Basic statistical methods in research and their interpretation

Stephen Hyer and Jyoti Balani

Epsom & St Helier University Hospitals NHS Trust

Correspondence:
Stephen Hyer
Steve.hyer@nhs.net

Submitted: May 2024
Accepted: July 2024
Published: November 2024

Citation: Hyer and Balani, Basic statistical methods in research and their interpretations, South Sudan Medical Journal, 2024;17(4):207-212 © 2024 The Author(s) **License:** This is an open access article under CC BY-NC DOI: https://dx.doi.org/10.4314/ssmj.v17i4.11

ABSTRACT

Whether quantitative or qualitative, research generates data that requires analysis and interpretation to derive insights. Statistical tests allow researchers to calculate how much the relationship between the variables they have investigated differs from that which might be expected by chance alone. In statistical terms, whether the null hypothesis of no significant relationship is accepted or rejected. This article will consider the common types of statistical tests applied to quantitative research data and their interpretation. By the end of this paper, readers should be better informed about the choice of statistical test for their research study and how to interpret the results.

Introduction

Statistical significance

The null hypothesis for statistical tests simply states that no significant association exists between the variables under consideration. We then employ statistical methods to *support* the null hypothesis, i.e., the findings are no more likely to occur than pure chance, or to *reject* it, i.e., the findings are unlikely to be due to chance.

A *p-value* measures the probability of obtaining the observed results, assuming that the collected data meet the null hypothesis expectation, i.e., there is no effect or relation between the variables. The level of statistical significance is expressed as a *p-value* between 0 and 1. The significance level is conventionally set at 0.05, meaning there is a 5% chance of the result occurring if the variables are not associated. Thus, a *p-value* <0.05 obtained after analysing the research data is statistically significant and unlikely to be a purely random (chance) occurrence. The closer the *p-value* to zero, the less likely it is to have occurred by chance. For a *p* of 0.001, there is a one in one thousand chance; for a *p* of 0.045, the chance is one in twenty-two. Values that round to 0.000 should be reported as <0.001, as they can never be zero.

A statistically significant result says nothing more than that there is an association between the variables. It does not imply a causal relationship.

It is important to recognise that a statistically significant result may, nevertheless, not be clinically significant. A large study can detect small, clinically unimportant findings that are statistically significant. P values are subject to several influences. Lower p-values are found in larger sample sizes when there is a greater spread of observations with large standard deviations and when the measured effect observed in interventions is very significant.

Choosing the correct statistical test (numerical data)

Selecting the right statistical test is often left until after the data has been collected and the study is completed. In fact, the appropriate statistical test should be considered when planning a research study so that the study is adequately powered to accept or reject a hypothesis. This has important implications, for example, for the study sample size. In more complex studies, it's best to get advice from a statistician *before* starting the study. Indeed, to register a clinical trial, a statistical analysis plan is required.

A few basic considerations will help in selecting the correct statistical test, and these will be outlined below.

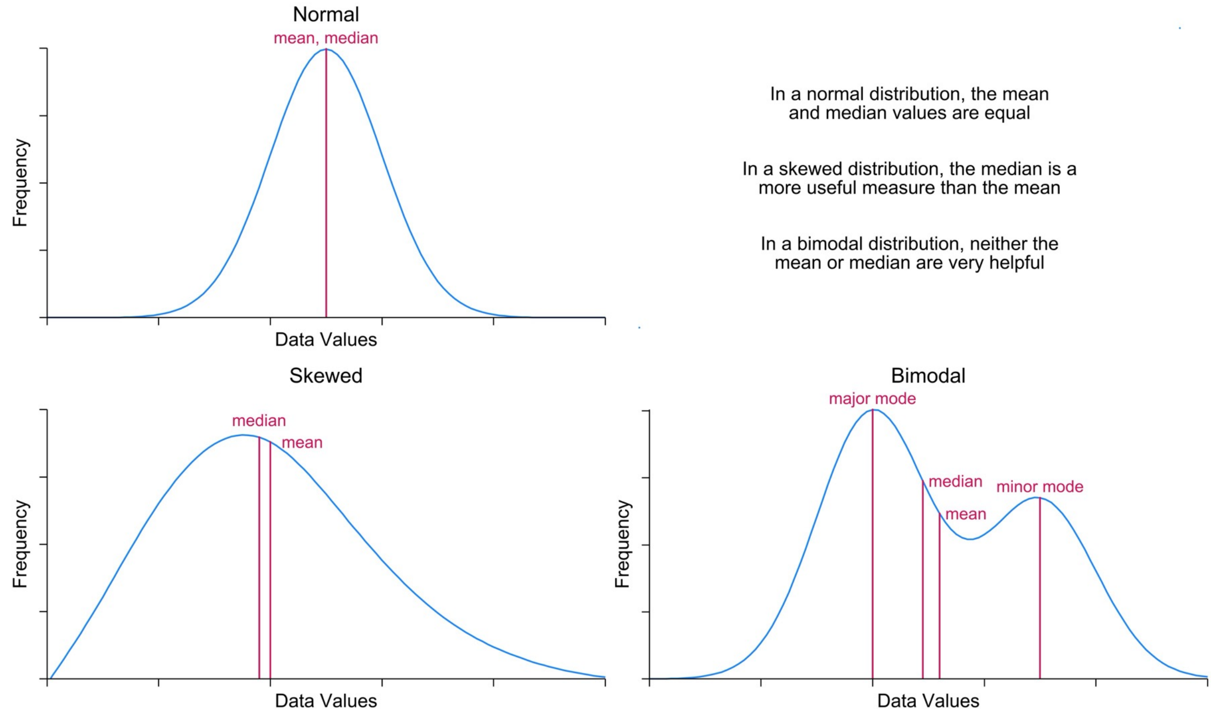
Q1. What types of data are being measured?

Raw data consists of variables or data items; the variable type is important when selecting the appropriate statistical test. *Numerical quantitative* variables (quantities) may be *continuous,* e.g., weight, height, or *discrete,* i.e., limited numbers in a defined collection, e.g., number of siblings. *Categorical* variables are values that are grouped together based on a particular characteristic or attribute, e.g., age group, sex, or educational level. Categorical variables that can be ordered or ranked are referred to as *ordinal* variables, such as the Likert scale of satisfaction rating (extreme dislike, dislike, neutral, like, extreme like). Categorical variables such as region, the categories of which have no obvious order or rank, are called nominal variables. Binary variables are categorical variables with exactly two categories, often yes and no, usually represented by 1 and 0.

Q2. Are the data paired or unpaired?

Consider a researcher undertaking a prospective study of a cohort of patients, making observations on them at two-time points (at the beginning and end of the study). For each individual, there will be two observations (*paired* data). Another study surveys a group with pre- and post-treatment samples, again producing paired data. In a further study, a researcher may compare observations at one point in time in one group of patients with a matched control group. Here, there will be *matched* data, which can also be considered paired.

It is important to determine if the data are paired or independent, as applying the wrong statistical test will give very different results. Independent, unpaired data collected from different populations can give valuable insights into baseline differences between them, which



In a normal distribution, the mean and median values are equal

In a skewed distribution, the median is a more useful measure than the mean

In a bimodal distribution, neither the mean or median are very helpful

Figure 1. Normal and non-normal distributions

can then be generalised. However, paired data in the same population are much more likely to give insight into the effects of a specific treatment or intervention. There are likely to be many other (unmeasured) factors when comparing two unrelated and unmatched populations.

Consider a study that compares results from two different populations that are not related in any significant way. The researcher wants to compare the differences between the two groups. The data will be independent (unpaired). Alternatively, another study comparing results from men versus women (unpaired). Unpaired samples also include a study in the same population when comparing results taken at different times.

Q3. Are the values of the outcome measure of the study in a normal (parametric) distribution or non-parametric?

The term *normal* distribution was introduced in the 19th century when it was believed that many natural phenomena, such as height, were distributed in a symmetrical 'bell-shaped' curve around the mean value. It can also be termed a *Gaussian* distribution (Figure 1). Statistical tests that assume the data are normally distributed are termed *parametric* tests, in contrast to *non-parametric* tests, where this is not assumed.

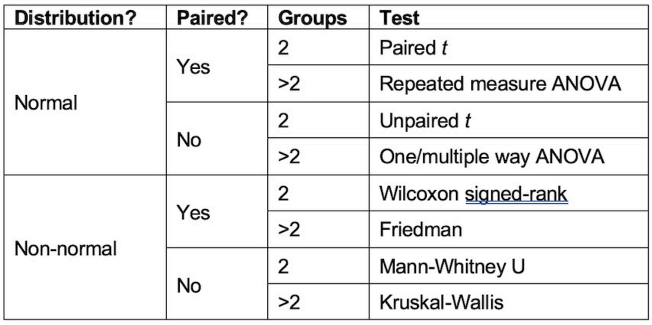
Parametric tests are applied when it can be assumed that the data of interest are at least approximately normally distributed. Depending on whether the data are paired or independent, the means of two groups can be compared using paired t-test or unpaired t-test. If there are more than two groups, the means can be compared by Analysis of Variance (ANOVA). A more sophisticated test, MANOVA (Multivariate Analysis of Variance), analyses multiple dependent variables.

Non-parametric statistical tests do not make any assumptions about the data distribution and are used, for example, where the data are likely to be skewed. The Mann-Whitney U test (also called the Wilcoxon rank sum test) is suitable for comparing two unpaired datasets and can also be used for paired data. If there are more than two sets, the Kruskal-Wallis test is employed.

A simple decision algorithm numerical outcome measures is shown in Table 1.

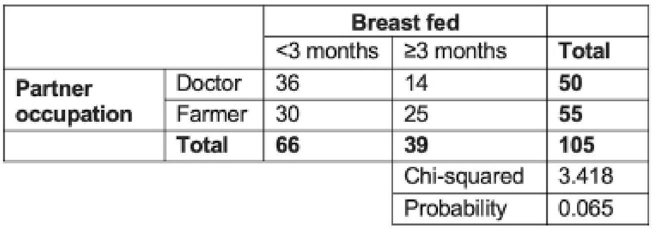
Chi-squared test to compare two categorical variables

Table 1. Decision algorithm for numerical data



Distribution?	Paired?	Groups	Test
Normal	Yes	2	Paired *t*
		>2	Repeated measure ANOVA
	No	2	Unpaired *t*
		>2	One/multiple way ANOVA
Non-normal	Yes	2	Wilcoxon signed-rank
		>2	Friedman
	No	2	Mann-Whitney U
		>2	Kruskal-Wallis

Table 2. Relationship between duration of breastfeeding and partner occupation (invented data)



		Breast fed		
		<3 months	≥3 months	**Total**
Partner occupation	Doctor	36	14	**50**
	Farmer	30	25	**55**
	Total	**66**	**39**	**105**
			Chi-squared	3.418
			Probability	0.065

The Chi-squared test is a commonly used test to determine whether observed data are significantly different from what would be expected if there were no association between the variables. It tests categorical data, which are usually displayed in a frequency distribution table (contingency table; Table 2). Note that the table contains actual numbers of occurrences and not percentages, means, proportions, or other calculated numbers.

Consider researchers interested in the length of breast-feeding (less than 3 months versus 3 months or more) comparing doctors' wives with farmers' wives; the null hypothesis being that there is no difference.

The chi-squared statistic measures the extent to which the observed values in the table differ from the expected values (the values if there were no association between the variables). The chi-squared probability tells us the probability of the observed values occurring under the null hypothesis of no association. Since 0.065 is greater than 0.050, we cannot reject the null hypothesis. The probability depends not only on the value of the chi-squared statistic but on the number of rows and columns in the table. This test should only be used if all of the expected table cell values are greater than one and 80% of the expected values are greater than five. Rows or columns can usually be combined to meet this requirement, or Fisher's exact test can be used instead. For neither test should "no answer" categories be included.

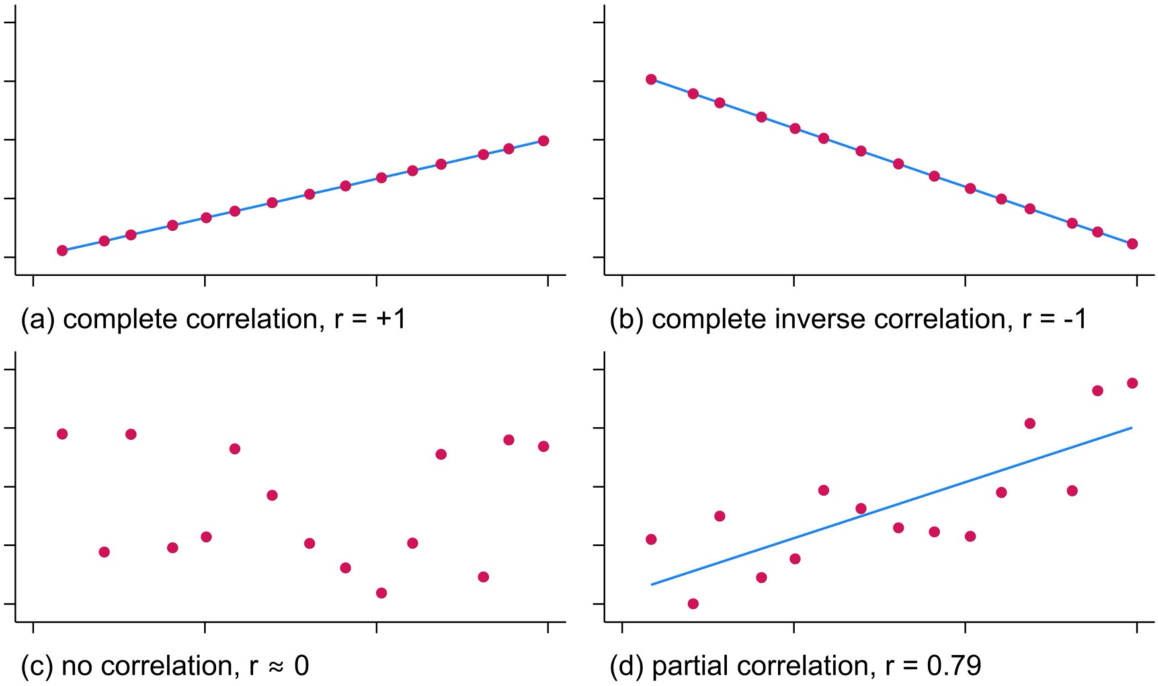


Figure 2. Illustration of different correlation coefficients. X axis independent variable, Y axis dependent variable

Regression analysis

In regression analysis, we are trying to quantify the relationship between a dependent variable (the variable you want to analyse) and at least one independent variable, the explanatory variable. This can be used to predict the extent to which changes in one variable will affect changes in the outcome variable of interest. We will look at an example where a researcher investigates the influence of prednisolone dose (the independent variable) on plasma glucose levels (the dependent variable of interest).

Simple Linear regression analysis

In linear regression analysis, we assume that the relationship between variables can be described by a straight line called the regression line. A simple linear regression analysis only looks at two variables (i.e., one independent and one dependent variable) and is sometimes called bivariate.

Linear regression is typically used with continuous variables, such as height, weight, and blood glucose level, but discrete and even some ordinal variables can be used, and variables can be transformed, for example, by taking logarithms. Traditionally, linear regression was thought of graphically, with the dependent variable plotted on the vertical (y) axis and the independent variable on the horizontal (x) axis. A positive slope shows that as x increases, y increases, whilst in a negative slope, y decreases as x increases (Figure 2). Computer software is used to fit a straight line to the data set.

Regression statistics

The strength and the direction of the relationship between the dependent and independent variables are given by the *correlation coefficient* (r), sometimes referred to as Pearson's correlation coefficient, provided both the dependent and independent variables are normally distributed. If either of the variables is not normally distributed, then Spearman's rho is the non-parametric equivalent of Pearson's r. A perfect direct relationship between the variables is denoted by an r value of +1. An r value of 0 denotes no relationship, while r = -1 indicates a perfect negative relationship (Figure 2). The correlation coefficient values near 1 indicate the strength of the relationship, while the '+' or '–' sign indicates the direction of the relationship. An increase in a dependent variable with an increase in the independent variable indicates a positive correlation, denoted by the '+' sign, while a decrease in a dependent variable with an increase in the independent variable is denoted by a '–' sign. The *p-value* determines the statistical significance of the correlation coefficient. Significant positive or negative correlation further needs to be assessed statistically by linear regression analysis after assuming certain prerequisites.

Another important statistic is the *regression coefficient* (β) which describes the change in the dependent variable (y) for each one-unit change in the independent variable (x). This corresponds to the gradient of the line using the equation for a straight line: $y = \beta x + c$ where c is the value of y when x is 0 i.e. the intercept (sometimes shown as β0).

Let us consider a hypothetical research project investigating the influence of prednisolone dose on plasma glucose (Figure 3). The linear regression analysis has shown a best-fit line with the equation $y = 1.01 + 0.60x$. The β value (slope, regression coefficient) of 0.60 indicates that for every one-unit increase in the independent variable (dose of prednisolone), the plasma glucose will rise by 0.60 units.

The *R2 value* (sometimes called the coefficient of determination) assesses the strength of the model. In the example, an R2 of 0.82 or 82% indicates that 82% of the variability observed in the dependent variable (y) (plasma glucose) is explained by the regression model, i.e., changes in prednisolone dose (x).

In Figure 3, the *Confidence Interval* (CI), which is typically set at 95%, means that we can be 95% confident that the regression line lies within this range (grey area).

It is important to note that a strong correlation (high r and R2 values that are statistically significant) does not prove cause and effect. For example, another variable that has not been measured (the hidden variable) may be the cause. Linear regression assumes a linear relationship when perhaps the data points would be best fitted on a curved line.

Multiple linear regression analysis is commonly used to examine multiple variables in relation to a single dependent variable. More complicated models use multiple dependent and independent variables (multivariate linear regression). These models provide a more realistic picture than simple linear regressions but still assume a linear relationship.

Logistic regression analysis

Clinical studies that evaluate the association between one or more factors and a single binary outcome, such as the presence or absence of death or disease, most often employ the method of logistic regression. Unlike linear regression, the relationship between the dependent and independent variables does not need to be linear. Whereas linear regression uses the best-fit straight line, logistic regression uses the S-shaped sigmoid curve, known as the logistic function (Figure 4).

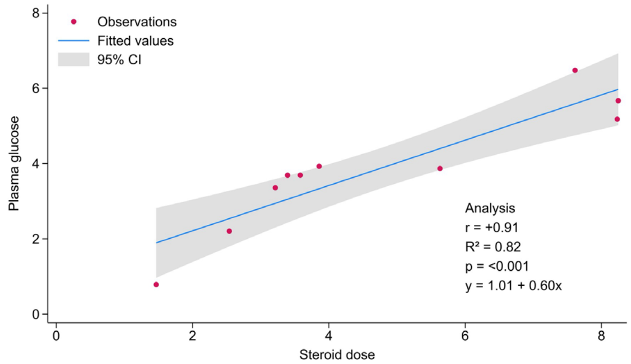


Figure 3: Hypothetical example of linear regression with statistics (explained in the text).

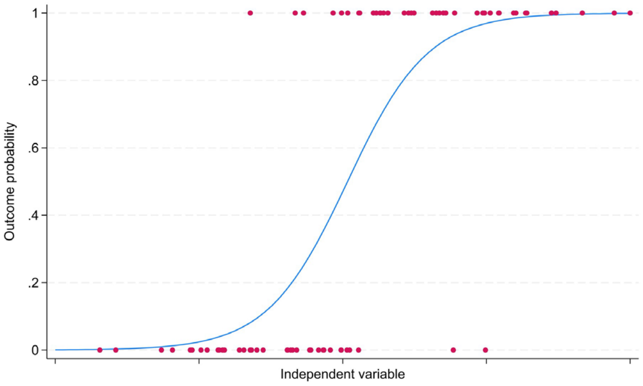


Figure 4: Sigmoid probability curve and example data points

Logistic regression calculates the probability of a binary (yes/no) event (the dependent variable) occurring based on one or more independent variables. For example, a researcher wants to know the likelihood of developing diabetes amongst South Sudanese children of different ages, ethnicities, weights, heights, social backgrounds, etc. These independent variables can also be termed risk factors.

The calculations used in logistic regression are complex and are nowadays performed by statistical software. The statistical outcome of logistic regression is usually expressed as the odds ratio (OR) for a unit increase in an independent variable and the 95% confidence intervals for the OR. We discuss this next.

Odds Ratio

$$\text{The odds ratio (OR)}=\frac{\text{odds of the outcome or event occurring}}{\text{odds of the outcome or event not occurring}}$$

The *odds* are the ratio of two probabilities: the probability that an event will occur divided by the probability that it will not occur.

For exposure to a risk factor, OR can be easily understood as the odds of an event after exposure divided by the odds of the event in the reference group who have not been exposed to the risk factor.

It is easier to think about OR in a contingency table. Consider a hypothetical research project investigating the risk of developing diabetes and exposure to cassava in diet (Table 3).

So the odds of developing diabetes after cassava consumption are (35÷45)÷(10÷45), that is 3.5, and the odds of developing diabetes with no exposure to cassava are (15÷55)÷(40÷55), that is 0.375. An OR of 1.0 suggests that exposure to the independent factor does not affect the probability of disease. OR<1 suggests that the independent variable is a protective factor, making the probability of developing the disease less likely. OR>1 suggests that the variable is a risk factor. In this example, the odds of developing diabetes after exposure to cassava in the diet is 9.33 times greater than the odds for those not consuming cassava.

Note that the OR is quite different from the relative risk or risk ratio (RR), except when the outcome or event is extremely rare. Using the same example, the risk after cassava consumption is 35÷45, and the risk after no consumption is 15÷55, so the relative risk or risk ratio is (35÷45)÷(15÷55), i.e., ≈2.85. Like ORs, RRs should be presented with confidence limits.

Summary

Statistical tests are powerful tools used for all types of clinical research. Selecting the appropriate test is crucial and depends on the study design. Computer programmes are widely available to do the calculations. Of the commonly used tests, the unpaired (standard) t-test is appropriate for comparing means from exactly two groups, such as controls versus experimental group, while the paired t-test is chosen for detecting differences in *before* and *after* type of studies in the same individuals/groups. T-tests should not be used repeatedly in the same study to compare different groups. Where there are more than two groups, the appropriate test is ANOVA: one-way ANOVA if one independent variable, and two-way if two different independent variables, e.g., two different treatments in the same study. Regression analysis allows the researcher to estimate the relationship between dependent variables and one or more explanatory variables. Correctly interpreting observed data provides useful insights for better clinical practice.

Table 3. Relationship between cassava consumption and diabetes (invented data)

		Diabetes		
		Yes	No	**Total**
Cassava in diet	Yes	35	10	**45**
	No	15	40	**55**
	Total	**50**	**50**	**100**

Sources of funding: none

Conflict of interest: none

Further reading

1. Heumann C, Schomaker M, Shalabh S. Introduction to statistics and data analysis: With exercises, solutions and applications in r. Springer International Publishing; 2017. https://doi.org/10.1007/978-3-319-46162-5.
2. Verma PJ, Abdel-Salam A-S. Testing statistical assumptions in research. Wiley; 2019. https://doi.org/10.1002/9781119528388.
3. Franke TM, Ho T, Christie CA. The chi-square test: often used and more often misinterpreted. Am J Evaluation. 2011;33(3):448-58. https://doi.org/10.1177/1098214011426594
4. The British Medical Journal (15. Study design and choosing a statistical test) https://thebmj-frontend.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/13-study-design-and-choosing-statisti
5. Pocock S. The simplest statistical test: how to check for a difference between treatments BMJ 2006; 332:1256 https://doi.org/10.1136/bmj.332.7552.1256
6. Shreffler J; Huecker MR. Types of Variables and Commonly Used Statistical Designs https://www.ncbi.nlm.nih.gov/books/NBK557882/