Update in Anaesthesia

Statistics for anaesthetists

Andrew Woodgate
Correspondence email: andrew.woodgate@nhs.net

Summary

Statistics are used to package raw data (often a large amount) into a form that is concise and has meaning to the reader. An understanding of statistics is important to medical practitioners to update clinical practice by analysis of experimental data and conclusions presented to us in medical journals or by company representatives. Summary conclusions should not be accepted at face value and, unless a study has been conducted rigorously, the results may be inaccurate and open to misinterpretation.

Andrew Woodgate
Core Trainee in Anaesthesia
Royal Devon and Exeter NHS Foundation Trust
Barrack Road
Exeter EX5 1JT
UK

INTRODUCTION

"Facts are stubborn things, statistics are pliable" (Mark Twain)

"There are three types of lies... Lies, damn lies and statistics" (often attributed to Benjamin Disraeli).

DATA COLLECTION

The basic premise of clinical research is to answer a question, for example, "Is treatment A more effective than treatment B?" In order to answer this question a researcher must identify a sample of individuals that is representative of the target population and use an appropriate study design to collect accurate and reliable data. If this is not done correctly, then the quality of statistical analysis is irrelevant - the findings of the study will be unreliable and potentially false, and so misleading and clinically harmful conclusions will be drawn.

SAMPLING

A target population is an entire group that a researcher is interested in, but the group is often too large to be accessed by the researcher (e.g. all patients undergoing elective total hip replacements in a country). To overcome the problem of accessibility, a sample is selected whose individual's characteristics are representative of the characteristics of the target population. The most representative sample is a simple random sample, in which every member of the target population has an equal chance of being selected. Statistical assessment of how well a sample's results match the entire target population's true value (sample error) is discussed later.

STUDY DESIGN

Study design falls into two broad categories:

1. Observational
2. Experimental (also called intervention or treatment).

Observational studies

Observational studies collect data from a sample group, but they do not introduce any intervention. Examples of observational studies include cross-section studies, cohort studies and case control studies.

Cross-sectional study

A cross-sectional study is a snap-shot of a situation at a certain point in time, where measured variables are compared and conclusions made (see Study 1, below).

Cohort study

Cohort studies (also called prospective or longitudinal studies) involve following a sample group over a period of time, with measurements taken at intervals. An example is follow up of patients after colonic resection using a novel technique, to define the incidence of cancer recurrence.

Case control study

Case control studies (also referred to as longitudinal or retrospective studies) involve selecting a sample group with a known outcome and comparing the subjects to a sample group who do not have the same outcome, but in other ways are similar. Inferences are then made from differences between the groups.

WORKED EXAMPLES

Study 1: Cross-sectional study into the risk factors for postoperative nausea and vomiting

If we wanted to investigate the causes of postoperative nausea and vomiting (PONV), an observational method is appropriate.

Sampling

- Defining our target population - this could be all patients undergoing general anaesthetic or be more specific to a certain type of surgery or anaesthetic technique.
- Sampling technique - random sample from the population, e.g. alternate patients.

Study design

- Identify an outcome measure - in this case a score of over 5 on the simplified PONV impact scale, would be used to define clinically significant PONV.
- A cross-sectional study could be used appropriately to identify likely risk factors. All possible risk factors should be predefined and that data collected (e.g. age, gender, weight, surgical procedure, past history of PONV).

The major drawback with observational studies is that confounding variables made lead us to draw false conclusion about the relationship between risk factors and outcome (see Box).

Experimental studies

In experimental studies the researcher measures the effect of an intervention they expose the sample to. For comparison, a proportion of the randomly selected sample becomes a control group, who do not receive the intervention. Selected subjects are entered into the treatment or control group in a random fashion; this process is called *randomization*, and reduces selection bias.

Confounding variables

These are unmeasured variables that influence the measured outcome. The confounding variable may be a third variable that correlates with both measured variables and so falsely leads us to conclude that there is a correlation between the two variables we have measured. For example, consider a study designed to identify a link between surgery performed as a neonate, with subsequent developmental delay. A confounding variable would be low birth weight, as this is likely to correlate with both the need for neonatal surgery and developmental delay, without itself being a causative factor.

Bias can also be reduced by blinding. In this process, the patient or the assessor is blinded from knowing which group a subject is/was allocated to. If both patient and assessor are kept unaware of allocation, this is called double blinding. Double blinded randomized controlled trials (RCTs) generally produce the most reliable results and are viewed as the gold standard. A practical issue associated with RCTs occurs when subjects drop out of either the intervention or control group - when this occurs it is viewed as good practice to analyse the data as if the lost subjects were still within the study - this is known as *intention-to-treat analysis.*

Bias is defined as an un-random influence which causes results to deviate from the true value. For example, if a study designed to assess recovery after knee surgery only enrols patients who are willing to come back to the hospital for assessment on day 3 post-operatively, this will introduce a bias towards enrolling fitter, more active patients. **Randomization** and **blinding** are techniques used to reduce both intentional and unintentional bias.

DESCRIPTIVE STATISTICS

Descriptive statistics are used to present large amounts of data in an interpretable form, using numerical or graphical methods.

Data types and their representation

Data can come in a number of forms that fit into different categories (Figure 1). The implication of the different categories is that data requires different methods of analysis and presentation depending on its form.

The basic division of categories is between whether the data is qualitative (also referred to as categorical) or quantitative (also referred to as numeric/metric).

Qualitative data

This is descriptive in nature and can be further sub-grouped into nominal and ordinal categories. Ordinal data is qualitative data that has a sequential order; examples include American Society of Anaethesiology (ASA) grades, Mallampati score and Visual Analogue

WORKED EXAMPLES

Study 2: Difference in speed of onset between two antiemetic drugs

This is a simple example of a study designed to answer the question - Does antiemetic A or antiemetic B have a faster onset?

Sampling

- Target population - all patients receiving a general anaesthetic who suffer PONV in recovery and require administration of an antiemetic.
- Random sample or all patients, depending on numbers.

Study design

- Randomised double blind study comparing the two groups. A control group is inappropriate in this scenario as it is unethical not to treat PONV. Random allocation of patients into group A (receive drug A) or group B (receive drug B).
- Blinding of assessor to patient group allocation. It is also possible to blind the person administering the drug if they are prepared in pharmacy and labelled as drug A and drug B. Complete blinding may not be possible - if drug A is prochlorperazine, it will cause tachycardia and so the administering nurse will be able to guess that drug A is indeed prochlorperazine.

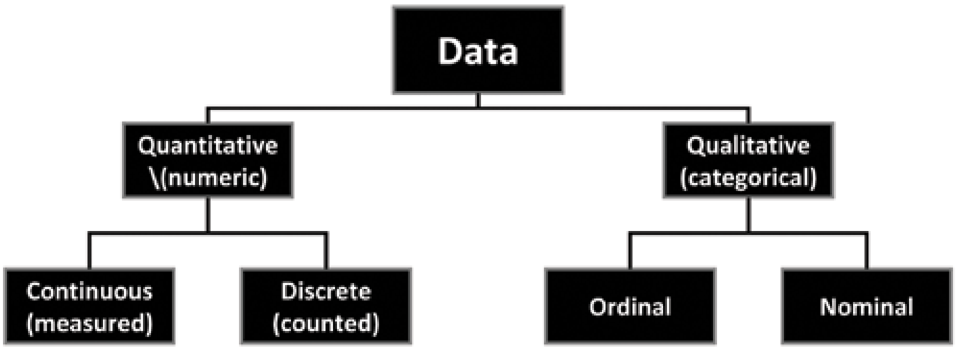


Figure 1. *Types of data*

(VAS) scores (e.g. for pain assessment). Nominal data is qualitative data that has no particular order; examples include type of operation, blood group and gender.

Quantitative data

This is numeric data that may be placed on a scale and that has units of measurement. Quantitative data can be further sub-grouped into continuous and discrete data. Continuous (also called measured) data can be any number including fractions of a whole numbers; for example height, weight, age. Discrete (also called counted) data can only be a whole number; for example heart rate, number of hospital admissions.

Table 1 shows the appropriate graphical forms of data presentation and display according to the data type.

Nominal	Ordinal	Discrete	Continuous
Pie Chart	Bar Chart	Bar Chart	Dot Plot
Bar Chart	Dot Plot	Dot Plot	Histogram
		Line Chart	
		Histogram	

Table 1. *Data presentation according to data type*

All types of data can be summarized in a tabulated format such as a frequency table (Tables 2 and 3 under 'Worked examples'). When tabulating continuous data there are often too many measured variables for practical use, therefore authors will often present the data in a grouped format (Table 3). The risk in this process is that, if there are too few groups, the level of detail in the data is lost.

Measures of central tendency, variability and distribution

As well as display in graphical and tabulated format, data can be summarized numerically. It is necessary to summarise any presented data in a way that enables it to be evaluated or compared to other data. In some instances this can be achieved by using percentages or proportions. Data may also be summarized by providing a measure of central tendency (mean, mode, median) along with a measure of the spread of data around it and the type of distribution.

Central tendency is defined as 'the tendency for the values of a variable to cluster around its mean, mode or median'. The type of data being summarized again dictates whether the mean, mode or median are used.

Mean

The mean is the sum of all values divided by the number of observations and is only used to summarise quantitative data. Qualitative data is not appropriately summarized with a mean as, although the data may be numeric (e.g. VAS pain score), there is a lack of 'objective proportionality' to it (i.e. a score of eight does not mean the pain is twice as bad as a score of four).

Mode

The mode is the value that occurs most frequently and is the only measure of central tendency that can be used with nominal data.

Median

The median is the middle value that separates the lower half and upper half of a data set (it will have a decimal place in an even numbered data set). The median is used to summarise ordinal and quantitative data.

There are advantages and disadvantages to using either the mean or median as a summary of quantitative data. The mean includes every value from the data set, which intuitively suggests it offers a more complete summary, however it will also be influenced by extreme outlier observations which may produce a misleading summary. The median is unaffected by extreme outliers. An example of this is length of stay after surgery, where one patient who has a stay of 4 months, amongst many who stay for 4-7 days, will greatly affect the mean but not the median. By convention it is usual to use the mean to summarise normally distributed data and the median for skewed (non-normally distributed) data (see below).

Spread

Spread refers to how tightly the measured variables of a dataset cluster around the value of central tendency - the tighter the cluster of variables around the central point, the narrower the spread. Spread is used as an indicator of how likely it is that the sample mean represents the true mean of the target population, i.e. a narrower spread corresponds with an increased likelihood of the sample mean representing the target population mean.

The choice of method of measuring spread of data in a sample is influenced by the type of data. As nominal data has no order, spread cannot be used. However spread is used with ordinal and quantitative data, where there is an order to data.

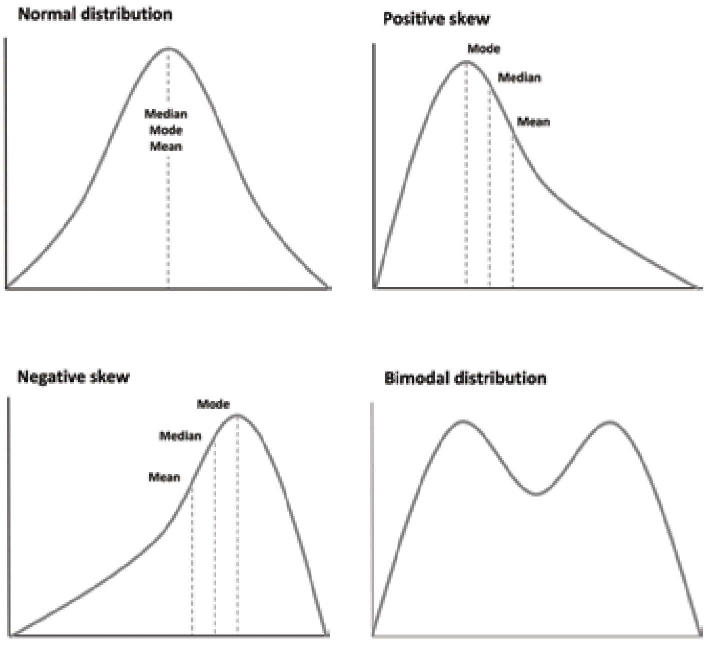


Figure 2. *Graphical representation of normal, positively skewed, negatively skewed and bimodal distribution*

When using the median as the description of central tendency i.e. for ordinal or quantitative data, then the measure of spread is the inter-quartile range. The *inter-quartile range* is defined as the range of values between the 75th centile and the 25th centile of the values measured. In other words it marks out the middle 50% of values within the set of data (with the median being the centre point). The inter-quartile range is used rather than the whole range to eliminate extreme outliers.

When using the mean to indicate central tendency, then the measure of spread is the *standard deviation* (SD). This represents the average distance of all the data values from the mean value, the smaller that value the narrower the spread. Calculating the standard deviation is relatively laborious and involves subtracting the mean from each individual value in the sample and squaring the difference (to eliminate the negative values of those data below the mean), all the squared values are then added with the total being divided by the number of values in the sample minus 1 (n-1) - this figure is called the *variance*. The square route of the variance is the standard deviation.

The type of data distribution influences the choice of the most appropriate method of analysis. Data may be distributed as (see Figure 2):

- normal (parametrically) where the mean, median and mode are all equal,
- positively skewed where the mean is greater than the median,
- negatively skewed where the mean is less than the median,
- bimodal where there are two peaks.

The distributive pattern of data can be assessed visually from graphical data or with statistical tests, such as the Shapiro-Wilkes test.

In normally distributed data, 67% of a sample's values will lie within one standard deviation either side of the mean, 95% lie within two standard deviations either side of the mean and 99% lie within three standard deviations either side of the mean.

DEDUCTION AND INFERENTIAL STATISTICS

Inferential statistics are used to assess the relevance of the findings of a study to a target population. This includes evaluating the element of chance for different findings between groups, evaluating the difference in baseline characteristics between groups and estimating how closely the sample represents the target population.

Standard error of the mean and confidence intervals

It is possible to use descriptive statistics (such as the mean) to make informed estimations of how accurate a representation a sample is of the target population. In other words, we can assess sample error using the mean. The sample error is quantified by calculating the standard error of the mean (SEM). The standard error of the mean is inversely proportional to the number of subjects within a sample and is calculated by dividing the standard deviation by the square route of the number of subjects in the sample.

A confidence interval is another method of estimating how similar the results of a sample are likely to be to the true population. The confidence interval is a range of values with a percentage estimation of how likely it is that values of the true population are to be found within this range. For example, if the average diastolic blood pressure in a sample of healthy pregnant women at term was analysed and was

WORKED EXAMPLES

Study 1: Cross-sectional study into the risk factors for postoperative nausea and vomiting

A sample of 100 subjects was identified as suffering PONV in the cross-sectional study. To present the relationship of PONV with the type of surgery we could include a frequency table (Table 2) and/or a pie chart (Figure 3). There were twenty-five subjects in each of the four groups. Due to the nominal nature of the data, markers of central tendency and spread are not indicated.

Table 2. *Frequency table of PONV according to type of surgery*

PONV	Orthopaedics	ENT	Plastics	General
Yes	2	15	1	5
No	23	10	24	20
Total	**25**	**25**	**25**	**25**

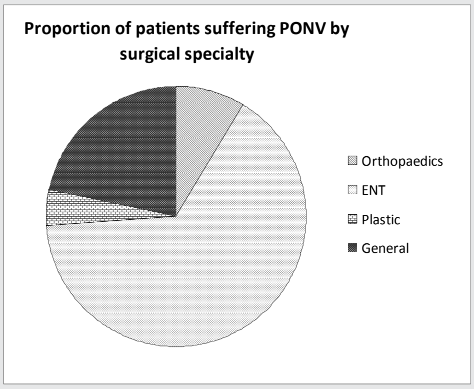


Table 3. *Pie chart of PONV according to type of surgery*

WORKED EXAMPLES
Study 2: Difference in speed of onset between two antiemetic drugs

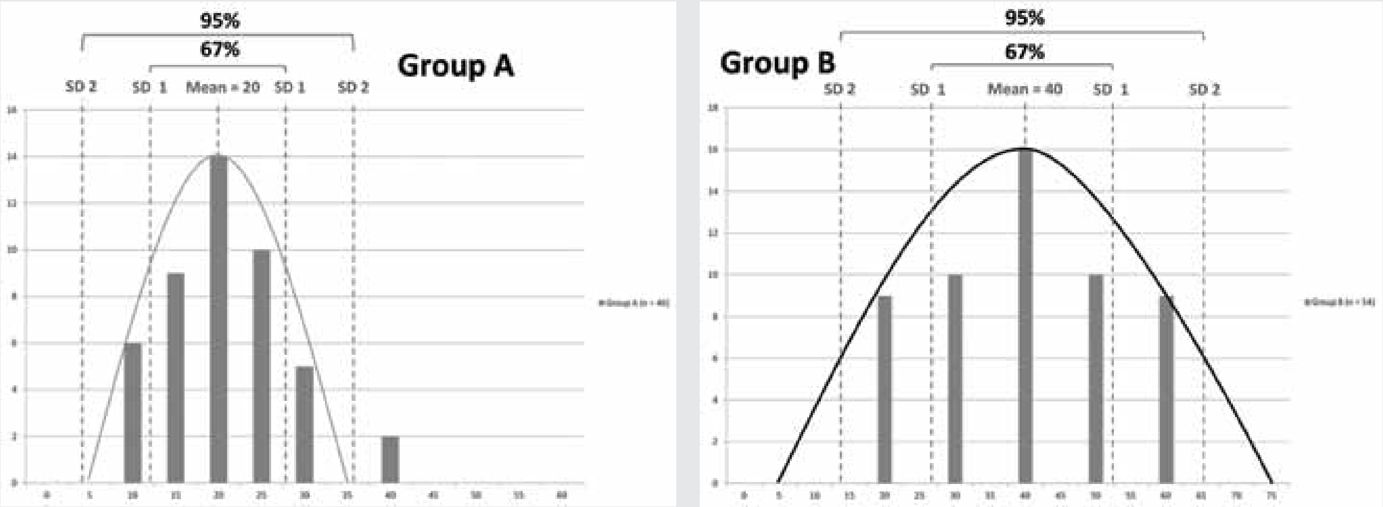
The measured variable is time, which is quantitative continuous data. However the presence of PONV was only measured at five minute intervals, effectively making the data discrete. Visual representation could be using a frequency table (Table 3), a histogram (Figure 4) or frequency curves. Summarising the data with a measure of spread and central tendency is influenced by the distribution of the data. The results (Table 3 and Figure 4) demonstrate a normal distribution, so the mean and standard deviation are used to summarise both groups.

Table 3. *Frequency table of time to resolution of PONV after antiemetic administration*

Minutes	**0**	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**	**70**	**75**
Group A (n = 46)	0	0	6	9	14	10	5	0	2	0	0	0	0	0	0	0
Group B (n = 54)	0	0	0	0	9	0	10	0	16	0	10	0	9	0	0	0

	Median	**Mode**	**Mean**	**Variance**	**SD**	**SEM**
Group A (n = 46)	20	20	20	60.9	7.8	1.1
Group B (n = 54)	40	40	40	170.4	13.1	1.8

Figure 4. *Bar chart (histogram) of time to resolution of PONV after antiemetic administration*



normally distributed we could infer, with 95% confidence, that the mean diastolic blood pressure for all healthy women in the population at term fell within two standard deviations either side of the sample mean (and be 99% confident that it fell within three standard deviations).

Confidence intervals can also be calculated from the median, a process that usually requires a computer program, which bases its calculation on the Wilcoxon signed-rank procedure.

Probability theory and statistical tests

Fundamental to the usefulness of a study is how confident we are that results are valid; in other words, how likely is it that any difference between groups is the result of the intervention or just down to chance? Inferential statistics rely on the use of probability (the p-value) to analyse the relationships between different variables. The p-value is defined as a measure of the chance of having a particular outcome in a given set of circumstances. If the outcome is certain then the probability is 1, if it is impossible then the probability is 0. The convention in clinical research is to accept that any differences with a p-value of less than 0.05 (i.e. 1 in 20) are likely due to the intervention rather than by chance. The smaller the p-value the more confident a researcher will be that any differences are not due to chance. So, if a study suggests that antiemetic A is better than antiemetic B, with a p-value of 0.05, then we can be 95% certain that this study reflects a true finding in the target population. Put another way, if we repeated the study 20 times, on average one of the studies would give us an untrue result (not representative of the true effect of the anti-emetics on the target population). So, if a journal publishes 20 studies, all with p values of 0.05, then one of these studies is likely to be misleading.

The p-value is calculated from data using statistical tests. The appropriate choice of test depends on a number of factors such as;

the type of data, distribution of the data, the number of groups and whether the data is paired or unpaired. Figure 5 identifies which statistical test is appropriate depending on the characteristics of the data.

Two common statistical tests are the chi square test and the student t test (Table 4). The chi square test is used on qualitative (categorical) data and can be used to test the equality of population properties in two or more independent groups. The student t test is used on quantitative data. Both the chi square and student t test use the sample data to calculate a figure which is used in conjunction with the degrees of freedom (Table 4) from the individual study to calculate a p-value from statistical tables (available on line).

The 'null hypothesis'

Traditionally all studies should propose a null hypothesis - an assumption that there is no difference between experimental interventions. Therefore if a significant difference in the measured out-come between groups is found then the null hypothesis is rejected.

Error and power

Error occurs when the wrong conclusions are drawn from a study

Table 4. *Statistical equations*

Term	Definition
Chi square test	$\sum (O-E)^{2} / E$ 0 = observed frequency E = expected frequency ∑ = sum of
Student t test	Difference in sample means / estimated standard error of the difference
Standard error of the difference	(SD of group A^{2} / n group A) + (SD of group B^{2} / n group B)
Degrees of freedom (df)	The number of means which are free to vary. 1. For student t test, df = n-1 2. For chi square, df = (number of possible outcomes -1) + (number of groups -1)

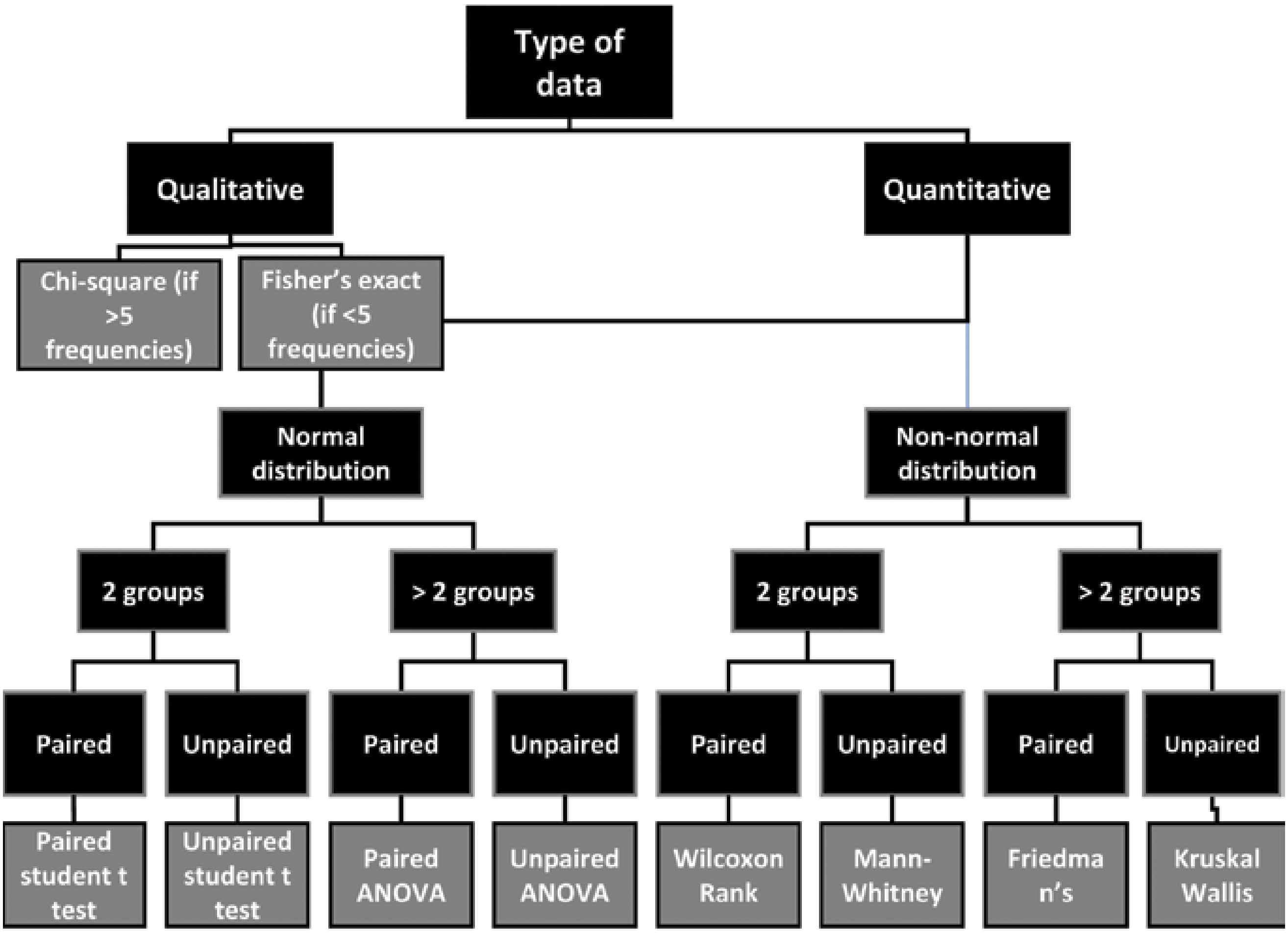


Figure 5. *Factors influencing choice of statistical test*

and are classified as type 1 and type 2 errors.

A type 1 error (alpha error or false positive) is the rejection of the null hypothesis inappropriately - assuming a significant effect from an intervention where one does not truly exist. Type 1 errors are related to the p-value. There is a 5% (1 in 20) chance of a type 1 error with a p-value of 0.05.

A type 2 error (beta error, false negative) is the acceptance of the null hypothesis when it is false - concluding there is no effect when there actually is. The chance of a type 2 error is related to sample size. Convention dictates that it is acceptable to allow a 20% chance of a type 2 error. The power of a study is the chance of avoiding a type 2 error and is defined as 1-beta (ie 0.2). Statisticians can calculate the number of patients required to ensure a study has sufficient power prior to the start of the study by using a combination of equations.

CONCLUSION

Statistical analysis is relevant to all medical practitioners. This article has covered some of the basic principles and enables the reader to have an understanding of the processes data is been put through and which processes are appropriate for different scenarios.

FURTHER READING

1. Medical statistics from scratch. David Bower. 2002. Wiley, UK.
2. Statistics at square 1. 10th ed. Swinscow TD, Campbell MJ. 2002. BMJ books.

WORKED EXAMPLES

Study 1: Cross-sectional study into the risk factors for postoperative nausea and vomiting

We can evaluate the nominal data collected from our observational study with the chi square test. This enables us to infer causation between the type of surgery and the likelihood of PONV by evaluating the element of chance in the results.

To calculate the expected frequency we divide the number of patients defined as suffering PONV by the total number of subjects.

- in this case 23/100 gives us a figure of 0.23;

To find the expected number of patients with PONV according to each surgical specialty we multiply the number of subjects in each group by 0.23;

- 25 x 0.23 = 5.75 as the expected number of patient s to suffer PONV in each group (as there are 25 subjects in each group.

We will also need to calculate the expected frequency of patients to not suffer PONV by following the same process;

- 77/100 = 0.77.
- 25 x 0.77 = 19.5

Table 5. *Chi square analysis of data*

PONV	Orthopaedics	ENT	Plastics	General	Expected Cases
Yes	2	15	1	5	5.75
No	23	10	24	20	19.25
Chi square score	3.2	29.1	5	5.5	-
p-value	>0.1	<0.001	>0.1	>0.1	-

We can now input the figures relevant to each group into the Chi square equation to calculate if the incidence of PONV is statistically significant compared to what would be expected. To calculate the chi square equation we subtract expected frequency from the observed frequency, square this number and divide the result by the expected frequency. The sum of the values gives the chi square score which is used (with the number of degrees of freedom) to read off a p-value from an appropriate statistical table.

- For the orthopaedic surgery group; ((2-5.75)squared/5.75) + ((23-19.5)squared/19.5) = 3.18

The chi square scores for each group are displayed in Table 5 along with the p-value indicating if observed frequency was statistically different to the observed. The results from this (fictitious) study suggest that, with all else being equal, patients undergoing ENT surgery are significantly more likely to suffer PONV compared to those undergoing general, orthopaedics and plastic surgery.

WORKED EXAMPLES

Study 2: Difference in speed of onset between two antiemetic drugs

- Null hypothesis: That there is no difference in the speed of onset of antiemetic A when compared to antiemetic B.
- Assess baseline variability between groups. Chi square can be used on multiple variables (age categories, ASA grade, type of surgery, group numbers etc) to assess if there is significant baseline variability between the two sample groups.
- Evaluate the difference in outcome measure between groups using the student t test to calculate a p-value.

Figure 4 and Table 3 demonstrate a difference in findings between the two (fictitious) groups. We use the student t test to evaluate the element of chance as the cause for this difference rather than the intervention.

The student t test formula (difference between sample means/estimated standard error of the difference) gives a "t value" which in conjunction with the number of degrees of freedom can be used to arrive at a p-value from appropriate statistical tables.

The standard error of the difference is calculated by: dividing the square of the standard deviation of group A by the number of subjects in group A then adding this number to the square of the standard deviation in group B divide by the number of subjects in group B

- ((7.8x7.8)/46)+((13.1x13.1)/54 = t value of 4.4

A t value of 4.4 with 98 degrees of freedom (df: n46 - 1 + n54 - 1) produces a p-value of <0.01. Therefore the null hypothesis could be rejected and antiemetic A claimed to have significantly faster onset than antiemetic B.

The standard error of the mean result suggests the results of group A are more likely to represent the mean of the target population than group B.