal to: } \sigma = \frac{N(N+1)(2N+1)}{24}$$

$$\text{The calculation of } Z: Z = \frac{T - \mu}{\sigma}$$

The normal distribution table can be used to determine the probabilities associated with the occurrence under H_0 of values as extreme as the obtained Z .

For the example data in Table 1, we compare Table 2. The Wilcoxon Table critical value for $N=16$ at 0.05 significance two-tail is 30. Since T at 10.5 does not exceed the critical value of 30 the null hypothesis is not rejected. It is concluded that it is likely there is no difference in the time taken by the two teams.

Table 2.

| <i>Difference (D)</i> | <i>Rank of D</i> | <i>Ranks with less frequent sign</i> |
|-----------------------|------------------|--------------------------------------|
| -2 | -3 | |
| 4 | 6.5 | 6.5 |
| -2 | -3 | |
| -4 | -6.5 | |
| 1 | 1 | 1 |
| -3 | -5 | |
| -8 | -8 | |
| 2 | 3 | 3 |
| | | $T = 10.5$ |

4.3 Unpaired two-sample non-parametric test

As discussed previously, for the case of two paired samples, the most common parametric technique is the t -test. There are three common methods available, the Kolmogorov-Smirnov two sample test, the Chi-square two sample test, and the Mann-Whitney U-test. The Kolmogorov-Smirnov technique is used to determine whether or not the two samples have been drawn from either the same population, or populations with the same distribution. The Chi-square technique can be used to test when frequencies recorded into discrete categories are divided between two independent groups with a significant degree of difference. The Mann-Whitney technique is equivalent to the Kolmogorov-Smirnov test, but unlike that test is best suited to large rather than small samples.

The Chi-square test is very commonly used and can be output from SPSS from the Frequency Tables analysis. The Mann-Whitney U-test is the most powerful of the tests and is the method discussed here. The Mann-Whitney U-test can be used to test whether two independent groups (measured at least at the ordinal scale) have been drawn from the same population. Fundamentally, the test examines the combined, ranked array of two samples in order to measure the degree of randomness in the array. Logically, if the two samples come from the same population the array will exhibit randomness. Samples drawn from two different populations would tend to diverge toward different extremes of the columned ranked array, resulting in little random variability of the array.

The two samples do not have to be of the same size. All values are ranked in order of increasing size, and each is identified as coming from either $N1$ or $N2$. The value of the U statistic is the sum of the number of times that a score in the sample $N2$ precedes a score in the sample $N1$.

For example, given:

| N1 | N2 |
|----|----|
| 5 | 8 |
| 3 | 1 |
| 2 | 4 |
| 7 | 6 |

Step 1, rank values, maintaining sample identity:

$N =$ 1 2 3 4 5 6 7 8

2 1 1 2 1 2 1 2 – sample identity either $N1$ or $N2$

Step 2, count the number of times $N2$ scores precede an $N1$ score:

$$1+1+2+3=7$$

That is, for the first $N1$ score (ranked 2), one $N2$ score precedes. For the second $N1$ score (3), still only one $N2$ score precedes. For the third $N1$ score (5), two $N2$ scores precede. For the fourth $N1$ (7), three $N2$ scores precede. The total number of $N2$ scores preceding $N1$ scores is seven.

Where ties occur, the generally accepted procedure is to use the average rank for all the tied values. Since the sample distribution for U is known, the probability of obtaining a value as extreme as 7 can be determined.

For samples smaller than 21 there are U tables. In our sample above $N1=4$, $N2=4$ and $U=7$. From the U table for sample sizes less than 9 with the probability at 0.05 (two-tail) the nearest p value is 0.057 with $U=2$. Since $U=7$ is greater than U critical at 2 we cannot reject the null hypothesis. The value searched for in the U table (in the above case from the column represented by $N1=4$ and $N2=4$) is determined by the level of significance and whether the test is one- or two-tail. Thus, by example:

One-tail 95%, $\alpha=0.05$ one-tail 90%, $\alpha=0.10$

Two-tail 95%, $\alpha=0.025$ two-tail 90%, $\alpha=0.05$.

On finding the closest α value the U critical value can be read off. Note here that if U -obtained is less than U -critical the null hypothesis is rejected.

For samples larger than 20 the normal distribution can be used where:

$$\mu = \frac{N_1 N_2}{2}$$

and:

$$\sigma = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}$$

If ties occur the standard deviation becomes:

$$s = \sqrt{\left(\frac{N_1 N_2}{N'(N'-1)} \right) \left(\frac{N'^3 - N'}{12} \right) - \sum T}$$

where: $\sum T$ = summing the T 's overall groups of tied observations.
 $N' = N_1 + N_2$

$$Z = \frac{U - \frac{N_1 N_2}{2}}{\sqrt{\frac{(N_1)(N_2)(N_1 + N_2 + 1)}{12}}}$$

and: $U = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1$

where: R_1 = sum of ranks assigned to group N_1

and where: T = a correction factor for the number of ties:

$$T = \sum \frac{t^3 - t}{12}$$

where: t = number of observations tied for a given rank.

4.4 Multiple paired sample test

In the previous sections statistical techniques for analysis of differences between a single sample and a specified population, and of differences between two samples have been discussed. This section examines techniques for testing for a significant difference between three or more samples, drawn from the same population, or from populations with the same parameters.

There are two common tests for related samples: the Cochran Q test and the Friedman Two-way Analysis of Variance. The Cochran test is particularly useful for nominally measured data, while the Friedman Analysis of Variance is best suited to ordinal scale measurement and is the more powerful test.

Data is cast into a table of K columns and N rows where the K columns represent the number of samples (three or more). Since the samples are matched, the number of cases (N) is the same for each.

The matching may be achieved by studying one group of individuals under several conditions (where the conditions are the samples), or by obtaining several sets each with K matched subjects. In this example, suppose the design comprises three sets under four conditions, where $K=4$ and $N=3$. Each group of subjects (three of) contain four matched subjects and one is assigned to each of the four conditions:

| | | Conditions (samples), $K=4$ | | | |
|-------|-------|-----------------------------|---|---|---|
| | | 1 | 2 | 3 | 4 |
| $N=3$ | Set 1 | 7 | 3 | 8 | 1 |
| | Set 2 | 6 | 8 | 2 | 4 |
| | Set 3 | 9 | 1 | 4 | 2 |

The conditions could be four different brands of the same package tour to the same destination and the test could be to determine whether there is any perceived difference in quality. Each set from 1 to N contains 1 to K individuals who have been matched on basic variables such as age, sex, and socio-economic status. Each individual in a set is randomly assigned to rate the quality of the brand using a specified ordinal scale. The null hypothesis is that all the samples (columns) come from the same population because there is no difference in the perceived quality of the given four brands.

In the Friedman test the first step is to rank the scores in each row. The lowest score is ranked one, the highest is K . Thus, for the example data:

| | | Package brands, $K=4$ | | | |
|-------|-------|-----------------------|---|---|---|
| | | 1 | 3 | 4 | 5 |
| $N=3$ | Set 1 | 3 | 2 | 4 | 1 |
| | Set 2 | 3 | 4 | 1 | 2 |
| | Set 3 | 4 | 1 | 3 | 2 |
| | R_j | 10 | 7 | 8 | 5 |

If the null hypothesis were to be accepted, we would expect the ranks 1 to K (1 to 4 above) to be represented equally in each column. Consequently, the totals (R_j) for each column would be approximately equal.

The Friedman test analyses the rank totals (R_j) to estimate whether they differ significantly. The value calculated is symbolized as Chi-r-square and is distributed approximately as Chi-square with degrees of freedom = $K-1$. The formula is:

$$\chi_r^2 = \frac{12}{NK(K+1)} \sum_{j=1}^K (R_j)^2 - 3N(K+1)$$

where: N =number of rows

K number of columns

R_j =total of ranks in j^{th} column

Thus:

$$\begin{aligned} \chi_r^2 &= \frac{12}{(3)(4)(4+1)} [(10)^2 + (7)^2 + (8)^2 + (5)^2] - (3)(3)(4+1) \\ &= 12 / 60[238] - 45 = 2.6 \end{aligned}$$

The significance of Chi-r-square is determined by examining the Friedman probability table. For $K=4$ and $N=3$ a value of Chi-r-square=2.6 has a probability of 0.524. Should the null hypothesis be rejected at a 0.524 level of significance? This is not high and does not justify the rejection of the null hypothesis. It is likely the four brands of tours are not perceived to have different quality.

If N and/or K is larger than those shown in the Friedman table the Chi-r-square table of critical values is used, with degrees of freedom at $K-1$.

4.5 Multiple unpaired sample test

This section examines the case of non-parametric tests for the significance of the difference between three or more independent samples. There are two common tests: the

Chi-square multiple independent sample test and the Kruskal-Wallis One-way Analysis of Variance test.

The Chi-square test is suitable when the data falls into discrete categories measured at either the nominal or ordinal scale. There is no difference in the computation of the Chi-square multiple sample test and the two-sample test. The Kruskal-Wallis technique tests the null hypothesis that the K (multiple) samples come from the same population or from populations with the same parameters. The test requires data at the ordinal scale of measurement and assumes that the variable measured has an underlying continuous distribution. The Kruskal-Wallis test is done here because although it is less commonly seen than the Chi-square test, it is more powerful.

Since the samples are unpaired there is no need to have the same number of observations in each sample.

The first step of computation is to rank the observations for each sample in an increasing series with the smallest ranked as one, and the largest ranked N . Ranking is carried out across all the samples to form one series.

For example, suppose a large company wishes to test the efficiency of its management teams in three different countries. Each team has been created independently by three different subsidiaries of the corporation. The staff of each management team is set a simple efficiency test and a percentage score results:

| Team 1 | | Team 2 | | Team 3 | |
|---------------|------|--------------|------|--------------|------|
| Test | Rank | Test | Rank | Test | Rank |
| 80 | 10 | 82 | 12.5 | 65 | 3.0 |
| 92 | 18 | 83 | 14.0 | 82 | 12.5 |
| 95 | 19 | 71 | 7.0 | 91 | 16.0 |
| 91 | 16 | 75 | 9.0 | 70 | 5.5 |
| 97 | 20 | 73 | 8.0 | 60 | 1.0 |
| 99 | 21 | 81 | 11.0 | 65 | 3.0 |
| 91 | 16 | 70 | 5.5 | | |
| | | 65 | 3.0 | | |
| $R_j=120$ | | $R_j=70$ | | $R_j=41$ | |
| $R_j^2=14400$ | | $R_j^2=4900$ | | $R_j^2=1681$ | |
| $n_j=7$ | | $n_j=8$ | | $n_j=6$ | |

Each score above is ranked in increasing order and where ties occur, each score (tied) is given the mean of the rank for which it is tied.

The null hypothesis states that the samples come from the same population or identical populations with regard to means. Therefore, the null hypothesis is that there is no difference in the average efficiency rating of the three different management teams.

The Kruskal-Wallis statistic is termed H and is calculated by the formula:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^K \frac{R_j^2}{n_j} - 3(N+1)$$

where: K =number of samples
 n_j =number of cases in the j^{th} sample
 N =the total number of cases ($\sum n_j$)
 R_j =sum of the ranks in the j^{th} sample

The H statistic is distributed approximately as Chi-square with degrees of freedom= $k-1$.

Consequently:

$$\begin{aligned} H &= \frac{12}{21(22)} \left[\frac{14400}{7} + \frac{4900}{8} + \frac{1681}{6} \right] - 3(22) \\ &= 12/462 [2057.1429 + 612.5 + 280.16667] - 66 \\ &= 10.61843 \end{aligned}$$

The sign of H is ignored.

There are several tied ranks in the previously presented data. Since the value of the H statistic can be slightly affected by ties it is necessary to correct for this factor via the formula:

$$1 - \frac{\sum T}{N^3 - N}$$

where: $T=t^3-t$ (where t is the number of tied observations in a tied group of scores)
 N =is the total number of cases for all samples
 $\sum T$ =to sum T for all groups of ties

The value of H is divided as:

$$1 - \frac{H}{N^3 - N}$$

The effect of the division is to increase the size of H slightly and this in turn makes H more statistically significant than it would otherwise have been.

Correcting for ties in the example, it is first necessary to recognize the tied groups:

| | | |
|-------------------|--------------|----|
| 65 thrice at rank | $3=T=27-3=$ | 24 |
| 70 twice at rank | $3=T= 8-2=$ | 6 |
| 82 twice at rank | $9=T= 8-2=$ | 6 |
| 91 thrice at rank | $11=T=27-3=$ | 24 |
| | | 60 |

Therefore, the correction factor for ties:

$$1 - \frac{\sum T}{N^3 - N} = 1 - \frac{60}{9261 - 21}$$

The H now becomes:

$$H = \frac{H}{1 - \frac{\sum T}{N^3 - N}} = \frac{10.61843}{0.9935065} = 10.6878$$

When the number of cases for the three or more samples is greater than five or there are more than three samples, the Chi-square table of critical values is used to determine the statistical significance of H at $K-1$ degrees of freedom. At 0.5 significance critical Chi-square at 2 degrees of freedom is 5.99.

If (H -obtained) exceeds critical Chi-square, reject H_0 .

Because 10.69 is greater than 5.99 the null hypothesis is rejected in favour of the alternative hypothesis. The conclusion is that the management teams do differ from country to country in terms of efficiency at 95 per cent statistical significance. For a higher level of significance at 99 per cent, critical Chi-square is 9.12 (at 2 degrees of freedom), and with H -obtained at 10.69, the null hypothesis can still be rejected.

If the number of cases in each of only three samples is five or less, the Chi-square approximation to the sampling distribution of H becomes less reliable and the critical value should be derived from the Kruskal-Wallis table. This table gives the possible sample sizes in the first columns, followed by values of H and p (probability under H_0) in the columns to the right.

5. Cross-cultural behaviour: example analysis

The following example analysis is part of a paper titled “Cultural Differences between Asian Tourist Markets and Australian Hosts: Part 1”, by Y. Reisinger and L. Turner reprinted from the *Journal of Travel Research* Vol. 40, number 3, 2002, pp. 295–315 with permission from Sage Publications. This study provides a good example of the use of the Mann-Whitney U-test on large samples.

The main research objectives of this study are to:

- 1) identify the key cultural differences between the Asian tourist markets and the Australian host population, as a representative of western culture,
- 2) determine the key dimensions of these differences, and their indicators, and
- 3) identify major cultural themes that should be included in every promotional strategy aiming at the Asian tourist market.

Only objective one is analysed here. The remaining objectives involve the use of Principal Components Analysis, a technique discussed in Chapter 7 of our book *Cross-Cultural Tourism Behaviour: Concepts and Analysis* (2003).

Culture, in this paper, refers to a stable and dominant cultural character of a society shared by most of its individuals, and remaining constant over long periods of time. Culture does not refer to the subcultures of many ethnic groups living in a society that may be distinguished by religion, age, geographical location or some other factor, nor the individual's character that can be influenced by environmental forces and easily changed over time.

The two distinct groups, tourists and hosts, were chosen for the study because these groups are the major tourism players. Hosts in this study are nationals of the visited country who are employed in the tourism industry and provide a service to tourists (e.g., front office employees, bus drivers, shop assistants, waitresses, custom officials). Knight (1996) referred to hosts as those who provide tourism services (e.g. shelter, accommodation and food), are in direct contact with tourists, and derive direct benefits from the tourists. Nettekoven (1979) referred to them as 'professional hosts' who are employed in the places of most frequent tourist visitation. These places offer maximum opportunities for a direct tourist-host contact. As a result, hosts represent the first contact points with tourists. Consequently, cross-cultural differences in the interpersonal interaction in the tourism context are most likely to be apparent in these two groups, tourists and hosts.

A sample of 618 Asian tourists visiting the Gold Coast region, Australia's major tourist destination, were personally interviewed in their own language, alongside 250 Australian service providers. Asian tourists were surveyed in a variety of locations on the Gold Coast, where there is a large concentration of Asian tourists. The total population of Asian tourists was divided into five mutually exclusive and exhaustive strata (Asian language groups), which represented distinct Asian tourist markets: Indonesian, Japanese, South Korean, Mandarin and Thai. The selection of cultural language groups was based on the statistical data showing the arrivals of international tourists to Australia from major countries of origin. A representative sample of respondents was chosen from each stratum. The sample elements were not selected in proportions that reflected the size of each major Asian tourist market on the Gold Coast, rather the emphasis was on getting a

maximum number of respondents from different language groups. An attempt was, however, made to choose respondents from a wide variety of socio-demographic backgrounds. This was done to ensure the samples were representative of the central tendency of their culture. The respondents were equivalent in their characteristics in terms of the purpose of travel and length of stay on the Gold Coast. Australian hosts were randomly selected from a variety of sectors of the tourism and hospitality industry on the Gold Coast such as accommodation, transportation, or entertainment in the same time period. Again, disproportionate samples were taken from each stratum because proportionate samples would have resulted in small samples. The study was conducted over the period 1994 to 1995.

Five measurement groups of cultural values, rules of social behaviour, perceptions of service, forms of interaction, and satisfaction with interaction were measured by a structured questionnaire. Personal values were measured using the Rokeach Values Survey (RVS) (Rokeach, 1973). The RVS was assessed as the best available instrument for measuring values, because it is 'based on a well-articulated conceptualization of value' and is successful in 'finding specific values that differentiate various political, religious, economic, generation and cultural groups' (Braithwaite and Law, 1985, p. 250). The RVS has been used in numerous studies to measure human values (e.g., Feather, 1980a,b,c; 1986a,b; Ng *et al.*, 1982) and it has identified cultural differences between countries, including differences between western and Asian countries. The 36-item RVS scale produced a Cronbach Alpha value of 0.9497 that indicated that the RVS was a very reliable instrument. In respect of validity, all items were adapted from the RVS (Rokeach, 1973), which measured human values. Rokeach (1973) selected only those values, which were considered to be important across culture, status and sex (Rokeach, 1971). The respondents were asked to indicate the importance of specific values by rating them on a 6-point scale, ranging from 1 (not important) to 6 (extremely important).

Rules of social interaction were measured using Argyle *et al.*'s (1986) list of thirty-four rules of social behaviour, which has also been widely used and assessed as a reliable and valid measure of the rules of social relationships. Only the rules that were applicable to

tourist-host interaction were included and the rules that governed family invitations, social visitation, or sexual activity were excluded. The rules specific to Asian cultures such as the clear indication of intentions, conforming to rules of etiquette and the status of the other person, having a sense of shame, and avoiding embarrassment were included. These rules were chosen from the literature on interpersonal relations in Asian cultures, and focus group discussions with Asian students. The Alpha Cronbach was 0.9048 indicating that the instrument was highly reliable. The respondents were asked to indicate the importance of specific rules on a 6-point scale, ranging from 1 (not important) to 6 (extremely important).

Perceptions of service were measured using a 22-item SERVQUAL instrument (Parasuraman's *et al.*, 1985, 1988), which has also been widely applied in empirical studies in various disciplines, including the hospitality and tourism industry, and assessed as reliable and highly valuable (Albrecht, 1992; Fick and Ritchie, 1991; LeBlanc, 1992; Luk *et al.*, 1993; Saleh and Ryan, 1991). However, the SERVQUAL scale was modified by eliminating all the positive and negative statements that made a comparison of responses difficult, including the word 'should'. Only the words describing service (adjectives) were used. Also, the scale was supplemented by additional items to reflect the distinctive features of a high quality service as perceived by Asian visitors (e.g., the ability of hosts to speak an Asian language, treat tourists as guests, know Asian culture and customs). These distinctive items of service were identified on the basis of the literature review about service quality. The responses were measured on a 6-point scale, ranging from 1 (least important) to 6 (extremely important).

Tourist-host interaction was measured using a list of various forms of interaction such as playing sport together, having a close relationship, or sharing a meal. These items were adapted from several studies' direct and indirect measures of social contact (Black and Mendenhall, 1989; Feather, 1980b; Gudykunst, 1979; Kamal and Maruyama, 1990; McAllister and Moore, 1991; Vassiliou *et al.*, 1972). Again, the responses were measured on a 6-point scale ranging from 1 (least preferred) to 6 (most preferred).

Satisfaction with interaction was measured using a list of various components of satisfaction with social interaction such as satisfaction with language spoken, conversation, or time spent together. The items were measured on a 6-point scale ranging from 1 (dissatisfied) to 6 (extremely satisfied).

The instrument was translated into Asian languages and back translated to the English language by a professional translating agency. The instrument was pre-tested twice in two pilot studies to ensure it was clear and understandable: once on a sample of twenty Australian tourists and twenty providers, and the second time on a sample of fifty Asian tourists. Professional native Asian language speaking interviewers were hired to collect data. In total the data collection process resulted in surveying 870 respondents: 250 Australian hosts and 618 Asian tourists from five language groups (106 Indonesian, 108 Japanese, 172 South Korean, 130 Mandarin-speaking, and 102 Thai).

The Mann-Whitney U-test is used instead of a z-test of the means of the variables, because not all variables were (or could be transformed to be) normally distributed. Analysis is done on the SPSS system. The results of the Mann-Whitney U-test identified significant differences in all five measurement groups (cultural values, rules of behaviour, perceptions of service, forms of interaction, satisfaction with interaction) between the total Asian and Australian populations with seventy-three out of 117 (62.4 per cent) areas of measurement showing significant cultural differences between the Australian and the total Asian samples (refer to Tables 3 and 4).

Table 3. Number of the significant differences between Australian hosts and Asian language groups

| <i>Group indicators</i> | <i>Max</i> | <i>Australian Total Asian</i> | <i>Australian Indonesian</i> | <i>Australian Japanese</i> | <i>Australian South Korean</i> | <i>Australian Mandarin</i> | <i>Australian Thai</i> |
|---------------------------|------------|-----------------------------------|----------------------------------|--------------------------------|--|--------------------------------|----------------------------|
| Cultural values | 36 | 18 | 14 | 26 | 26 | 12 | 14 |
| Rules of interaction | 34 | 22 | 24 | 22 | 18 | 18 | 20 |
| Perceptions of service | 29 | 23 | 18 | 23 | 20 | 15 | 24 |
| Forms of interaction | 11 | 7 | 5 | 8 | 7 | 6 | 3 |
| Satisfaction | 7 | 3 | 3 | 4 | 3 | 2 | 3 |
| Total | 117 | 73 | 64 | 83 | 74 | 53 | 64 |

Table 4. The Mann-Whitney U-test of the differences in cultural values, rules of social interaction, perceptions of service, forms of interaction and satisfaction with interaction between Australian hosts ($n=250$) and total Asian tourists ($n=618$)

| Measurement Group | z-test | | z-test | | z-test |
|---|------------|-------------------------------|------------|------------------------------------|------------|
| <i>Cultural values</i> | | | | | |
| a comfortable life | -1.4703 | pleasure | -2.0275* | forgiving | -0.1657 |
| an exciting life | -1.1237 | salvation | -5.5966*** | helpful | -0.0107 |
| a sense of accomplishment | -1.8044 | self-respect | -6.7442*** | honest | -3.1904** |
| a world of peace | -1.4887 | social recognition | -1.6728 | imaginative | -1.0280 |
| a world of beauty | -0.5049 | true friendship | -2.4880* | independent | -3.7725*** |
| equality | -2.8793** | wisdom | -0.0306 | intellectual | -2.1053* |
| family security | -2.8891** | ambitious | -2.2367* | logical | -0.7708 |
| freedom | -6.5026*** | broad-minded | -0.4982 | loving | -2.6717** |
| happiness | -3.2357** | capable | -1.1952 | obedient | -1.0057 |
| inner harmony | -2.1927* | cheerful | -1.7364 | polite | -3.3062*** |
| mature love | -3.4440*** | clean | -2.7369** | responsible | -0.0102 |
| national security | -1.3962 | courageous | -1.1777 | self-controlled | -2.4574* |
| <i>Rules of social interaction</i> | | | | | |
| should address by first name | -5.6758*** | should take others' time | -3.2099** | should have a sense of shame | -5.6699*** |
| should shake hands | -6.8501*** | should develop relationship | -5.6425*** | should ask for financial help | -0.6472 |
| should look in the eye | -9.5694*** | should touch the other person | -1.2410 | should ask for advice | -1.4431 |
| should think about own needs | -0.4808 | should acknowledge birthday | -2.5842** | should ask personal questions | -5.2496*** |
| should express opinion | -0.5112 | should be neatly dressed | -1.2572 | should respect others' privacy | -6.3187*** |
| should show intentions clearly | -5.8349*** | should conform to etiquette | -3.9182*** | should show interest in other | -7.1359*** |
| should obey instructions | -6.3291*** | should conform to status | -2.0481* | should show respect to other | -8.1729*** |
| should criticize in public | -7.0109*** | should swear in public | -5.4757*** | should show affection | -1.2262 |
| should compliment other | -7.9697*** | should not make fun of other | -0.3453 | should show emotions | -3.9084*** |
| should apologize if not at fault | -0.4428 | should avoid arguments | -6.6947*** | should talk about sensitive issues | -0.8220 |
| should compensate if at fault | -3.0403** | should avoid complaining | -2.9989** | | |
| should repay favours | -0.6371 | should avoid embarrassment | -0.6598 | | |
| <i>Perceptions</i> | | | | | |
| neatly dressed | -5.2277*** | polite | -3.0413** | keep tourists informed | -4.8190*** |
| perform service required | -6.2694*** | respectful | -5.0529*** | listen to tourists | -5.7517*** |
| responsive to tourists' needs | -6.6850*** | considerate | -3.2617** | need adequate explanations | -3.2087** |
| require help | -1.2015 | treat as guests | -4.8425*** | understand tourists' needs | -5.6261*** |
| prompt service | -7.5167*** | trustworthy | -3.7197*** | anticipate tourists' needs | -5.0024*** |
| service on time | -7.5816*** | confident | -0.4764 | need individualized attention | -4.2763*** |
| find solutions to problems | -5.0824*** | concerned about welfare | -5.4387*** | know Asian culture/customs | -0.4982 |
| answer questions | -4.5351*** | approachable | -3.5354*** | speak Asian languages | -0.3821 |

| | | | | | |
|------------------------------|------------|----------------------------------|------------|---------------------------------|-----------|
| provide accurate information | -6.4345*** | easy to find | -7.7714*** | know Australian history/culture | -0.7542 |
| friendly | -0.9894 | easy to talk to | -2.9905** | | |
| Forms of interaction | | | | | |
| invite home | -9.8994*** | have close relationship | -7.8447*** | exchange gifts | -2.4978* |
| play sport together | -3.6514*** | share a meal | -5.6582*** | have business contact only | -2.6874** |
| share recreation facilities | -0.5862 | chat on a street | -0.0996 | have no contact at all | -0.6818 |
| take part in family parties | -7.9068*** | talk in shops | -0.5012 | | |
| Satisfaction | | | | | |
| with tourists/hosts | -4.9786*** | with time spent together | -3.4804*** | with service provided | -0.7144 |
| with conversation | -0.2667 | with language skills | -1.9804* | | |
| with friendship | -0.8191 | with knowledge of others culture | -0.3272 | | |

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The Mann-Whitney U-test also identified significant differences in all five measurement groups between the individual Asian language groups and the Australian sample. The Japanese were the most distinct from the Australian hosts. Out of 117 areas of measurement, eighty-three significant differences were found between Australian hosts and Japanese tourists, seventy-four between the South Korean sample, sixty-four between the Indonesian and Thai samples and fifty-three between the Mandarin sample (refer to Table 3).

Summary

Hypothesis testing is useful for comparing sample means from tourism surveys. Difference between parametric and non-parametric analysis were described. The text explained how to conduct hypothesis testing and when to use the z - and t -tests. One- and two-tailed tests relate to the direction of the comparison between means. Single sample means test compares a single sample mean with a population mean. Two- sample means test compares two sample means against each other. Two samples can be paired (matched or same individual cases) or unpaired (independent cases). Non-parametric tests can be useful where a normal distribution cannot be assumed. One-sample non-parametric Chi-square test is useful. A paired two-sample test example is the Wilcoxon Signed Ranks test. The unpaired two-sample non-parametric test is the Mann-Whitney U-test. The multiple paired sample test is the Friedman Two-way Analysis of Variance test and the

multiple unpaired test the Kruskal-Wallis test. Example application to tourism culture analysis using the Mann-Whitney U-test was presented.

Discussion points and questions

1. Explain the difference between parametric and non-parametric statistical tests.
2. Why is a quantitative procedure useful in testing the relationship between samples?
3. State three assumptions of parametric statistical tests.
4. What is the difference between unpaired and paired samples?
5. If the null hypothesis is accepted, why is the significance level not used?
6. What is post hoc error in hypothesis testing?
7. Create an example of a problem that would need to be analysed by the Friedman Two-way Analysis of Variance test.
8. Define Type II error.

Further reading

- Brownlee, K. (1965) *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.
- Dietrich, F. and Shafer, N. (1984) *Business Statistics: An Inferential Approach*. San Francisco: Dellen Publishing Co.
- Harrison, S. and Tamaschke, R. (1993) *Statistics for Business, Economics and Management*. New Jersey: Prentice-Hall.
- Hays, W. (1981) *Statistics*. 3rd edition. New York: Holt, Rinehart and Winston.
- McClave, J. and Benson, P. (1991) *Statistics for Business and Economics*. 5th edition. New York: Macmillan.
- Petocz, P. (1990) *Introductory Statistics*. New York: Nelson.
- Siegel, S. (1956) *Nonparametric Statistics for the Behavioural Sciences*. Tokyo: McGraw Hill.
- Sincich, T. (1992) *Business Statistics by Example*. New York: Macmillan.
- Swift, L. (1997) *Mathematics and Statistics for Business Management and Finance*. London: Macmillan.