Available online at www.sciencedirect.com
ScienceDirect
British Journal of Oral and Maxillofacial Surgery 46 (2008) 38–41
BRITISH Journal of Oral and Maxillofacial Surgery
www.bjoms.com

Which is the correct statistical test to use?

Evie McCrum-Gardner *

Schools of Nursing & Health Sciences, Health & Rehabilitation Sciences Research Institute, University Of Ulster, Shore Road, Newtownabbey BT37 0QB, United Kingdom

Accepted 6 September 2007
Available online 24 October 2007

Abstract

This paper explains how to select the correct statistical test for a research project, clinical trial, or other investigation. The first step is to decide in what scale of measurement your data are as this will affect your decision—nominal, ordinal, or interval. The next stage is to consider the purpose of the analysis—for example, are you comparing independent or paired groups? Several statistical tests are discussed with an explanation of when it is appropriate to use each one; relevant examples of each are provided. If an incorrect test is used, then invalid results and misleading conclusions may be drawn from the study.

Keywords: Statistical test; Parametric; Nominal; Ordinal; Interval

Introduction

In this paper I explain how to select the correct statistical test depending on the type of data and purpose of the analysis. When choosing the appropriate statistical test, the first step is to decide what scale of measurement your data is as this will affect your decision. The next stage is to consider the analysis required—for example, are you comparing independent or paired groups? I discuss when it is appropriate to use a range of parametric and non-parametric tests including examples of each.[1] If an incorrect test is used, then invalid results and misleading conclusions may be drawn from the study.

Definitions of elementary statistical concepts and a useful statistical glossary are provided in "The Statistics Homepage" (http://www.statsoft.com/textbook/stathome.html).[2]

Scales of measurement

There are three main types:

Nominal: categories but no order such has sex (male/female), marital status (single/married/divorced/widowed), location of lesion (tongue, lip, floor of mouth, palate and so on).

Ordinal: ordered categories such as pain (mild/moderate/severe), Likert scale (strongly disagree/disagree/neutral/agree/strongly disagree), stage of tumour (grades I–IV), visual analogue scale (VAS).

Interval: including age (years), weight (kg) or length of osteotomy (cm).

It is important to distinguish between a rating scale (such as the shoulder scale shown in Table 1) and an interval scale such as weight. For weight, the difference between 0 and 30 kg is the same as between 70 and 100 kg. However, this would not apply to the shoulder scale. Similarly for Likert scales such as the 5-point scale, the difference between 1 and 2 is not necessarily the same as between 4 and 5.

Parametric assumptions

It is important to examine the distribution of interval-scale data to check if they are normally distributed—that is, bell-shaped, symmetrical about the mean. Examples of an

* Tel.: +44 2890366852.
E-mail address: ee.gardner@ulster.ac.uk.

0266-4356/$ – see front matter © 2007 The British Association of Oral and Maxillofacial Surgeons. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.bjoms.2007.09.002

Table 1
The University of Washington Quality of Life questionnaire shoulder domain[6]

No problem with shoulder	100
Shoulder stiffness with no effect on activity or strength	70
Pain or weakness in the shoulder that has caused a change in work or hobbies	30
An inability to work or do hobbies because of problems with the shoulder	0

interval-scale variable that is approximately normally distributed (diastolic blood pressure) and one that is skewed (tricyclerides) are shown in Fig. 1. In addition to the histogram, normality tests such as Kolmogorov–Smirnov and Shapiro–Wilks can be used to decide if the distribution is normal. For more detailed information on parametric assumptions refer to Bland.[1]

General points about non-parametric methods

Non-parametric methods are typically less powerful and less flexible than their parametric counterparts. Parametric methods are preferred if the assumptions can be justified. Sometimes a transformation can be applied to the data to satisfy the assumptions, such as log transformation. Siegel gives more information on the non-parametric tests discussed below.[3]

Comparison of two groups

First of all we will consider the situation when two groups are to be compared: firstly independent groups and secondly paired (before/after) groups. Table 2 gives the appropriate test to use depending on whether the data are interval and approximately normally distributed, ordinal, interval and skewed, or nominal.

Table 2
Selecting the appropriate test for comparisons between two groups

Scale of measurement	Independent samples	Paired samples
Interval scale (parametric assumptions satisfied)	Independent samples *t*-test	Paired samples *t*-test
Ordinal scale or interval scale (parametric assumptions not satisfied)	Mann–Whitney *U*-test	Wilcoxon signed rank test
Nominal scale		
Two categories	χ^2-Test for 2×2 table	McNemar's test
C categories ($C>2$)	χ^2-Test for $2 \times C$ table	–

We will now look in more detail at each of these tests including an example.

Two independent groups

Independent samples t-test

The independent samples *t*-test is used to compare sample means from two *independent* groups for an *interval*-scale variable when the distribution is approximately normal.

Example: to compare the mean duration of follow-up (months) between location of the teeth (mandibular, maxillary) if duration is approximately normally distributed.

Mann–Whitney U-test

The Mann–Whitney test is used to compare two *independent* samples when data are either *interval* scale but assumptions for *t*-test (normality) are not satisfied, or *ordinal* (ranked) scale. The hypothesis being tested is whether the two medians are equal (as opposed to two means in the independent-samples *t*-test).

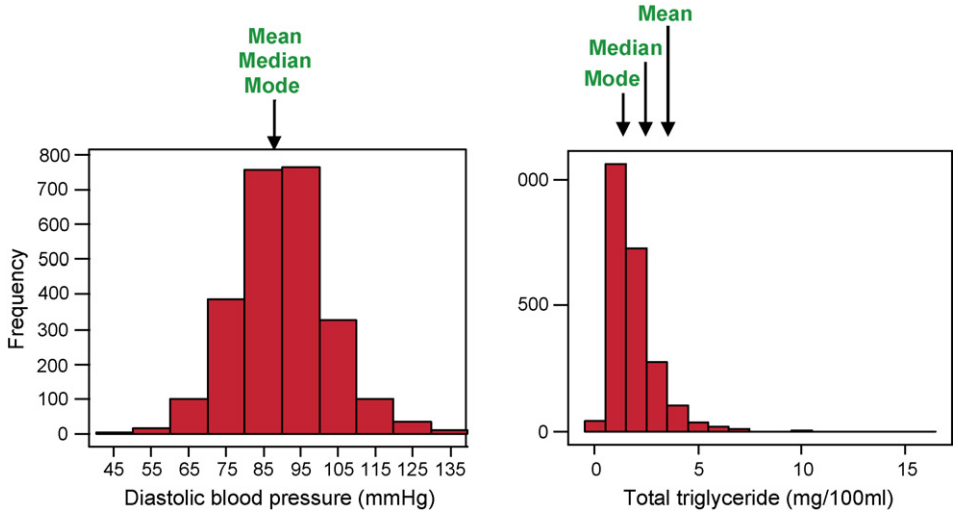

Fig. 1. Histograms to examine the distribution for normality: diastolic blood pressure normally distributed, tricyclerides skewed.

Example: to compare length of osteotomy (cm) (if skewed) between surgical operator (consultant, non-consultant).

Chi-square (χ^2) test

The chi-square (χ^2) test is used to compare *proportions* between two or more independent groups or investigate if there is any association between two *nominal*-scale variables. The data can be presented as a contingency table with one of the variables as rows and the other as columns in the table. If the sample size is less than 20 in a 2×2 table then the Fisher's exact test should be used.

Example: to examine the association between treatment group (drug A, drug B, placebo) and prognosis (alive/dead) or to compare the proportion of those alive between the three treatment groups.

Two paired groups

Paired samples t-test

The paired *t*-test is used to compare two sample means where there is a one-to-one correspondence (or *pairing*) between the samples. It is appropriate for an *interval*-scale variable when the distribution (of within-pair differences) is approximately normal.

Example: to compare the maximum interincisal distance (mm) before and after operation (if the distance is approximately normally distributed).

Wilcoxon signed rank test

The Wilcoxon signed rank test is used to compare two *paired* samples when data are either *interval* scale but assumptions for the paired *t*-test (normality of within-pair differences) are not satisfied or *ordinal* (ranked) scale. The hypothesis being tested is whether the median difference is zero (as opposed to mean difference in the paired *t*-test).

Example: to compare the amount of bone removed (%) before and after treatment (small sample size so amount of bone skewed).

McNemar's test

McNemar's test is used to compare two *paired* samples when the data are *nominal* and *dichotomous*.

Example: to investigate the outcome (success/failure) of a paired experiment using two drugs.

Comparisons of more than two groups

Table 3 gives the appropriate test to use when comparing more than 2 groups.

Independent groups

One-way analysis of variance (ANOVA)

If there are more than two *independent* groups being compared the one-way ANOVA is used if the parametric assumptions are satisfied—that is, *interval*-scale variable approximately normally distributed.

Example: to compare mean bone density (%) between pre-cancerous lesions (mild, moderate, severe dysplasia) if the density is approximately normally distributed.

Table 3
Selecting the appropriate test for comparisons between more than two groups

Scale of measurement	Independent samples	Paired samples
Interval scale (parametric assumptions satisfied)	One-way ANOVA	Repeated measures analysis of variance
Ordinal scale or interval scale (parametric assumptions not satisfied)	Kruskal–Wallis one-way ANOVA	Friedman's test
Nominal scale	χ^2-Test for RXC table	Cochran's Q

R rows, C columns.

Kruskal–Wallis one-way ANOVA

The non-parametric alternative is Kruskal–Wallis one-way ANOVA and is used for *ordinal* data, or an *interval*-scale variable, which are not normally distributed.

Example: to compare the shoulder scale (Table 1) between three age groups.

Chi-square (χ^2) test

Again the chi-square test is used for *nominal* data with R rows and C columns in a contingency table where $R > 2$ and $C > 2$.

Example: to examine the association between treatment group (drug A, drug B, placebo) and obesity (not overweight, overweight, obese).

Paired/related groups

Repeated-measures analysis of variance

If there are more than two *related* groups (one-to-one correspondence) being compared, repeated-measures analysis of variance is used if the parametric assumptions are satisfied—that is, the *interval*-scale variable approximately normally distributed.

Example: to compare blood pressure between the 4 time periods—before treatment, 1 day after treatment, at 1 week, and at 1 month (if blood pressure is approximately normally distributed).

Friedman's test

The non-parametric alternative is Friedman's test and is used for *ordinal* data or an *interval*-scale variable that is not normally distributed.

Example: to compare the shoulder scale between the 3 time periods—before operation, after 7 days, and after 14 days.

Cochran's Q-test

Cochran's *Q* is used for *nominal dichotomous* data when there are more than two *related* groups.

Example: to compare the proportion of pass/fail in a group of dental students for a series of six examinations.

Association between two interval or ordinal variables

The *correlation coefficient* is used to investigate the association between two *interval or ordinal* variables. If both variables are interval and approximately normally distributed then the *Pearson's* product-moment correlation coefficient is used. If either variable is ordinal or interval and skewed then the non-parametric equivalent is *Spearman's* rank correlation coefficient.

Example of Pearson's: to examine the relationship between length of osteotomy (cm), amount of bone removed (%), cost of treatment (£) (all approximately normally distributed).

Example of Spearman's: to examine the relationship between survival time (years, skewed) and each of the following: age, weight, maximum interincisal distance (mm).

This can be extended to *multiple regression* analysis with an *interval*-scale response (dependent) variable and several predictor (independent) variables. *Logistic regression* is used when the response variable is *dichotomous,* that is two categories only. Both multiple and logistic regression are covered in Bland.[1]

Example of multiple regression: to investigate which factors (length of osteotomy, amount of bone removed) are associated with cost of treatment, taking age and body mass index into account.

Example of logistic regression: to investigate which factors (sex, age, number of teeth extracted) are associated with having postoperative complications (yes/no).

Statistical software

All the above tests can be done using a number of statistical packages including Statistical Package for the Social Sciences (SPSS). Pallant[4] gives clear, step-by-step instructions how to perform each statistical test.

Power, *P* values, and percentages

Before undertaking a research study it is important to make a sample size calculation so that the study will have sufficient power to detect significant differences.[5] In addition to this it is often necessary to combine categories so that there are sufficient numbers in each group for comparison. For example, for the chi-square test to have valid results there needs to be an expected frequency in each cell of more than 5. Specific guidelines on cell size for both the chi-square and Fisher's exact tests are given on page 110 of Siegel.[3]

When quoting percentages, it is essential to also quote the numerator and/or the denominator so that it is clear how the percentage has been calculated. Percentages based on small numbers such as less than 10 are not meaningful.

A significance level (*P* value) is considered significant if it is less than 0.05. It is sufficient to quote *P* values to two decimal places if greater than 0.01. However, if the *P* value is very small then $P < 0.0001$ should be used.

References

1. Bland M. *An introduction to medical statistics*. 3rd ed. Oxford University Press; 2000.
2. The Statistics Homepage http://www.statsoft.com/textbook/stathome.html.
3. Siegel S. *Nonparametric statistics for the behavioral sciences*. 2nd ed. London: McGraw-Hill; 1988.
4. Pallant J. *SPSS survival manual*. 2nd ed. Maidenhead: Open University Press; 2005.
5. Interactive Statistical Calculation Pages http://www.statpages.org/#Power.
6. Laverick S, Lowe D, Brown JS, Vaughan ED, Rogers SN. The impact of neck dissection on health-related quality of life. *Arch Otolaryngol Head Neck Surg* 2004;**130**:149–54, http://archotol.ama-assn.org/cgi/reprint/130/2/149.pdf.