3: 201–212 (2005)
DOI:10.1002/msc.11

Main article

An introduction to medical statistics for health care professionals: Basic statistical tests

Elaine Thomas PhD, MSc, BSc
Senior Lecturer in Biostatistics, Primary Care Sciences Research Centre, Keele University, North Staffordshire, UK

Abstract

This article, the third and final article in the series, aims to give health care professionals (HCPs) a sound and helpful introduction to medical statistics (Thomas, 2004, 2005). A brief summary of the content of the previous articles is given in Table 1. The current article will cover the area of basic statistical tests with the aim of guiding HCPs to the correct test for a particular research question and dataset. The article will not go into great depth of the formal methods of calculation required for all the tests covered but I would suggest that the reader refer to standard textbooks (Jordan et al., 1998; Swinscow, 1998; Altman, 1994; Bland, 2000), the help sections of statistical packages (SPSS or Stata), or consult a statistician.

For ease of reference within the article the tests have been grouped by the data type, i.e. numerical or categorical. Further separation within each data type has been carried out depending on the number of groups being compared, whether the groups are independent, the size of the sample, and, in the case of numerical data, the distribution of the variable. For quick reference, two other tables are also presented which summarize which analysis methods should be used in each situation. Copyright © 2005 John Wiley & Sons, Ltd.

Key words: statistical association, correlation, regression

Numerical data

This section will concentrate on the tests available for examining numerical variables (Table 2). As seen previously for summary measures, the distribution of the values under examination determines the method of analysis (Thomas, 2004). Tests applied to variables where an assumption is made about their distribution, e.g. the values are normally distributed, are also known as parametric methods, and tests where no assumptions are made about the

TABLE 1: Topic areas covered in the first two papers in the series

Paper 1	Paper 2
Descriptive statistics	Hypothesis tests and estimation
Data types	Hypothesis testing
Summary measures	p-values
Data presentation	Confidence intervals
	Statistical vs clinical significance
	Generalizability

TABLE 2: Tests for comparison of numerical data across two distinct groups

	Are the groups independent or paired?	
What is the distribution of the data?	Independent	Paired
Parametric	Two sample t-test	Paired t-test
Non-parametric	Mann-Whitney U-test	Wilcoxon matched pairs

distribution of the variable, e.g. values are from skewed distribution, are known as non-parametric methods. Steps should be taken to assess the distribution of the data prior to any data analysis. To assess for normality the data should be plotted and compared to the standard normal shape (symmetrical, bell-shaped curve). Moreover, for the normal distribution, the mean, median and mode all take approximately the same value. More formal mechanisms to assess normality are available in most statistical packages.

Throughout this section the following abbreviations will be used: n = sample size and SD = standard deviation.

Two independent groups of observations

One of the most commonly used statistical tests is that comparing two independent groups of observations, i.e. two groups of individuals that are not related to one another. With independent data the interest lies in the average difference between the groups and the variability in these differences, i.e. the between-subject variability.

Data are available from a cohort population-based study of older adults with chronic knee pain – age and gender, body mass index (BMI), and a measure of knee functioning (0 = no problems to 68 = worst problems) (WOMAC; Bellamy, 1996). Participants were recruited and followed up three years later. One question of interest to the researchers is whether levels of functioning are the same at recruitment between males and females.

Parametric methods

If both sets of data can be assumed to be sampled from the normal distribution and the variability in the two datasets is similar then the difference in means between the groups is the effect of interest. In this situation, the two independent sample t-test is carried out which results in the test statistic, *t*, which can then be compared to standard tables for the relevant sample size to determine the associated p-value. The magnitude of the test statistic, *t*, is dependent on three factors: (1) the difference in means between the groups, (2) the variability of the data in each group, and (3) the two sample sizes. For a particular sample size, a greater value of *t* indicates a smaller p-value and hence more evidence to reject that there is no difference between the groups.

So, using terminology previously described (Thomas, 2005) for the example above:

H_0: Mean level of knee function, as measured by the WOMAC, is the same in males and females.

H_A: Mean level of knee function, as measured by the WOMAC, is not the same in males and females.

In the study data there are 60 females with chronic knee pain with a mean (SD) WOMAC score of 37.2 (17.3) and 53 males with a mean (SD) WOMAC score of 32.1 (14.5). Hence the best estimate of the difference in WOMAC function scores between female and male chronic knee sufferers is 5.1. The test statistic for this comparison is calculated as $t = 1.69$ and when compared to tabled values indicates $p = 0.09$.

As described previously, the p-value indicates the amount of evidence available to reject the null hypothesis. Comparing the calculated p-value to the conventional cut-off for statistical significance of $p < 0.05$, this study suggests there is little evidence to reject H_0, and therefore we accept that there is no statistical difference in WOMAC function scores in males and females with chronic knee pain. However, the observed mean difference between the genders was 5.1 and may be a clinically relevant figure and the lack of evidence to reject H_0 may be as a consequence of the relatively small samples being compared.

Non-parametric methods

If the two sets of data are not normally distributed then the appropriate test is the Mann-Whitney two sample U test. Whereas the t-test was concerned with the mean value of the difference between the two groups, this test is concerned with the ranks of the differences and assesses the hypothesis that the two sets of data are from populations with the same distribution.

The data for both groups are combined with the values ranked in order of size, irrespective of which group they are from. The ranks for each group are then summed. If the mean rank for the two groups is similar, then the groups will have a similar distribution. This relationship can be formally tested by calculating a test statistic, U, which is simply the number of times observations in one sample precede observations in the other sample when the observations are ranked together. The minimum value for U = 0, when all observations in group 2 are larger than group 1, and the maximum value for U = (size of sample 1 x size of sample 2) and is achieved when all observations in group 1 are larger than group 2. So the calculated value of U for the specific set of observations can be compared to these two extremes in standard tables to derive an associated p-value.

Two groups of paired observations

Paired data arise when the same individuals are studied at more than one time point, or when participants are individually matched to another group of individuals. With paired data the interest lies in the average difference between the observations for each individual and the variability in these differences, i.e. the within-subject variability. Hence, the objective is to assess whether the difference in the measure between the groups is zero.

In the follow-up of the chronic knee pain study, participants completed the WOMAC index at two time-points and a question of interest is whether the level of knee function changes in chronic knee pain sufferers over a three-year period.

Parametric methods

If the individual differences between the two sets of data can be assumed to be normally distributed then the appropriate test is the paired t-test. This test examines whether the mean difference between the two sets of data is different from zero, i.e. no difference. Again, the magnitude of the test statistic, *t*, is dependent on three factors: (1) the difference within each participant, (2) the variability within each participant, and (3) the sample size. As for the independent t-test, the calculated value of *t* is then compared to standard tables for the relevant sample size to determine the associated p-value.

H_0: The level of knee function at recruitment is the same, within an individual, as the level of function at three-year follow-up.

H_A: The level of knee function at recruitment is not the same, within an individual, as the level of function at three-year follow-up.

As it was previously shown that function scores were not different between males and females, the data for the two genders had been combined and WOMAC data were available for 105 participants at both time points. The mean WOMAC function score at recruitment was 36.8 and this rose to 42.7 at three-year follow-up. Hence the best estimate of the difference in WOMAC function scores over a three-year period is –5.9, i.e. an average worsening of 5.9 points. The test statistic for this comparison is calculated as $t = -7.4$. Comparing this to tabled values indicates $p < 0.0001$, i.e. there is substantial evidence to reject H_0 and hence to accept that there is a statistically significant deterioration in WOMAC function in older adults with chronic knee pain over a three-year period.

Non-parametric methods

If the within-subject differences are not normally distributed then the appropriate test is the Wilcoxon matched pairs signed rank sum test. As for the case of independent data, the interest lies in ranks and not the values themselves.

For paired data, the differences (sample 1 – sample 2) for each individual is taken and their absolute values, ignoring the sign, are ranked. Then the sum of the ranks is calculated for those individuals where the difference is positive, i.e. sample 1 > sample 2, and for those individuals where the difference is negative, i.e. sample 1 < sample 2. If the two sets of observations are from the same population, the sum of the ranks for positive and negative differences would be about the same. The test statistic, T, is taken to be the smallest of these two sums of ranks and is compared to standard tables to give an associated p-value.

Three or more groups

The examples above can be extended to a situation where there are three or more sets of observations which can be derived either from a single sample or from independent samples. Although the specifics of the analysis methods needed to examine three or more groups of data are beyond the scope of this article, it is important to understand that different analysis methods are required for these two different situations. An example of data from a single sample may be multiple records of WOMAC function in a single cohort of knee pain patients measured at several time-points during the disease course, this is also known as repeated measures data. Alternatively, data on three or more independent groups of subjects can be compared such as to address the question whether WOMAC function scores vary across ethnic groups. A potential analysis method for this form of data would be analysis of variance or ANOVA.

Relationships between two numerical variables

The previous examples in this section have been concerned with comparing a single numerical measure across groups of patients. A natural extension to this is to examine the relationship between two numerical variables within a single sample (Table 3). There are three main reasons why comparisons such as this are carried out:

1. To determine whether the two measures are associated – correlation.

TABLE 3: Analysis methods for examining relationship between two variables

		What is the purpose of the analysis?	
Data type	Data distribution	Determine magnitude of association	Predict value of one variable based on other
Numerical	Parametric	Pearson's correlation	
	Non-parametric	Spearman's rank correlation	Simple linear regression
Categorical		Chi-square test Fisher's exact test	

2. To determine whether the value of one of the measures can be predicted for a known value of the other measure -- linear regression.
3. To compare how well two different approaches are at measuring the same factor – agreement.

Each of these three scenarios addresses a different research question, and hence, each requires distinct statistical methods that serve the relevant purpose. When embarking on this form of data comparison it is vitally important that the objective of the research is clear. The first two scenarios will be covered in this article with readers directed towards specific texts to cover the area of agreement (Bland and Altman, 1986; Dunn, 1989).

Correlation

One simple research question may be, 'Is variable A related to variable B in a set of patients?' The simplest method of analysis to determine the association is correlation. The magnitude of this association can be calculated and this leads to a quantity know as the correlation coefficient which takes values from –1 to +1. This value quantifies the degree of linear association between the two variables, i.e. if we plot the points on a graph, how close to a straight line do they fall? Again, there are different forms of correlation analysis depending on the distributions of the measures being compared. It is also important to remember that correlation assesses the linear relationship between the two variables and hence if a non-linear relationship is present this will not be detected by this specific analysis method. Therefore it is vitally important that the data are assessed in a scatterplot first to determine the overall nature of any association. The calculation of correlation coefficients is not straightforward by hand so the equation will not be presented here as most statistical packages and spreadsheets, including Excel, will perform the calculation.

Two extreme examples are given in Figure 1. In the graph on the left we see an example of perfect positive correlation which is defined as correlation = +1 in which

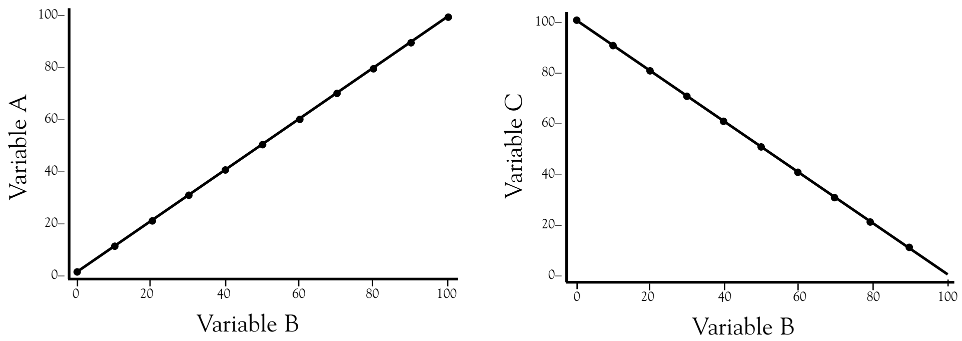

FIGURE 1: Scatterplots representing two extremes of correlation. Left: Perfect positive correlation, r = +1. Right: Perfect negative correlation, r = –1.

the value of the two variables increase together. Conversely, the graph on the right shows variables that are perfectly negatively correlated, correlation = –1 in which as the value of one of the variables increases, the values of the other variable decrease at the same rate. In addition to quantifying the magnitude of the association by a correlation coefficient it is also possible to test the null hypothesis of no association between the variables, and hence derive a p-value. An important issue to remember is that for a particular correlation value, the statistical significance of that correlation is related to the size of the sample it is calculated from, i.e. for the same correlation value, the associated p-value will be smaller for a larger sample size. As an example, a correlation coefficient of 0.3 will indicate a significant linear association for a sample size of 50 ($p < 0.05$) but not for a sample size of 40.

Data from the cross-sectional part of the study were used to examine the research question, 'Does an association exist between the degree of knee function and body mass index (BMI) in older females with chronic knee pain in the general population?'

Parametric methods. If both variables are normally distributed the statistic of interest is the Pearson's correlation coefficient, r. Looking at the data in Figure 2 we can assess by eye that there appears to be some linear association between knee function and BMI, although there appear to be some participants with low function scores but high BMI and vice-versa. The resulting calculation for this dataset gives r = 0.13, i.e. there is some association between knee functioning and BMI. Comparing the calculated value for r together with the sample size against standard values results in a p-value of p = 0.013, i.e. there is evidence to accept that there is a linear relationship between degree of knee functioning and BMI in this population.

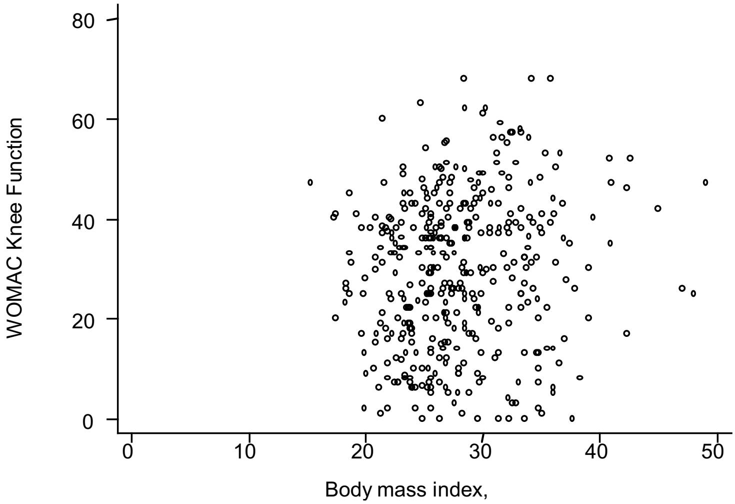

FIGURE 2: Scatterplot of body mass index (kg/m^2) and knee function score (WOMAC) in older females with chronic knee pain.

Non-parametric methods. When the assumption of normality does not hold, the most commonly used non-parametric alternative is the Spearman's rank correlation coefficient, r_S. In common with most other non-parametric methods of analysis, r_S is based on the rankings of the observations rather than the values of the observations. As with the parametric method, r_S takes values from –1 to +1 and a hypothesis test, with the null hypothesis of no association, can be carried out.

Simple linear regression

With a particular set of data it may be important not only to assess the relationship between two variables but also to be able to predict the value of one variable from knowing the other. It is easy to see that the correlation coefficient cannot address this question as it only determines the magnitude of the linear association between the two variables. The appropriate form of analysis to answer this more complex question is regression. There are many forms of regression analysis, the choice of which is dependent on the type of data collected and the exact research question being posed. This article will concentrate on the situation where the interest lies in a linear relationship between two variables, i.e. simple linear regression.

Generally there are two variables, x and y, which can be plotted together as in Figure 2. The x-variable is termed the independent variable and the y-variable the dependent variable. Linear regression is then used to determine the line of best fit through the points on the scatterplot. Looking at the scatterplot of any two variables, there may be several possible lines that can be drawn through the points. Figure 3 shows an example of a scatterplot and the line of best fit, i.e. the regression

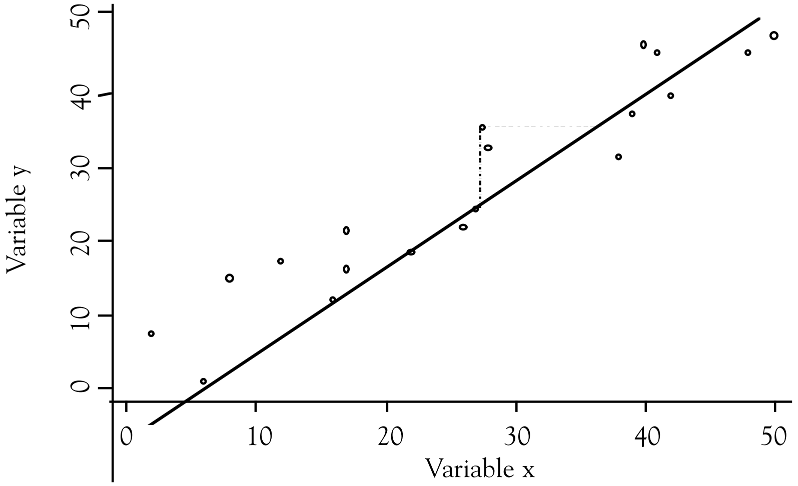

FIGURE 3: Scatterplot including the regression line.

line, through the points. The objective of the regression is to determine the single line that minimizes the distances of all data points from that regression line; the dotted lines on Figure 3 show these distances for one specific observed data point.

All regression lines can be characterized by two measures:

1. The point at which the line would cross the *y*-axis if the line was extended – termed [α].
2. The slope of the line – termed [β].

In general, all regression lines can be illustrated by a simple equation as follows:

$y = [\alpha] + [\beta]\, x$

Returning to the data from the knee pain study, the relationship between knee function severity and body mass can be further investigated using linear regression. In this example, the objective of the analysis is to predict the magnitude of knee function problems for a given body mass value, i.e. x = body mass index and y = WOMAC function score. From the correlation coefficient calculated above, it is already known that a strong linear relationship exists between the two variables.

The regression analysis determines the values of [α] and [β]:

[α] = 11.8 and [β] = 0.73, i.e.

knee function score = 11.8 + 0.73 x body mass index.

From this equation, it is possible to determine the predicted values for knee pain severity for any body mass index value. For example, for a BMI of 25 kgs/m^2 the predicted value of the WOMAC function score would be 30.1.

This is the simplest form of regression with only one independent variable, i.e.

x. The technique can be extended to the case where more than one independent variable is used to estimate or predict the value of the dependent variable, i.e. multiple linear regression, but this is beyond the scope of this introductory series.

Categorical data

This section will concentrate on the tests available for comparing categorical variables. The simplest case is for two by two (2 x 2) tables where both factors of interest are on two levels, e.g. gender (male/female) and disease status (present/absent) and this will be the focus of this section. This can be extended to larger tables where one or both of the variables are measured on more than two levels.

Two by two tables

Participants were recruited to a clinical trial that compared the effectiveness of physiotherapy and steroid injection for treating patients with shoulder pain who presented to primary care (Hay et al., 2003). Participants underwent a clinical examination, in which restriction in active external rotation of the involved shoulder was determined as either present or absent, and completed a self-report question regarding the presence of concomitant neck pain. A research question of interest within this study was, 'Is there an association between concomitant neck pain and restriction in active external rotation in patients recruited to a primary care trial of treatments for shoulder pain?' See Table 4.

To determine the evidence to reject the null hypothesis that there is no association between neck pain and restriction, a chi-square test can be performed which compares the observed data with data that would be expected if H_0 were true, i.e. no association between the two variables. A measure of the differences between the observed and expected data is then calculated, the chi-squared statistic (χ^2), which represents the magnitude of the association between the variables. This test statistic can then be compared to tabulated values to determine the statistic significance of the association.

For the above data $\chi^2 = 2.16$ and when compared to standard tables this is

TABLE 4: Example of a two by two table

		Neck pain		
		Yes	No	Total
Restriction in active external rotation	Yes	21	9	30
	No	94	75	169
	Total	115	84	199

associated with a p-value = 0.142. As this is less than the conventional level for statistical significance, $p < 0.05$, there is no evidence to confirm an association between concomitant neck pain and restriction in active external rotation in participants recruited to a trial with shoulder pain.

Large tables

The methodology presented above for the simple 2 x 2 table can then be extended to larger tables in which either one or both of the variables are measured on more than two levels. Firstly, in the case where one variable is measured on more than two levels the table format would be generally called 2 x k, where k is the number of categories for the second variable. As an example, gender (male/female) and disease severity (mild/moderate/severe) would be specifically called a 2 x 3 table. A second case would be where both variables are measured on more than two levels, generally termed p x k, where p is the number of levels in the first variable. An example here would be smoking status (never/previous/current) and disease severity (mild/moderate/severe). Moreover, as a special example of the 2 x k table, a chi-squared test for linear trend can be carried out where the 'k' variable is ordinal, i.e. has a natural order to it, such as disease severity.

Small samples

There are certain requirements for the chi-square test to be appropriate. The standard criteria are:

1. All cells should have expected values of greater than 1.
2. No more than 20% of the cells can have an expected value of less than 5.

Statistical packages, such as SPSS, inform you if either of the criteria is breached for any particular calculation. In the case were the criteria do not hold, it may be necessary to merge categories, although this will lead to a loss of the detail from the original data. In the case when these criteria cannot be met, for the simple 2 x 2 table an alternative method of analysis, the Fisher's exact test, is available.

Matched data

All the above sections examining the relationship between categorical data have been based on independent groups. As was seen for the numerical data methodology, it is possible to have matched data, such as variables recorded on the same group of patients before and after an intervention. In this situation there is a specific analysis technique, the McNemar's test.

Conclusion

The aim of this third article was to give the reader an introduction to the methods available to assess associations between groups of participants and between variables within groups of participants. The analytical methods covered here were chosen as those most commonly used in the preliminary analysis of data and they also form the basis for more advanced forms of statistical analysis. The aim of the series is to give health care professionals a greater understanding of the use of statistics in health research, so that they may feel confident when reading the literature in their area of interest. It is hoped that this introduction to statistics will encourage health care professionals to be more aware of statistics and less fearful.

Acknowledgements

I would like to thank Sarah Ryan for asking me to write this series, the two anonymous reviewers for their constructive comments on an earlier draft, and Professor Peter Croft for allowing me the time to complete this series.

References

Altman DG (1994). Practical Statistics for Medical Research. London: Chapman and Hall.

Bellamy N (1996). WOMAC Osteoarthritis Index. A user's guide. Ontario: London Health Services Centre, McMasters University.

Bland JM (2000). An Introduction to Medical Statistics (3rd Edn.) Oxford: Oxford University Press.

Bland JM, Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1: 307–310.

Dunn G (1989). Design and Analysis of Reliability Studies: The statistical evaluation of measurement errors. Newcastle upon Tyne: Edward Arnold.

Hay EM, Thomas E, Paterson SM, Dziedzic K, Croft PR (2003). A pragmatic randomised controlled trial of local corticosteroid injection and physiotherapy for the treatment of new episodes of unilateral shoulder pain in primary care. Annals of the Rheumatic Diseases 62: 394–399.

Jordan K, Ong BN, Croft P (1998). Mastering Statistics: A guide for health service professionals and researchers. Cheltenham: Stanley Thornes (Publishers).

Swinscow TDV (1998). Statistics at Square One (9th Edn.) London: BMJ Publishing Group.

Thomas E (2004). An introduction to medical statistics for health care professionals: Describing and presenting data. Musculoskeletal Care 2: 218–228.

Thomas E (2005). An introduction to medical statistics for health care professionals: Hypothesis tests and estimation. Musculoskeletal Care 3: 102–108.

Address correspondence to Elaine Thomas, Primary Care Sciences Research Centre, Keele University, North Staffordshire, ST5 5BG. Tel: +44 1782 583924, Fax: +44 1782 583911. E-mail e.thomas@keele.ac.uk

Received January 2005
Accepted March 2005