5

OVERVIEW OF SEVERAL EXPERIMENTAL DESIGNS AND ANALYSES.

There is a close relationship between the design of an experiment and the statistical methods that will be used to analyse the resulting data. A brief overview of the most common experimental designs and analyses are thus presented in the next two sections. From The Design of Animal Experiments by Festing *et al (2002)* – section A – and Intuitive Biostatistics by Motulsky (1995) –Section B.

A) EXPERIMENTAL DESIGN

The purpose of experimental design is to provide guidelines on data collection in such a way that causation may be established beyond a reasonable doubt and the residual variation minimised so that most of the variation is explained by the variables that are deliberately manipulated. The designs may also need to incorporate some extra considerations, such as the efficient use of more than one sex or strain, so that the interpretation of the results may make it possible to draw wider conclusions about the underlying science. A summary of the designs most commonly used in biomedical research is given in the Table on the next page.

Principles of Design and Analysis : some reminders

Good experimental design relies on two principles.

Replication:. The more times something is repeated, the greater the confidence of ending up with a genuine result. For example, with two choices of routes to drive into work, each of about the same distance, somebody would usually want to drive each more than once before being happy to choose one route over another. However, this is not a justification for running very large experiments. As chapter three explained, replication should be across the source of variation which is likely to be the biggest cause of variability in the experiment. The determination of sample size is discussed in more detail in Chapter 5. Replication also makes it possible to detect abnormal observations or outliers, perhaps due to errors of measurement or recording. Statistical analysis is used to make estimates of the probability of error in reaching conclusions from the data, and in particular to prevent claims of an important treatment effect when the results could well be due to sampling variation.

Randomisation: Experimental subjects (e.g. animals) must be allocated to treatment groups at random, and the order in which subjects are tested must be at random in order to avoid biases which would be almost certain to occur with any other method of allocation.

Considerations in the Choice of design

The following Table summarises some of the more common experimental designs used in biology. All researchers should be aware of these before embarking on experiments with animals.

Starting with the very simplest design (a completely randomised, single-factor design), there have been two major ways in which more complex designs have developed:

1. Designs for Improving precision: The precision of an experiment is increased by better control of variation and/or by taking account of some natural structure in the experimental material. Randomised block and Latin square designs eliminate some sources of heterogeneity by choosing subsets of uniform experimental units which are then assigned at random to the treatment groups. Crossover designs involve sequential experimentation on the same subjects, which should eliminate some inter-individual variation.

2. Designs for Increasing the amount of information from each experiment : Factorial designs, which involve two or more independent variables such as some treatments and both sexes, or several treatments and termination times will usually produce more information without necessarily increasing the size of the experiment. The aim of these designs is to find out the effects of each factor separately, and their joint effects. Advanced versions of these designs can be used, example, to find the optimum combination of a large number of factors. For a given input of resources (animals, reagents, time etc.), factorial designs will normally provide more information than a single factor design, at little or no extra cost. There seems to be a widespread misunderstanding among scientists that if an experiment needs, say, 24 male mice, then an experiment involving both males and females will need 48 mice. This is not true. In many cases the experiment involving 24 male mice could be done using 12 male and 12 female mice (a factorial design) with little loss of precision, and with a useful increase in the amount of scientific information which is produced.

Summary of the most common experimental designs
(With slight modifications from **The Design of Animal Experiments by Festing, Overend, Gaines Das, Cortina-Borja & Berdoy (2002)**

Type of Design	When to use	Advantage	Disadvantage	Notes
Single factor design (fixed effects; random effects)	*Fixed effects*: when uncontrollable sources of variation are unlikely to be important *Random effects:* when simply interested in quantifying source of random variation (e.g. surveys) rather than compare treatments. Useful if we have within-subject measures	Simplest "designs". Easy to use and analyse. Less affected by unequal sample size.	No control of additional "nuisance" variation caused by uncontrolled time and space variation etc which might affect results.	Typical designs used when doing simple *t*-tests.
Block designs (including Latin square)	When wanting to remove the effects of known or suspected "nuisance" variation. The evidence shows that it is under used in the bio-medical literature, particularly in its simplest form (controlling at least *one* source of unwanted variation).	Deals with known or suspected "nuisance" variation in a systematic and powerful way by breaking the experiment into a series of subunits which are analysed as a whole. Unequal sample sizes present some problems.	Can be sensitive to missing values. Becomes increasingly complex when dealing with two (Latin square) or three sources (Graeco-Latin squares) of unwanted variation.	These designs aim to improve the *precision* of the experiment. They can be used in conjunction with others (e.g. factorial designs). Analysis of covariance, another way of increasing precision.
Factorial Designs	When wanting to increase the "generality" of the results, by testing *simultaneously* the potential effect of several factors (e.g. treatment, sex, strain) and their interactions on the response (dependent variable[s])	A more powerful alternative to doing several smaller experiments for each factor. Allows testing for interactions between these factors	*More a note of caution than a disadvantage: beware of understanding the biological significance of interactions; Interaction between more than two factors are often difficult to interpret.*	These designs aim to increase the *amount of information* yielded by the experiment. They can be used in conjunction with other (e.g. blocking designs)
Repeated measure designs (cross over, split plot, mixed effects designs)	When using several measures on the same individual, either because we are interested in change over time or as a way to deal with strong individual variation.	A more powerful alternative to the dreaded repeated t-tests! The individual can act as its own control (cross over)	Crossover: not valid if there is a strong order effect.	High precision as a result of eliminating inter-individual variation.
Sequential designs	When results with individual the experimenter wants to stop collecting data as soon as possible, perhaps because the use of animals is difficult (e.g. labour intensive, costly) or involves important welfare issues.	Results are analysed as the experiment unravels, enabling the experimenter to stop as soon as significance is obtained.	Requires expert advice. Logistics may be a problem.	Up-and-down method may be replacing the classical LD50 test in te USA

b) ANALYSIS: CHOOSING A STATISTICAL TEST

"Which test should I use?". The following, from H. Motulsky 1995, is a nice summary with helpful explanations and tips. It is worth highlighting however (not clear in the text) that parametric tests rely on two assumptions: Normal (i.e. Gaussian) distribution and homogeneity of variance - people tend to ignore the second. Note also that whilst one might therefore feel that the non-parametric route is the easy route (and in some ways it is), parametric tests are worth the effort, even if it means transforming the data to satisfy the above assumptions, because they allow the testing of several variables and their interactions all at once (ANOVA). When this is appropriate for your data, this is a more powerful way of analysis…and will allow you to detect an effect that you may otherwise have missed.

This extract can also be found on: www.graphpad.com/www/Book/Choose.htm

REVIEW OF AVAILABLE STATISTICAL TESTS

Choosing the right test to compare measurements is a bit tricky, as you must choose between two families of tests: parametric and nonparametric. Many -statistical test are based upon the assumption that the data are sampled from a Gaussian distribution. These tests are referred to as parametric tests. Commonly used parametric tests are listed in the first column of the table and include the t test and analysis of variance.

Tests that do not make assumptions about the population distribution are referred to as nonparametric- tests. You've already learned a bit about nonparametric tests in previous chapters. All commonly used nonparametric tests rank the outcome variable from low to high and then analyze the ranks. These tests are listed in the second column of the table and include the Wilcoxon, Mann-Whitney test, and Kruskal-Wallis tests. These tests are also called distribution-free tests.

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: THE EASY CASES

Choosing between parametric and nonparametric tests is sometimes easy. You should definitely choose a parametric test if you are sure that your data are sampled from a population that follows a Gaussian distribution (at least approximately). You should definitely select a nonparametric test in three situations:

- The outcome is a rank or a score and the population is clearly not Gaussian. Examples include class ranking of students, the Apgar score for the health of newborn babies (measured on a scale of 0 to 1O and where all scores are integers), the visual analogue score for pain (measured on a continuous scale where 0 is no pain and 10 is unbearable pain), and the star scale commonly used by movie and restaurant critics (* is OK, ***** is fantastic).
- Some values are "off the scale," that is, too high or too low to measure. Even if the population is Gaussian, it is impossible to analyze such data with a parametric test since you don't know all of the values. Using a nonparametric test with these data is simple. Assign values too low to measure an arbitrary very low value and assign values too high to measure an arbitrary very high value. Then perform a nonparametric test. Since the nonparametric test only knows about the relative ranks of the values, it won't matter that you didn't know all the values exactly.
- The data are measurements, and you are sure that the population is not distributed in a Gaussian manner. If the data are not sampled from a Gaussian distribution, consider whether you can transform the values to make the distribution become Gaussian. For example, you might take the logarithm or reciprocal of all values. There are often biological or chemical reasons (as well as statistical ones) for performing a particular transform.

Table 37.1. Selecting a statistical test

	Type of Data			
Goal	**Measurement (from Gaussian Population)**	**Rank, Score, or Measurement (from Non-Gaussian Population)**	**Binomial (Two Possible Outcomes)**	**Survival Time**
Describe one group	Mean, SD	Median, interquartile range	Proportion	Kaplan Meier survival curve
Compare one group to a hypothetical value	One-sample *t* test	Wilcoxon test	Chi-square or Binomial test **	
Compare two unpaired groups	Unpaired *t* test	Mann-Whitney test	Fisher's test (chi-square for large samples)	Log-rank test or Mantel-Haenszel*
Compare two paired groups	Paired *t* test	Wilcoxon test	McNemar's test	Conditional proportional hazards regression*
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test	Chi-square test	Cox proportional hazard regression**
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test	Cochrane Q**	Conditional proportional hazards regression**
Quantify association between two variables	Pearson correlation	Spearman correlation	Contingency coefficients**	
Predict value from another measured variable	Simple linear regression or Nonlinear regression	Nonparametric regression**	Simple logistic regression*	Cox proportional hazard regression*
Predict value from several measured or binomial variables	Multiple linear regression* or Multiple nonlinear regression**		Multiple logistic regression*	Cox proportional hazard regression*

* Only briefly mentioned in book ** Not discussed in book

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: THE HARD CASES

It is not always easy to decide whether a sample comes from a Gaussian population. Consider these points:

• If you collect many data points (over a hundred or so), you can look at the distribution of data and it will be fairly obvious whether the distribution is approximately bell shaped. A formal statistical test (Kolmogorov-Smirnoff test, not explained in this book) can be used to test whether the distribution of the data differs significantly from a Gaussian distribution. With few data points, it is difficult to tell whether the data are Gaussian by inspection, and the formal test has little power to discriminate between Gaussian and non-Gaussian distributions.

• You should look at previous data as well. Remember, what matters is the distribution of the overall population, not the distribution of your sample. In deciding whether a population is Gaussian, look at all available data, not just data in the current experiment.

• Consider the source of scatter. When the scatter comes from the sum of numerous sources (with no one source contributing most of the scatter), you expect to find a roughly Gaussian distribution. When in doubt, some people choose a parametric test (because they aren't sure the Gaussian assumption is violated), and others choose a nonparametric test (because they aren't sure the Gaussian assumption is met).

CHOOSING BETWEEN PARAMETRIC AND NONPARAMETRIC TESTS: DOES IT MATTER?

Does it matter whether you choose a parametric or nonparametric test? The answer depends on sample size. There are four cases to think about:

• Large sample. What happens when you use a parametric test with data from a nongaussian population? The central limit theorem (discussed in Chapter 5) ensures that parametric tests work well with large samples even if the population is non-Gaussian. In other words, parametric tests are robust to deviations from Gaussian distributions, so long as the samples are large. The snag is that it is impossible to say how large is large enough, as it depends on the nature of the particular non-Gaussian distribution. Unless the population distribution is really weird, you are probably safe choosing a parametric test when there are at least two dozen data points in each group.

• Large sample. What happens when you use a nonparametric test with data from a Gaussian population? Nonparametric tests work well with large samples from Gaussian populations. The P values tend to be a bit too large, but the discrepancy is small. In other words, nonparametric tests are only slightly less powerful than parametric tests with large samples.

• Small samples. What happens when you use a parametric test with data from nongaussian populations? You can't rely on the central limit theorem, so the P value may be inaccurate.
• Small samples. When you use a nonparametric test with data from a Gaussian population, the P values tend to be too high. The nonparametric tests lack statistical power with small samples. Thus, large data sets present no problems. It is usually easy to tell if the data come from a Gaussian population, but it doesn't really matter because the nonparametric tests are so powerful and the parametric tests are so robust. Small data sets present a dilemma. It is difficult to tell if the data come from a Gaussian population, but it matters a lot. The nonparametric tests are not powerful and the parametric tests are not robust.

FISHER'S TEST OR THE CHI-SQUARE TEST?

When analyzing contingency tables with two rows and two columns, you can use either Fisher's exact test or the chi-square test. The Fisher's test is the best choice as it always gives the exact P value. The chi-square test is simpler to calculate but yields only an approximate P value. If a computer is doing the calculations, you should choose Fisher's test unless you prefer the familiarity of the chi-square test. You should definitely avoid the chi-square test when the numbers in the contingency table are very small (any number less than about six). When the numbers are larger, the P values reported by the chi-square and Fisher's test will he very similar.
The chi-square test calculates approximate P values, and the Yates' continuity correction is designed to make the approximation better. Without the Yates' correction, the P values are too low. However, the correction goes too far, and the resulting P value is too high. Statisticians give different recommendations regarding Yates' correction. With large sample sizes, the Yates' correction makes little difference. If you select Fisher's test, the P value is exact and Yates' correction is not needed and is not available.

REGRESSION OR CORRELATION?

Linear regression and correlation are similar and easily confused. In some situations it makes sense to perform both calculations. Calculate linear correlation if you measured both X and Y in each subject and wish to quantity how well they are associated. Select the Pearson (parametric) correlation coefficient if you can assume that both X and Y are sampled from Gaussian populations. Otherwise choose the Spearman nonparametric correlation coefficient. Don't calculate the correlation coefficient (or its confidence interval) if you manipulated the X variable.
Calculate linear regressions only if one of the variables (X) is likely to precede or cause the other variable (Y). Definitely choose linear regression if you manipulated the X variable. It makes a big difference which variable is called X and which is called Y, as linear regression calculations are not symmetrical with respect to X and Y. If you swap the two variables, you will obtain a different regression line. In contrast, linear correlation calculations are symmetrical with respect to X and Y. If you swap the labels X and Y, you will still get the same correlation coefficient.

PAIRED OR UNPAIRED TEST?

When comparing two groups, you need to decide whether to use a paired test. When comparing three or more groups, the term paired is not apt and the term repeated measures is used instead. Use an unpaired test to compare groups when the individual values are not paired or matched with one another. Select a paired or repeated-measures test when values represent repeated measurements on one subject (before and after an intervention) or measurements on matched subjects. The paired or repeated-measures tests are also appropriate for repeated laboratory experiments run at different times, each with its own control.
You should select a paired test when values in one group are more closely correlated with a specific value in the other group than with random values in the other group. It is only appropriate to select a paired test when the subjects were matched or paired before the data were collected. You cannot base the pairing on the data you are analyzing.