

Essential statistics for the pharmaceutical sciences

Dr Phil Rowe BSc, MSc, PhD, FRSS

Publisher: John Wiley & Sons Ltd (<http://eu.wiley.com>)

Published: 2007

ISBN: 978-0-470-03468-2

Additional topics as supplements to the book:

Selecting a statistical method

Selecting a statistical method

This document describes the analysis of experiments, trials or surveys where the questions to be answered concern possible cause and effect relationships. In statistics, the usual terminology is that we are investigating the possible effect of one or more 'Factors' on an 'Outcome' (See [Section 13.1.1 of book](#)). To identify an appropriate statistical test we need to follow these three stages in the order shown ...

1. Identify the factor(s) and the outcome
2. Determine what type of data would be used to record the above
3. Identify the appropriate test

An example – Effect of rainfall on fungal toxin contamination of a drug source

We will apply this approach to the analysis of the possible effect of rainfall on the fungal toxin contamination of drug-containing nuts (Described in [Section 14.2.2 of the book](#)).

Identify the factor(s) and the outcome.

If there is a cause and effect relationship it will be the rainfall that is the cause and the toxin the effect (Greater rain might cause more toxin, but more toxin is hardly likely to make it rain more!). So, rainfall is a factor and toxin the outcome. In this study, we are not checking for the effect of any other factor upon the toxin, so there is just one factor.

Determine what type of data would be used to record the factor(s) and outcome

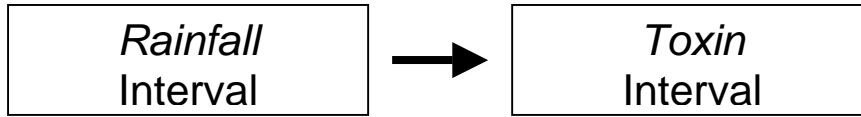
The factor (Rainfall) is a measured parameter capable of taking a range of values. Measured data of this type constitutes interval data ([See Section 1.2](#) for definition)

The outcome (Toxin concentration) is also a measurement – Interval data.

In summary: A single, interval factor may affect an interval outcome.

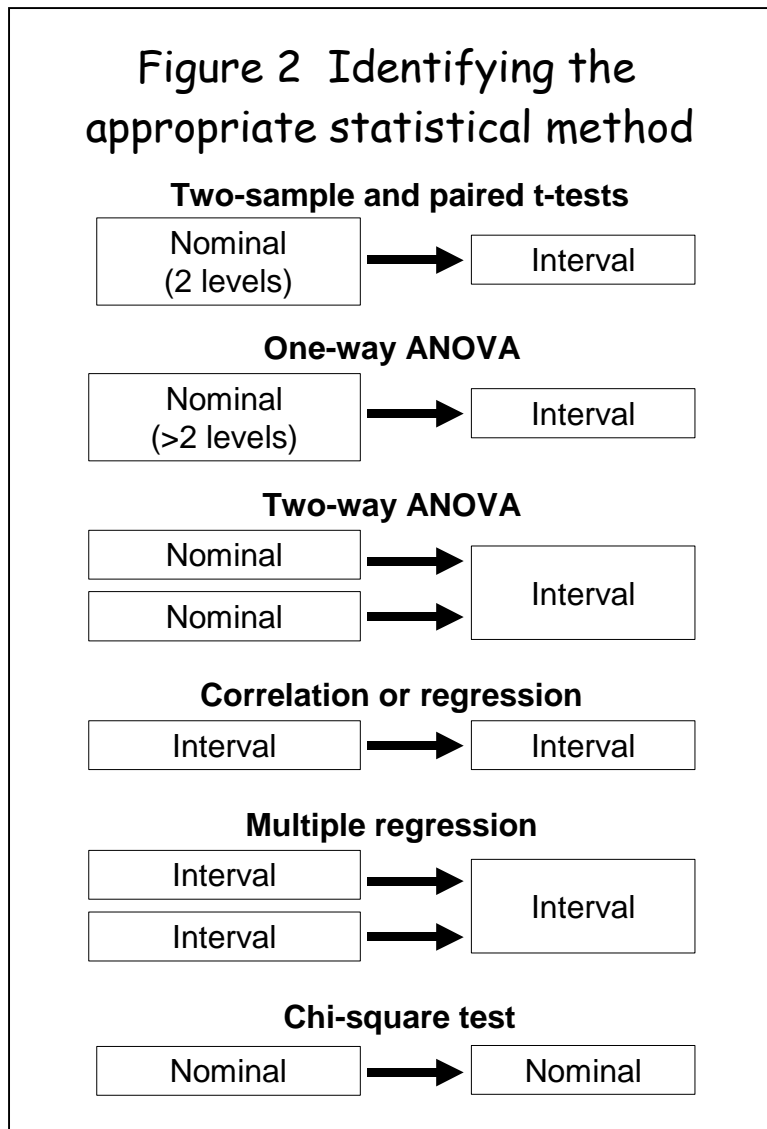
Diagrams such as Fig 1 will be used to summarise various experimental structures. The box on the left represents the factor under consideration, the arrow indicates a possible cause and effect relationship and the box on the right represents the outcome. A single box on the left implies that only one factor is under consideration; two boxes mean that we are investigating two or more factors.

Figure 1. A single, interval factor may affect an interval outcome



Identify the appropriate test

Figure 2, shows the types of factors and outcomes handled by various commonly used tests. Inspection of the figure shows that only correlation or regression analysis matches the pattern we have identified. Regression analysis would generate an equation allowing us to predict fungal toxin concentration from rainfall, whereas correlation would just tell us whether they are related without producing an equation. The aim of the work was to predict contamination of crops grown at various sites with known rainfall. As the equation will be required, we would select regression analysis.



Another example – effect of rifampicin on theophylline clearance

[Section 6.1.2](#) describes an investigation of the possible effect of rifampicin treatment on the clearance of theophylline. Proceeding as previously ...

Identify the factor(s) and the outcome.

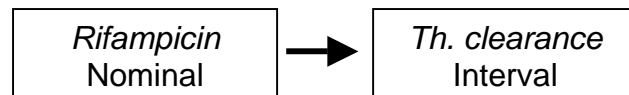
There is again just one factor (Rifampicin) that may affect the outcome (Theophylline clearance).

Determine what type of data would be used to record the factor(s) and outcome

Subjects' status with respect to rifampicin treatment would be recorded as a categorisation (Either 'Control' or 'Rifampicin treated'.) This is nominal data.

The outcome – theophylline clearance – is a measured variable capable of taking any value (within some reasonable range). This type of measurement data is interval in nature.

In summary: A single nominal factor may affect an interval outcome.



Identify the appropriate test

Fig 2 shows that both the t-tests and the one-way analysis of variance fit our requirements and some further details need to be considered.

First consider the number of levels that the factor can take. Rifampicin treatment can take just two levels (Control or Treated). Figure 2 show that one-way ANOVA is only used when there are more than two levels, so the appropriate method is a t-test.

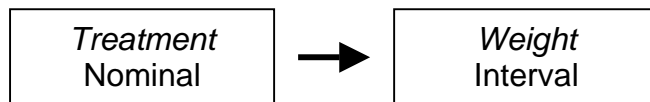
Finally we need to decide whether the experiment was carried out in a paired or unpaired manner. [Section 12.7](#) explains the nature of pairing. If we had used a single group of patients all of whom were studied on two occasions (Once under control conditions and once after rifampicin treatment), the data would be paired. But, here we used two separate sets of patients each of which received just one of the two treatments and there is no pairing; the two-sample t-test should be used.

Other examples

Paired t-test

[Section 12.1.1](#) describes a single group of subjects being weighed twice – once after placebo and once after drug treatment.

- The factor is drug treatment and the endpoint is weight.
- There is just the one factor.
- The factor (Drug treatment) is either placebo or active – nominal data
- The endpoint (weight) is a measured parameter – interval data.

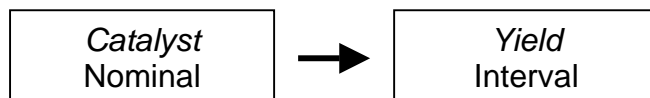


- Figure 2 suggests a t-test or one-way analysis of variance.
- The drug treatment takes two levels (Control or actively treated), so we need a t-test.
- The experimental structure is paired (Each subject studied under both conditions) so the final choice is a paired t-test.

One-way analysis of variance

[Section 13.2.1](#) describes a comparison of the efficiency of five different catalysts.

- The factor is type of catalyst and the endpoint is yield of product.
- There is just one factor.
- The factor (Catalyst) is nominal
- The endpoint (Yield) is a measured parameter – interval data

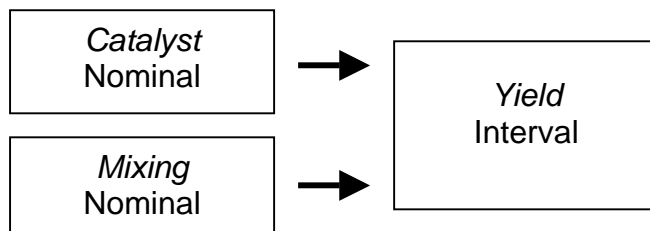


- Figure 2 suggests a t-test or one-way analysis of variance.
- The catalyst has five levels, so we need a one-way analysis of variance.

Two-way analysis of variance

[Section 13.3.1](#) describes an extension of the above experiment which simultaneously studied the effects of both five different catalysts and two different mixing methods on yield.

- There are two factors (Catalyst and mixing method) and the endpoint is yield
- The first factor (Type of catalyst) is nominal
- The second factor (Mixing method) is also nominal
- The endpoint (Yield of product) is interval

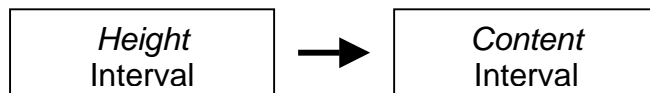


- Figure 2 indicates a two-way analysis of variance
- Whenever you decide that a two-way ANOVA is appropriate, a good check is that the experiment should have used a full factorial design (Studying all combinations of both factors). That is true for this experiment.

Correlation

[Section 14.1.5](#) describes a study of the possible relationship between the height at which leaves grow in a tree and the content of a drug

- The factor is height and the endpoint is drug content.
- There is just the one factor.
- The factor (Height) is interval
- The endpoint (Drug content) is interval



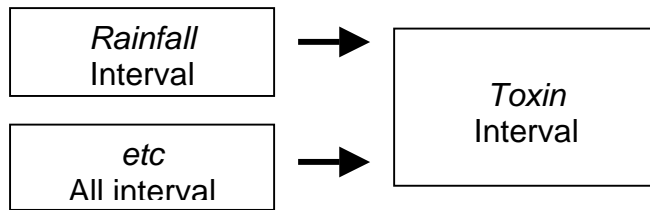
- Figure 2 suggests correlation or regression analysis.
- We only want to know *whether* these are related; we do not need an equation. Correlation analysis is adequate.

Sometimes correlation analysis is applied to two sets of data which we suspect may be associated but we do not believe there to be any cause and effect relationship. In such cases the important thing is that we have two sets of interval data; there is no need to try to specify which is factor and which outcome.

Multiple regression

[Section 14.3.2](#) describes a study of the possible effects of four meteorological factors (Rainfall, Temperature, Sunshine and Wind) on fungal toxin contamination.

- There are four factors (as above) and the endpoint is toxin
- All four factors are measured variables - interval
- The endpoint (Toxin) is interval

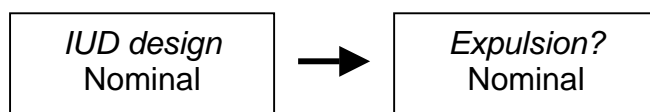


- Figure 2 indicates multiple regression analysis.
- Note that Fig 2 shows multiple regression with just two factors, but there can perfectly well be any number.

Chi-square test

[Section 16.1.1](#) describes a trial comparing rates of expulsion of two designs of IUD.

- The factor is IUD design and the endpoint is expulsion.
- There is just the one factor.
- The factor (Design of IUD) is nominal
- The endpoint (Expelled or Not expelled) is also nominal



- Figure 2 indicates a contingency chi-square test.
- Whenever you decide that a contingency chi-square test is appropriate, a good check is that the results should be capable of being presented in a contingency table. That is true for this trial (See Table 16.1).

Sometimes the contingency chi-square test is applied to two sets of data which we suspect may be associated but we do not believe there to be any cause and effect relationship. In such cases the important thing is that we have two sets of nominal data; there is no need to try to specify which is factor and which outcome.

Non-parametric tests

Any of the above tests that involve interval scale data will have an assumption that such data is normally distributed and in the case of t-tests and analyses of variance, there will also be an assumption that the sets of data being compared have similar standard deviations.

If the data fails to meet these requirements of normality and similar standard deviations

If data transformation ([Section 17.1](#)) cannot resolve the problem, there are non-parametric tests that duplicate much of the functionality of the parametric tests mentioned above. See [Sections 17.2 – 17.4](#) for details of the following tests which are equivalent to ...

- Mann-Whitney (Two-sample t-test)
- Wilcoxon paired samples (Paired t-test)
- Kruskal-Wallis (One-way analysis of variance)
- Spearman correlation (Pearson correlation)

If measurement data is ordinal in nature (Section1.3)

You are probably best to use a non-parametric test as above.

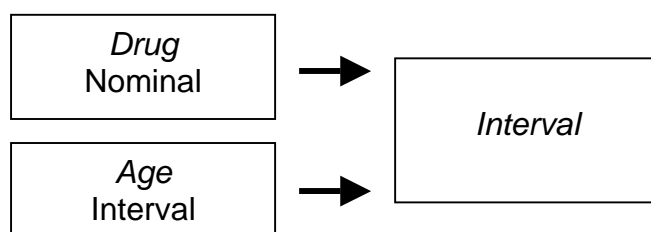
Experimental designs not covered in the book

It is pretty obvious that there are many combinations of factors and outcomes which are not covered by Fig 2. The appropriate approach to two strikingly obvious examples are described below.

Analysis of covariance

All of the tests in Fig 2 either consider a single factor or if there are several factors, they are all of the same type (All nominal or all interval). However, there are situations where a measured outcome may be influenced by both a nominal and an interval factor. For example, in a study comparing blood pressure reduction with two different drugs we may suspect that the response is also affected by the patients' ages.

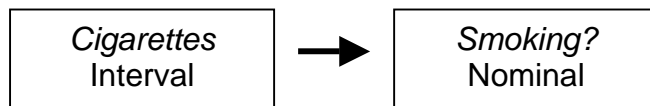
- The outcome is reduction in blood pressure – Interval
- First factor affecting outcome is type of drug – Nominal
- Second factor affecting outcome is age - Interval



This should be analysed by an analysis of covariance which is [fully described](#) in another document on this site.

Logistic regression

Another combination not considered is the possible effect of a measured (Interval) factor upon a categorical (Nominal) endpoint. For example, a smoking cessation programme might have an endpoint of smoking status three months after completion of the programme with outcomes recorded as 'Smoking' or 'Not smoking' which forms nominal data. One factor that might well have an influence upon participants' success is how heavily they smoked prior to entry onto the programme. If this was recorded as number of cigarettes smoked per day this would be nominal data.



This should be analysed by a logistic regression which is [fully described](#) in another document on this site.