Experimental Design[1]

John Maindonald

Statistical Consulting Unit of the Graduate School,
Australian National University

> The methods of science, with all its imperfections, can be used to improve social, political and economic systems, and this is, I think, true no matter what criterion of improvement is adopted. How is this possible if science is based on experiment? Humans are not electrons or laboratory rats. But every act of Congress, every Supreme Court decision, every Presidential National Security Directive, every change in the Prime Rate is an experiment. Every shift in economic policy, every increase or decrease in funding for Head Start, every toughening of criminal sentences is an experiment. Exchanging needles, making condoms freely available, or decriminalizing marijuna are all experiments. . . . In almost all these cases, adequate control experiments are not performed, or variables are insufficiently separated. Nevertheless, to a certain and often useful degree, such ideas can be tested. The great waste would be to ignore the results of social experiments because they seem to be ideologically unpalatable. [Sagan 1997, pp. 396-397]

> *There is no more effective way to settle a disputed question than to do an experiment, when an experiment is possible. When fire-walkers walk across hot charcoal and emerge unharmed, it provides a pretty effective demonstration that such a thing is possible. When one plant grows like crazy in a bed of compost, while its neighbour has no compost and wilts, it seems like a convincing demonstration that compost helps growth. It seems convincing even though this is a rather poorly designed experiment.*
>
> *The aim of experimental design is to ensure that the experiment is able to detect the treatment effects that are of interest, and that it uses available resources to get the best precision possible. The choice of design can make a huge difference.*

We wish to compare two technicians who will use a pressure tester to compare apple firmness. How should we do the comparison? Should we give the testers separate samples of perhaps twenty apples? Or should we use one sample of twenty apples, with both technicians making firmness measurements on each apple?

In a clinical trial that compares two different therapies for treating arthritis, right and left hand grip strength will be among the outcome measurements. The measurements are highly variable. Is it useful to increase the precision by making repeated grip strength measurements? Or is the variation in measured grip strength for an individual patient of minor consequence relative to variation between patients? If it turns out to be useful to make repeated measurements on individual patients, should the repeat measurements be made at the same session, or at different sessions that are separated by a few days or weeks?

We plan on undertaking an experiment in which trays of fruit are the experimental unit. In each of several coolstores, different treatments will be applied to different trays. Should we

[1] These notes have still to be converted to a free-standing document. At present the figure numbers start with Fig. 13.

opt for lots of trays with a small number of fruit on each, or for a small number of trays with a large number of fruit on each? Which is the better design?

The initial discussion will focus on two widely used types of experimental design – Completely Randomised Designs, and Randomised Block Designs. The emphasis will be on designing experiments so that we get the best possible value for the resources used. There'll also be a brief introduction to incomplete block designs, both balanced and approximately balanced.

I begin with definitions of the terminology that will be used through this chapter.

The Language of Experimental Design

You will learn about

a. treatment units and measurement units. They are not necessarily the same!

b. randomisation, especially as opposed to haphazard assignment of treatments

c. replication – genuine replication, effective replication and bogus replication

d. blocking and other forms of local control

e. levels of variation – these are sometimes called strata.

12.1 Multiple Levels of Variation – Blocks

Let us first of all remind ourselves of the issue that arises when we make multiple measurements on an experimental unit. We have then introduced another, lower, level of variation – within experimental units as well as between experimental units. One can also group experimental units together into blocks. Where experimental units are grouped together into blocks, blocks become another, now higher, level of variation. The simplest type of one-factor block design, the randomised complete block design, has one experimental unit from each of the treatment levels in each block, e.g.

	Block 1	Block 2	Block 3
Treatments	A, B, C	A, B, C	A, B, C

N. B. Treatments should be randomly allocated to experimental units, independently for each block

Also possible are block designs where a subset of the treatments appear in each block. For example, we might have

	Block 1	Block 2	Block 3
Treatments	A, B	B, C	C, A

One treatment has been left out in each block, in a balanced way. This is a *balanced incomplete block* design. I have used this type of design for comparing the readings of different firmness testing devices on the same fruit. Each fruit was in effect a block. We did two sets of two readings, one pair with each of the devices, on the one fruit.

Block designs are widely used in agriculture, where the aim is to maximise the precision of treatment comparisons. Thus each block is chosen to be as uniform as possible. In the simplest form of randomised block design, all treatments occur once in each block. Blocks should be sampled from the wider population to which it is intended to generalise results, so that they might be on different sites.

In clinical trials blocks are more often used as a way of making it hard to predict treatment allocations for individual patients. Allocation of treatments to patients is random within

blocks, subject to devices that achieve a roughly equal numbers in the different treatments. (ICH 1998, p.21). Where a surgical trial involves several different surgeons, blocking may be highly desirable as a mechanism for controlling variation. The patients that are allocated to a surgeon form a block, with random allocation to treatments within those blocks.

12.2 Trade-Offs From Different Design Possibilities

In order to illustrate some of the different design possibilities, and the possible impact on precision, I will demonstrate two ways to do a taste experiment. Section 12.5 will extend these ideas further, with a further example.

The Standard Deviation of a Difference

If you take two independent samples of size *n*, each from distributions with standard deviation σ, then

(i) $SD[x_1 - z_1] = \sqrt{2}\,\sigma \quad (= \sqrt{\sigma^2 + \sigma^2})$

(ii) $SE[\,\bar{x} - \bar{z}\,] = \sqrt{2}\,\sigma / \sqrt{n}$

A Simple Taste Experiment

The first is a completely randomised design, of a kind that is sometimes used in clinical trials. It was a taste experiment. To get an indication of what panelists think about the sweetness of a product, they are asked to mark off their response on a so-called Likert scale, thus:

Not sweet enough Too sweet

1 3 x 5 7 9

One uses a ruler to read off the results. One way to make this easy is to place the 1 at 10mm, the 30 at 30mm, and so on. The x is at about 36mm. A reasonable way to do the experiment is to give each person both products. Here then is a set of results from such an experiment:

```
   Person 1  2  3  4   5  6   7  8  9 10 11 12 13  14 15 16 17
4 units  72 74 70 72  46 60  50 42 38 61 37 39 25  44 42 46 56
1 unit   58 69 60 60  54 57  61 37 38 43 34 14 17  54 32 22 36
Diff.    14  5 10 12  -8  3 -11  5  0 18  3 25  8 -10 10 24 20
```

The `units' were amounts of an additive. Here (Fig. 13) are boxplots that show the spread of results for the two products and for the differences:

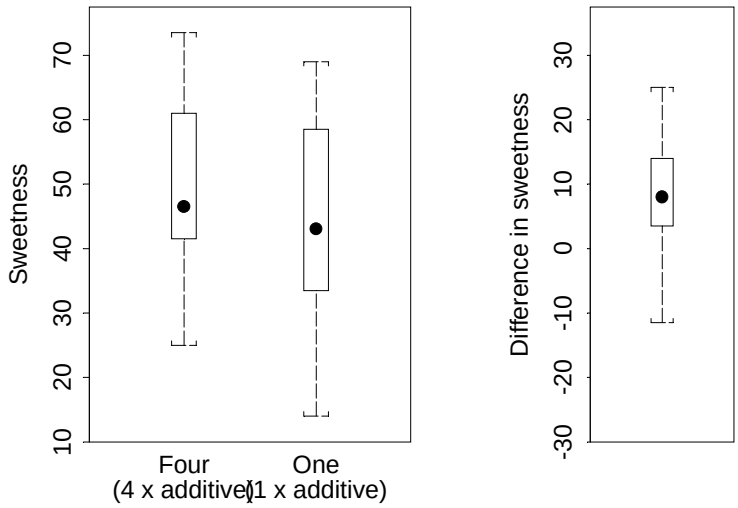

Fig. 13: Distributions of assessments for the two individual products (left panel)compared with the distribution of the difference (right panel Correlation between the two sets of results leads to a small standard deviation for the difference.

Notice the very much smaller spread of values for the differences.

Another way to do the experiment would have been to take 34 people, choose 17 people at random and given them the first product, and give the other product to the remainder. We thus have two possible types of experiment. The alternatives – an "independent samples" or *completely randomised design*, and a *paired comparison design*, are shown diagrammatically in Figures 14 and 15.

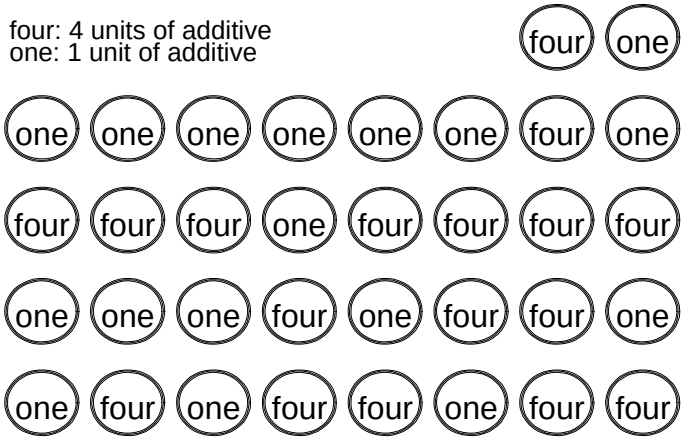

Fig. 14: Completely randomised (independent samples) design, i. e. the 34 tasters are allocated at random to one of two groups, in such a way that there are 17 in each group. Those labelled `one' get one unit of additive, while those labelled `four' get 4 units of additive.

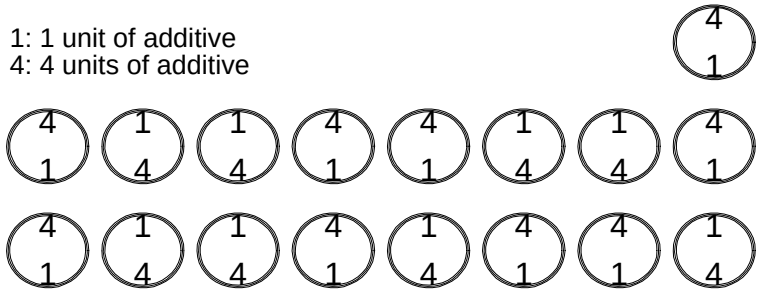

Fig. 15: Paired comparisons (Dependent sanples) design. Here each of 17 tasters is given both products. The order of presentation should be random. [N.B.: This is a simple example of a block design, with 2 treatments per block.]

We have enough information, in the data from the paired comparison experiment, to compare the precision of the two alternative experimental designs. In the individual samples experiment, the SE is $\sqrt{2}$ SD/$\sqrt{17}$ = 3.8 .

Independent Samples Experiment (2 groups of 17 panelists)	Paired Samples Experiment (17 panelists)
SD = 15.7 (pooled estimate; individual SDs are 14.6, 16.8)	SD = 10.7 (This is the SD of the 17 sample differences.)
SE = 3.8 (individual SEs are 3.6, 4.1)	SE of differences = 2.66 (i.e. 10.7/$\sqrt{17}$)
SED = $\sqrt{3.8^2 + 3.8^2}$ = 5.46	

Under what conditions is the paired comparison experiment better? The answer hinges on the correlation between the two sets of results. Fig. 16 summarises this information. If there is a strong correlation, then it pays to pair, or to "match".

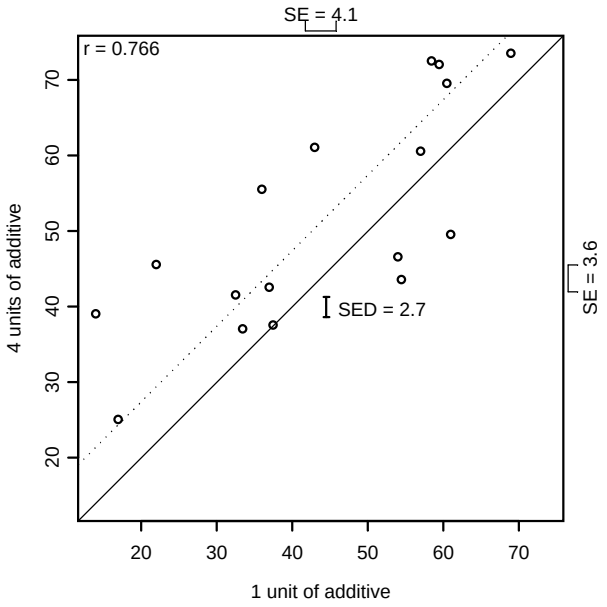

Fig. 16: Sweetness for the sample `four' (4 units of additive) versus Sweetness for the sample `one' (1 unit of additive).

It is useful to look at this experiment in the light of the experimental design principles.

1. It must be possible to assess the consistency of treatment comparisons, when the experiment is repeated. (There should be a valid estimate of the standard error of treatment effects.)
2. Results should be free of bias.
3. Results should be as precise as possible, given the available resources of time and materials.

A good way to demonstrate that results are repeatable is, not surprisingly, to repeat the experiment. Generally all we insist is that the experimenter assess the extent to which results are consistent then they repeat the experiment. At this point it becomes reasonable to expose the work to a harsher test – can other experimenters get similarly consistent results that point in the same direction?

> Long ago Fisher . . . recognised that solid knowledge came from a demonstrated ability to repeat experiments . . . This is unhappy for the investigator who would like to settle things once and for all, but consistent with the best accounts we have of the scientific method. . . .
> [Tukey 1991.]

Repeatability is right at the heart of science. There are however weak and strong repeatability tests. The practice of statistics is beginning to reflect distinctions that, in the past, have often been ignored.

Freedom from bias is achieved by making the two samples look totally alike – by what the clinical trials people call concealment.

The matched pairs design greatly assisted precision. I have described a simple example of blocking. One compares treatments under conditions that are as nearly identical as possible.

12.3 Randomised Controlled Trials

> One day when I was a junior medical student, a very important Boston surgeon visited the school and delivered a great treatise on a large number of patients who had undergone vascular reconstructions. At the end of the lecture a young student at the back of the room timidly asked, "Do you have any controls?" Well, the great surgeon drew himself up to his full height, hit the desk, and said, "Do you mean did I not operate on half the patients?" The hall grew very quiet then. The voice at the back of the room very hesitantly replied, "Yes, that's what I had in mind." Then the visitor's fist really came down as he thundered "Of course not. That would have doomed half of them to their death." It was absolutely silent then, and one could scarcely hear the small voice ask, "Which half?"
> [Peacock, E. E. 1972. Medical World News Sept 1 1972, p.45.]

> What makes it possible to write a long article on controversies in controlled clinical trials without writing a much longer article on uncontrolled trials or uninvestigated therapies? Essentially this paradox arises because in controlled trials we have a model of perfection and we can discuss departures from it with ease. Without such a model, one tends to emphasise only major difficulties --- having swallowed a camel, why strain at a gnat?
> [Mosteller, Gilbert & Lewis, p. 14, in Shapiro & Lewis 1983.]

By contrast with much agricultural experimentation, the design of a randomised controlled trial may be very simple in concept. In a randomised controlled trial with two treatment groups, subjects are randomly assigned to one or other treatment. The randomisation commonly designed so that roughly equal numbers of patients are assigned to each treatment. Complications arise from the ethical and logistical difficulties of conducting a properly designed clinical trial. It is ethical to involve

patients in a clinical trial only if it is unclear which is the most effective treatment. A clinical trial is both ethical and necessary in those cases where there are differences of opinion among medical specialists.

A minor elaboration of the two-sample trial arises when subjects are matched, or when treatment comparisons can be made within subjects. In this case it may be possible to perform the analysis on the difference between the responses or on log(ratio) of the reponses, or on some other measure of the difference. The analysis then reduces to a single sample analysis.

The simplicity of the analysis, and perhaps the simplicity of interpretation, may be compromised when an adjustment for covariate effects seems necessary. There are particular problems of interpretation when results are different depending on whether or not there is a covariate adjustment.

Proper controls are essential. Synthetic estrogen (diethylstilbestrol or DES) injections in pregnancy were at one time thought to prevent miscarriage A randomised double-blind study published in 1953 showed no effect, compared with placebo injections (Dickmann and Davis 1953). This result seems to have attracted little attention, and DES continued in use for another two decades or more. This unproved therapy later proved to give an excess of cases of vaginal carcinoma and of breast cancer (Gehan & Lemak 1994, p.159.) Irwig et al. (1999, pp.7-11 and elsewhere) give other such examples.

Randomised controlled trials where there is matching provide a simple example of a block design. The individuals who are matched form a single block.

The Importance of Strict Protocols

Randomisation ensures that all units have an equal chance of receiving all treatments. This reduces opportunities for unconscious bias in the assignment of subjects to treatments. Also important may be procedures that reduce opportunities for bias when treatments are
applied or results are assessed. In a clinical trial, the ideal is that neither patients nor clinicians should know which treatment has been used. The broad term `concealment' is used for devices that ensure this. The double blind randomised controlled trial, where
allocations are randomised and neither patient not doctor knows which treatment has been assigned, sets the standard for clinical trials.

Some Noteworthy Clinical Trials

We have come a huge distance from the standards of evidence of earlier centuries. Often effective comparative evidence of the effectiveness of treatments was provided as a result of chance occurrence. Thus Ambroïse Paré (1510-1590) believed from what he had read that gunshot wounds ought to be cauterised with scalding hot oil of elders to which a little theriac had been added. When unable to get oil of elders, he applied a dressing made of yolk of egg, oil of roses and turpentine. To his surprise, those who were given the makeshift treatment fared much better.

John Hunter (1728-1793) described five cases in which, because of delay in seeing the patients, he had been unable to follow his standard practice of removing the musket balls or shrapnel from gunshot wounds. All wounds healed promptly, and Hunter discovered that ``balls seldom or ever did any harm when at rest''. One of the earliest published studies that has a claim to be a clinical trial was conducted by British navy

surgeon James Lind in 1753. Because of the small numbers of subjects, present day researchers might prefer to call it a pilot clinical trial. Lind assigned two scurvy victims to each of six treatments

1. 1 quart of cider per day
2. 25 gutts of elixir vitriol 3x daily
3. 2 spoonfuls of vinegar 3x daily
4. ½ pint of seawater per day
5. 2 oranges & 1 lemon per day
6. an electuary (garlic, mustard seed, . . .) recommended by a hospital surgeon.

The seamen on the oranges and lemon did best, to the extent that one of them was able to resume duty, and the other was appointed nurse to the remaining 10 scurvy patients. The two seamen who had been given cider showed some improvement. In spite of the clear win for the orange and lemon treatment, Lind continued to recommend the standard `dry air' treatment.

Pierre Charles Alexander Louis (1835), compared the outcomes for pneumonia patients who had been bled with the outcome for patients who had not been bled. He concluded that there was no appreciable difference in mortality or in duration or in duration of illness or in other clinical indicators. His results were so far out of line with general opinion that he had misgivings about publishing them.

Note the consistency with which these early trials, with all their defects, overturned conventional medical wisdom. Those early experimenters were by and large not willing to allow their experimental results to challenge that wisdom. They were inclined to argue that the fault lay with the experimental method, or with their use of it. Indeed the experimental method used for clinical trials has required huge refinement to bring it to the point where it is now a credible instrument for comparison of treatments.

The first random allocation of treatments may have been that of Amberson (1931), in a study of sanocrysin in the treatment of pulmonary tuberculosis. Amberson took two carefully matched groups of ten patients each, then tossing a coin to decide which patient received which treatment. two carefully matched groups of 12 patients each,

Diehl et al.'s (1938) trial may have been the first trial that randomly assigned individuals to treatments. A total of 1640 volunteer students were each assigned to one of four treatments for the common cold – three different vaccines and a placebo. Moreover it was a double blind trial.

Randomised controlled trials did not become common till the 1950s. A major stimulus for the conduct of clinical trials in the United States arose from 1962 Kefauver-Harris amendments to the United States Food, Drug and Cosmetic Act of 1938. Approval of a drug for human use was to require ``adequate and well-controlled investigations".

Today standards for clinical trials are under continual review, a result of extensive and well-documented evidence of the misleading results that may be obtained when trials do not follow strict protocols. There is now a huge literature that gives advice on the conduct of clinical trials. See in particular Begg et al. (1996), ICH (1998) and related ICH documents, Piantadosi (1997) and Senn (2000),

There are difficult conduct and analysis issues – the ethics of random allocation, and the use of covariate adjustments – on which there may never be complete agreement.

There remains room for improvement – in getting different researchers to co-operate and follow compatible protocols when investigating similar research questions, in paying better attention to the time course of results, in the validation of measurement procedures and in the further development of analytical methods for bringing together results from multiple trials. Senn (2000) has wider relevance than drug trials.

The Consequences of Methodological Defects

Commonly shortcuts are taken. Schulz et al.(1995) took trials that had been examined in 33 meta-analyses, classified them according to methodological quality, and used logistic regression to estimate the bias that resulted from each methodological defect. The
methodological defects examined were inadequate allocation concealment, exclusions after randomisation, and lack of double-blinding. As an indication of the split between categories,
steps taken to conceal treatment allocation schedules were adequate in 79 trials, unclear in 150, and inadequate in 21. By comparison with trials where steps taken to conceal treatment allocation schedules were adequate showed a reduction in the treatment effect odds ratios
reduced by a factor of 0.67 (95% CI 0.60 - 0.75) for trials where the schedule was unclear, and by a factor of 0.59 (95% CI 0.48 - 0.73) in trials where the schedule was inadequate.

For trials that are not double-blinded the estimated reduction in odds ratio is a factor of 0.83 (95 % CI 0.71 - 0.96). Overall, the evidence is that inadequacies in procedures may, as other authors have claimed, lead to serious overestimates of treatment effects. Crude attempts at meta-analysis that do not consider trial quality may overestimate treatment effects.

I am not aware of studies of the agricultural literature that attempt to estimate bias as a function of methodological adequacy.

Drug Trials

Many different designs are used in drug trials. One distinction is between parallel designs and changeover designs. In parallel designs each patient receives one treatment only. In a changeover design each patient receives two or more of the treatments in turn, and the response is recorded following each treatment. Thus a changeover design allows comparisons of the results of different treatments on the same patients.

Here is a three-treatment parallel design, with four patients per treatment. Each cell in the table is a different patient.

Treatment A	Treatment B	Treatment C
A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4

Two possible changeover designs are

1. Each patient receives the three treatments A, B, C. The order is randomised.

2. Each patient receives the four treatments A, B, C, D. Now however we note that the treatments may appear in each of the orders ABC, ACB, BAC, BCA, CAB, CBA. We randomly assign two patients to each of these orders of treatment.

The changeover design may look good, until we start to worry about carry-over effects. With 12 patients, either of the above designs gives 36 results, instead of the 12 that we get from the parallel design. In addition we can compare the effects of each pair of drugs by comparing effects on the same patients. So comparisons should be precise. Unfortunately however, we do have to worry about carry-over effects. The second design makes it possible, under certain assumptions, both to compare drugs and to estimate any effect from order. We lose degrees of freedom in order to estimate order effects, losing some of the advantage we had gained by testing different drugs on the same patient.

For further discussion see Senn (1998).

Ethical Issues in Experiments with Human Subjects

It is unethical to conduct a trial unless the outcome is in doubt. If there is doubt, the proper ethical action is to initiate a clinical trial. Consider a life-threatening disease where some specialists give one treatment and some another, with no convincing evidence to support the choice. If one is superior to another, the patients of half the specialists will get the inferior treatment. For convenience, assume this is about half of the patients. The ethical approach is, for the duration of a clinical trial, to make the assignment randomly. Half of the patients will still be disadvantaged, but only until such time as the trial shows a clear difference.

There are strict legal requirements for trials with human (and also for animal) subjects. Participants must give informed consent. Detailed requirements are set out in Therapeutic Goods Administration (1991) and in ICH (1996). There are potential ethical issues that go beyond the considerations in these documents. Poor trial design or analysis inadequacies may vitiate results, depriving future patients of the use they might have received from results from a well-conducted trial.

12.4 Allocation of Resources

Here I return to the questions I asked at the beginning.

1. We have an experiment where a tray of fruit is the experimental unit. What is the best way to increase precision – to increase the number of trays, or to increase the number of fruit?
2. In a randomised trial for comparing two treatments for arthritis, with grip strength as the main outcome measure, what is the best way to increase precision – to increase the number of patients, or the number of observations per patient?
3. In a randomised controlled trial for comparing two treatments for epileptic fits, with number of fits per two-week period as the main outcome measure, what is the best way to increase precision? Do we need more patients, or more observations on the patients that we have?

I will use the third example (from Thall & Vail 1990) for illustration. Data are from a randomised controlled trial intended to test the usefulness of progabide in controlling epileptic fits. For each of 28 patients in the placebo group and 31 patients in the progabide group, there are four sets of differences from baseline, one for each of four two-weekly periods[2]. Here is the analysis of variance table:

	Sum of squares	d.f.	Mean square
Difference between treatments	6.4	1	6.38
Between patients	162.0	57	2.84
Within patients	117.8	177	0.665

If there were no between patient component of variation, additional to what can be explained by differences at different two-week periods for the same patient, then the *between patients* mean square and the *within patients* mean square would be, to within statistical variation, equal. The difference between 2.84 and 0.665, i.e. 2.84-0.665 = 2.175, is the part that is explained by the between patient component of variation. The between patient mean square is thus in two parts

1. Due to the between patient component of variance = 2.175
   Then as there are four values for each patient, the between patient component of variance is $s_b^2 =$ 2.175/4
2. Due to the within patient component of variance = s^2 = 0.665

The variance of the mean for one patient is then $s_b^2 + s^2/m$ = 0.665 + 0.544/*m*

The second part, i.e. 0.544/*m*, already with *m* = 4 contributes a relatively small part of the total, variance. There is very limited scope for reducing $s_b^2 + s^2/m$ by increasing *m*. The variance of the mean of *n* patients is $(s_b^2 + s^2/m)/n$. In order to reduce it substantially, we need to increase *n*.

We will meet the same formula in the discussion of sample size calculations for cluster samples. There *m* is the cluster size. The quantity that we have called s^2 will be written s_w^2. (Here *w* = within).

The take-away message is that we can always reduce the standard error of the mean (for the placebo group or for the treatment group) by increasing the number of patients. The variance is inversely proportional to *n*, with the standard error inversely proportional to $\sqrt{n}$. There are severe limits to our ability to reduce this variance by increasing the number of observations per patient.

Note: We have ignored the complication that the measurements are made in four successive two-weekly periods. There is likely to be a relatively stronger correlation between successive two-week periods (between 1&2, 2&3, and 3&4), a weaker correlation between 1&3 and between 2&4, and a weaker correlation still between

[2] The baseline was established by following all patients, initially, for an eight-week pre-treatment period. The data are differences, on a square root scale, from (baseline count)/4.

1&4. The conclusions above are affected only to the extent that increasing *m*, while decreasing the variance, will not decrease it proportionately to *m*.

*12.5 Two Ways to Compare Instruments for Measuring Fruit Firmness

Here we extend the paired comparison idea to >2 treatments. There are two ways one could do this. There could be more than two treatments per block. Or one can go to an experiment where there is one block for every pair of treatments. Again, I want to start by talking about a less precise complete randomised design experiment that we could have, but did not, do.

The experiment I now describe was designed to compare instruments, known as penetrometers, for measuring fruit firmness. My description, and the data that I will present, come from the larger experiment described in Harker et al. (1996). Dr. Harker had charge of the experiment. I helped design it, and did the analysis.

A dial records the pressure that is needed for the probe, fixed via a spring to the handle, to penetrate the fruit. The aim of the experiment was to compare four different designs of penetrometer, plus another piece of equipment designed on a different principle that was known as a twist tester. I will describe and compare two experiments – the experiment we in fact did, and a different simpler experiment that we could have done. Just for simplicity, let's assume that there were four machines to compare. In the experiment we actually did, we compared either 10 or 13 device-operator combinations.

A Complete Randomised Design

Here is an experiment that we might have done (Fig. 17):

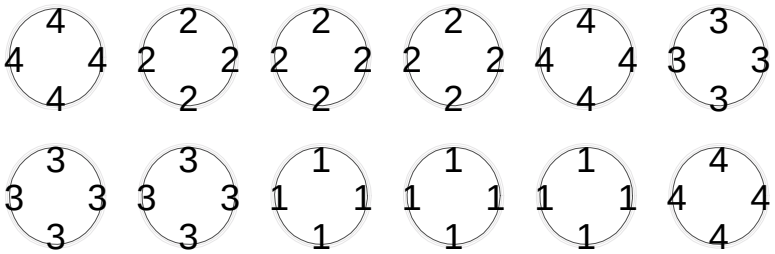

Fig. 17: A Complete Randomised Design for comparing four instruments for measuring fruit firmness. In order to improve accuracy for the result for each fruit we make four measurements per fruit.

The experimental procedure is to take 12 fruit, and divide them up randomly into four sets of three. The first set is tested with penetrometer 1, the second with penetrometer 2, and so on. In the technical jargon, this is a completely randomised design. I've made four measurements for each treatment unit. We could get one average for each fruit, making the data very easy to analyse.

The All Possible Pairs Experiment

Here (Fig. 18) is the "all possible pairs" experiment that we in fact did, though we had 9 (in 1991) or 13 (in 1992) device-operator combinations, rather than just 4 devices:

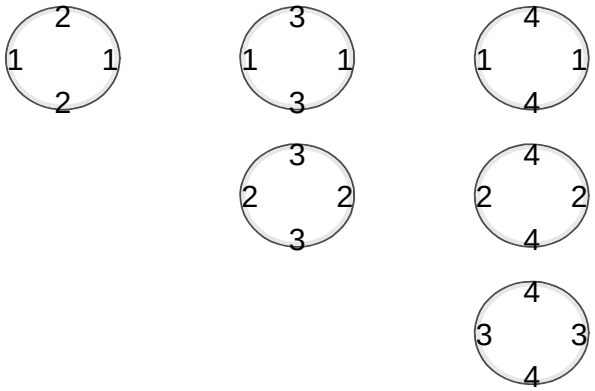

Fig. 18 : A design for an experiment to compare four instruments for measuring fruit firmness. The design makes each of the six possible pairs of comparisons: 1 versus 2, 1 versus 3, and so on. Each measurement was repeated twice on each fruit, allowing an assessment of the variability for one instrument on the same fruit.

Each of the measurements was taken twice. This was not strictly necessary. The "all possible pairs" experiment has a built-in redundancy that allows us to assess repeatability. In addition to the direct comparison between 1 and 2, there are the comparisons :

1 versus 3 and 3 versus 2 (Then $y_1 - y_2 = y_1 - y_3 + y_3 - y_2$.)
1 versus 4 and 4 versus 2
1 versus 3, 3 versus 4, and 4 versus 2
1 versus 4, 4 versus 3, and 3 versus 2

The details are not important. What is important is that are several different ways in which we can get a comparison between machine 1 and machine 2. This built-in redundancy would have allowed us to get an estimate of the standard error of treatment effects, even if we'd done just one repeat of the total experiment.

So both experiments satisfy the requirement of demonstrable repeatability, at least to the extent that we could assess the consistency of results when the experiment was repeated by the same group of scientists on the same batch of fruit on the same day.

Fig. 19 shows one set of results:

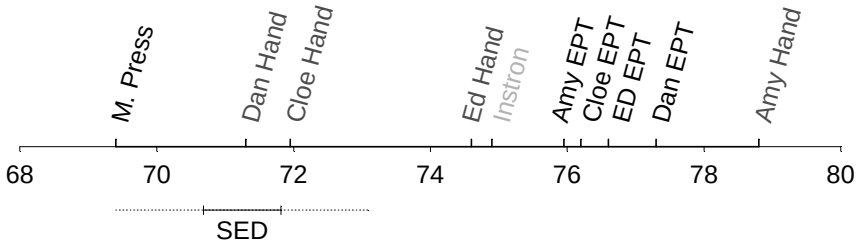

Fig. 19: Results from an experiment that compared device-operator combinations for measuring fruit firmness. Results were for firmness of apples at harvest, 1991.

This demonstrated what was already known, that one can get a huge variation between hand operators. Operators are, though, relatively consistent within themselves. The EPT pressure tester, which is also hand operated, seems less affected

by the operator, at least for the relatively firm apples. We were however rather more interested in the variability. We'll look at that shortly.

Accuracy

I will make a comparison with the other type of experiment, the experiment we did not do. If we'd used a complete randomised design, it would have taken 3.7 times as many fruit to get the same precision, for comparing means of device-operator combinations. People would have been punching away at penetrometers almost four times as long. Here are details:

		Number of Fruit	
		All possible pairs design	Complete Randomised Design
Kiwifruit 1991	Harvest	90	90×9.3
	Storage	90	90×7.3
Apples 1991	Harvest	90	90×1.5
	Storage	90	90×3.7
Kiwifruit 1992	Harvest	156	156×5.6
	Storage	156	156×1.7
Apples 1992	Harvest	156	156×1.1
	Storage	156	156×2.6

Table 1: Comparison of number of fruit required in "all possible pairs" design with number required in a (less efficient) complete randomised design.

Observe that when fruit are firmer, within fruit variation is relatively more similar to between fruit variation. So the gain from the "all possible pairs" design is less marked.

Which instrument was the most accurate?

	Harvest		Storage	
	1991	1992	1991	1992
Instron	.11	.17 (.10)	.18 (.10)	.10
M. Press	.13 (.10)	–	.072	–
Hand	.082	.10	.071	.082 (.062)
EPT	.085	–	.080	–

			(.07)	
Twist	–	.073	–	.092

Table 2: The above figures are standard deviation estimates, when there are repeated within fruit measurements using that device. The bracketed figures are robust estimates of standard deviation. These are given only if the robust estimate is more than 10% less than the crude estimate.

Points to emerge are that the hand machines do well, providing one stays with a single operator, and that the Instron seems prone to occasional aberrant readings.

What did we learn?

Experimentation is guided learning. What did we learn?

One important point was that we'd not thought too carefully about the way that firmness may vary around the equator of the apple or kiwifruit. The indications from our results are that such variation is inconsequential, and certainly much smaller than fruit to fruit variation. However I'm sure that some of the variation is systematic within fruit variation. If we knew more about it, we could come up with a more precise design. Someone (Roger Harker or someone else) may in future do such an experiment. For the present, much more interesting experiments are occupying his attention. Thus he has been fixing small radio receivers inside teeth, to record the noises that people make when they bite.

These results can help illuminate what other researchers have found or failed to find. Thus Bongers (1992) found no operator differences. But

- Bongers used soft apples (50 - 58N)
- Bongers used an imprecise design.

There are many reasons why researchers may fail to find an effect! Be careful how you interpret negative results. Two other older papers did find operator differences.

12.6 Multiple Factors

In the above, we have focused on the kinds of complication that arise when we introduce multiple levels of variation, albeit in a fairly simple way. We may also have multiple factors. In fact, experiments that examine all combinations (or perhaps a subset of combination) of levels of several factors in a systematic way have huge advantages over experiments that vary levels of one factor at a time. Replication can be reduced to a minimum (e.g. two replications of the total experiment), or may not even be necessary, when there are three or more factors. Technically, what happens is that high-order interactions, i.e. highly complicated forms of interaction between factors that are assumed unlikely, are used to estimate the random error. This is a conservative procedure; it will tend to over-estimate the error variance.

Hidden Replication in Multi-Factor Experiments

Replication is not an end in itself. It is designed to ensure that there are enough degrees of freedom to estimate variances that are relevant to the calculation of standard errors of parameter estimates and of fitted values. Especially in multi-factor experiments, it is often possible to get the needed degrees of freedom from within the experimental design, without

formal replication of the total design. The degrees of freedom that one needs are available from within a design in which each combination of factor levels may occur once only. It may be helpful to speak of this as `hidden replication'.
In multi-factor experiments that have many factors, some factors, and/or their interactions, will have little or no effect. Suppose that each combination of factor levels has a separate experimental setup. Then degrees of freedom that correspond to factor interactions and main effects that appear negligible are available for the estimation of error. There is a risk that the error estimate will be biased upwards because these factors or interactions really did have some small effect. Often, the benefits outweigh any actual or potential loss of precision.

Thus consider an experiment where a fixed amount of water is put in a jar in a microwave oven and heated. The main aim is to check out the effect of position on the turntable – near the centre or on the edge. Rather than just repeating it, you do it once in a red plastic container and once in a yellow plastic container. If as you expect the different colour of the containers makes no difference, this is just as good as repeating it with the same colour of container. At the same time you have checked out that the colour of the plastic really does make no difference. Perhaps there was a sneaking suspicion that the two plastics were slightly different materials.

Or consider heating water in an electric jug for varying times. The increase in temperature is measured. The results are set out in a table:

Time (sec.) . .
Temperature change (°C) . . .

Over a small range of times, one expects the temperature increase to be proportional to the time in the oven, resulting in a straight line relationship. So the different points on the line can all be seen as checking out the relationship. Having the different points is as good as having replicates all at the one temperature, providing the points are obtained quite independently of one another. The pattern must be regular enough that departures from it stand out with reasonable clarity.

Note however that the measurement at different times of effects that develop or evolve over time (e.g. growth) does not give hidden replication. The points are not independent. The present measurement depends to an extent on the measurement at the previous time-point.

Question: Which of the following have satisfactory hidden replication?

1. Apples, all of similar firmness, are dropped from ever greater heights and the extent of bruising noted.
   [Question: Suppose two assessments, by two different technicians, are available for each apple. Does that have anything to do with replication?]
2. Asparagus plants are set out in plots 1, 2, ..., side by side along a row, with one plot for each different level of fertilizer. There is a random allocation of fertilizer level to plot.
3. Asparagus plants are set out in plots 1, 2, ..., side by side along a row. Plot 1 gets the highest level of fertiliser, plot 2 gets the next highest level, and so on.
4. The design is as in 3., but this is repeated for several rows. A coin is tossed to decide the end of each row at which to start with the highest level of fertilizer.
   [Should there be an equal number of rows in each direction?]
5. Two graphs, one for each variety of apple, show how a biochemical measurement changes as the season progresses. The aim is to compare the patterns of change for the two varieties. Each result, for each variety at each time, is a single bulked measurement from 10 apples.
6. In a spacing trial, apple trees are arranged in concentric circles, in such a way that the inter-tree spacing increases as one moves out from the centre.
7. The layout 6 is repeated several times.
   [Is this really necessary?]

Large is good. Is larger better?

One should not make experiments too large. Very large experiments bring a seriously increased risk that the experiment will not go according to plan. On balance, unless it takes a long time to give results, it is usually best to spread resources over several experiments. The experimenter may learn something from the initial experiments that leads to carrying out quite different subsequent experiments.

In general reduction in the number of replicates, providing it can be achieved while still retaining enough degrees of freedom to estimate error, is preferable to reducing the number of factors. Unless the experimental methodology is well established, or cheap to repeat, it is well to do a pilot experiment first. This is particularly important for large experiments. The initial experiment may aim to identify major effects only. Various elaborations on 2^n factorial experiments are available for this purpose.

The initial experiment may aim to identify major effects only. It may try to narrow down the range of factors to be investigated. Various elaborations on 2^n factorial experiments are popular for this purpose.

*12.7 Fractional factorial designs

Many of the designs used in industry, and suitable also for laboratory use, are an adaptation of the 2^n factorial design. These designs are useful for exploration, for determining which factors should be examined further. They are not usually intended to provide final answers.

Consider for example a 2^4 design. There are four factors, each at two levels. For example, you are studying the heating of water in a microwave oven. The four factors are:

1. Location on turntable: Centre / Outside
2. Is container covered: Yes / No
3. Nature of material of container: Plastic / Glass
4. Shape of container: Tall and narrow / Low and squat

There are $2 \times 2 \times 2 \times 2 = 16$ combinations, and it would be quite reasonable to look at all 16 combinations of factor levels.

With factorial experiments where the number of factors is large (say > 5), replication happens without planning for it. It is likely that one or more of the factors will have negligible effect. Results from repeating the experiment over levels of that factor give what are, in effect,
replicate results. Or if all main effects are substantial, certain of the interaction effects will be so small that they can be neglected. This makes it possible, by mathematical juggling, to get `effective replication'. Factorial experiments have surprising bonuses.

Often 2^n experiments are used for initial `look see' purposes. The aim is to pick out the one or two factors, or perhaps combinations of two factors, that have the major effect. Even with just three or four factors, it may not be necessary to replicate – second and/or third order interactions can be used to estimate “error”. The aim will be to determine which factor effects and interactions have a substantial effect and should be investigated first.

In fact for an initial `look see' we might decide, in the above experiment in heating water in a microwave oven, that we'd like to get away with fewer units. We'd like to know which

factors should be investigated further. It is possible to do an experiment, using just half of the full sixteen units, that will provide limited information. Such an experiment is known as a half factorial of a 2^4. The trick lies in knowing which eight, out of the total of 16 combinations of factor levels, should be chosen.

There are various elaborations of 2^n designs that add a small number of further design points. An additional central point, with each factor midway between its low and high levels, is often used.

*12.8 Further Issues in the Design and Analysis of Experiments

Response surface designs

Where two or more factors are quantitative, results are appropriately presented as a response surface. Special design considerations arise, on which there is a large literature.

The analysis should try to identify such features of the shape of the response surface as the data allow. Often it is possible to make only gross distinctions, to determine whether it is shaped like a plane, like a saucer, like a hill, like a saddle, or like a hill with a plateau.

If the fitted surface is three-dimensional and provides a good fit to the shape of the response surface consider presenting it as a contour plot. Contour plots give level contours on the surface. Problems of perspective make it difficult or impossible to read precise information from wireframe diagrams. Contour plots may be a better alternative.

Researchers will often use significance tests to compare results from design points, two at a time. This makes for numerous tests, and loses the power that response surface analysis offers to detect effects. It also treats the design points as though they are somehow special. Usually they have been chosen only to give a reasonable spread over the ranges of factor levels that are of interest. The experiment was looking for the total pattern of response, and chose these particular design points as narrow windows into a view of the response surface. The aim of the analysis should be to enlarge these windows as much as possible, to get a picture of the whole surface.

Analysis of Unbalanced Experimental Data

Often the unbalance results from a small number of missing values. One then inputs these as missing values and the analysis proceeds as normal. Estimates are unbiased, but analysis of variance residual mean squares are biased downwards.

REML (Residual Maximum Likelihood) provides an output similar to analysis of variance when the unbalance is too great for the satisfactory use of an analysis of variance with missing values.

Where the unbalance is severe, Quantile-Quantile plots provide a means for identifying data points that stand out as different from the rest.

Quantile-Quantile Plots

Normal probability and other quantile-quantile (Q-Q) plots check whether data follow some assumed distribution. Normal probability plots check whether the data follow a normal distribution. They are often used to examine residuals from regression or from analysis of variance. The data values, or the values of the residuals are taken in order. Assume there are n values. Along the x-axis one marks the normal deviates that correspond to cumulative probabilities $(i\text{-}0.375)/(n+0.25)$, $i = 1, 2 \ldots, n$. The

lowest data value is plotted against the first of these points, the next lowest against the next point, and so on. If the data really are from a Normal distribution then, aside from random sampling variation, the points lie close to a straight line. The effect of random sampling variation is a difficulty in the use of these plots when the number *n* of sample points is small. One needs to calibrate the eye by examining a number of plots of data that have been generated to be random normal. Fig. 19 shows two probability plots, one for random normal data and the other for data in which one value is clearly an outlier.

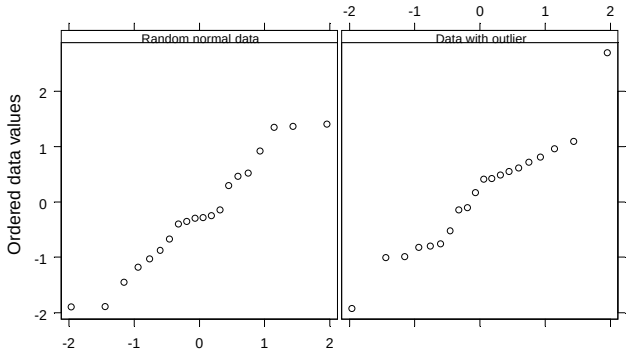

Fig. 20: Normal probability plots: for random normal data, and for data with outlier.

Here are a couple of specific possibilities for the appearance that a normal probability plot may present:

1. The points lie on a curve rather than on a line. This is what one expects if the distribution is e.g. lognormal or gamma rather than Normal.
2. The points fall on two or more distinct lines or curves. This is what one expects if the data is from two or more distributions, e.g. two Normal distributions with different means and/or with different variances.

Where the data is from two Normal distributions with quite different means one expects two well separated parallel lines, one for a set of ``low" data values, and one for a set of ``high" data values.

Where one distribution has both a larger mean and a larger standard deviation than the other, the points tend to separate out onto two lines. The second of the lines will have a higher slope than the first. If most points are from one distribution, with just a small number of points from another distribution, they may show up as `outliers'.

Datasets of perhaps six or nine points allow only very limited inferences, of the type just noted. The patterns must be very clear if they are to stand out above against random variation.

The best way to get an idea of the effect of sampling variation on a Q-Q plot is to take repeated samples of the requisite size from a Normal distribution, and examine the Q-Q plots. Examination of repeated plots for random normal samples helps calibrate the eye! In small samples, the plot may be quite irregular. In large samples, it will be close to a line.

Specific forms of Spatial Dependence

If spatial or temporal variation can be modelled, this has implications both for design and analysis. Row and column designs are a generalization of Latin squares that are appropriate when plot effects can be expressed as the sum of a row and of a column effect. Row and column designs aim to allow for field fertility gradients in both the row and column directions. See Williams & Matheson (1997) and John & Williams (1995). Software is available that will generate efficient row and column designs. More generally, in experiments that have a spatial layout (e.g. in field or storage chamber or sensory experiments), plots that are close together may be more similar than those that are widely separated in ways that can be modelled. Here the design issues seem not to have been much explored.

12.9 Design Questions – Examples to Ponder

1. A commercial supplier of processed chicken intends to change its order forms, to make them more customer friendly. It has three alternative new form designs to test. Design a comparative trial, which can be implemented with minimum interference with current procedures, for deciding between the three forms. What would you use as outcome variables?

2. Several of your customers have complained about the form and have offered help from their staff in improving it. One of them suggests an experiment with a block design, i.e. each person who orders chicken tries out all three forms. Can you make this work? They do not, of course, want to fill out three different forms for the same order. In any case, this would probably not work, as it might be much easier to fill in whichever form were filled out after the first.

3. A commercial firm, with a number of branches nationwide, is considering sending regular newsletters to customers, believing that this will lead to increased sales and greater customer loyalty. However they want to be sure that the newsletters do serve their intended purpose, and are prepared to undertake an experiment where the initial mailings go only to a subset of customers. The firm is willing to wait for up to a year for results. Customers in the same city might become aware of any different treatment for other customers in the same city, and this might limit your choice of design. What design would you suggest?

4. An advertising firm wishes to determine which of two TV advertisements is more effective. Suggest a design for an experiment.

References and Further Reading

Experimental Design

Box, G.E.P., Hunter, W.G., and Hunter, J.S. 1978. Statistics for Experimenters. Wiley, New York.

Cox, D.R. 1958. Planning of Experiments. Wiley, New York.

John, J.A. and Williams, E.R., 2nd. ed. 1995. Cyclic and Computer Generated Designs. Chapman & Hall, London.

Williams, E. R. and Matheson, A. C., 2nd edn. 1997. Experimental Design and Analysis for Use in Tree Improvement. CSIRO, Melbourne.

Experiments Comparing Instruments for Measuring Fruit Firmness

Blanpied et al. 1978. A standardised method for collecting apple pressure test data. New York's Food & Life Sciences Bulletin 74:1-8.

Bongers 1992. Comparison of three penetrometers used to evaluate apple firmness. Washington State Tree Postharvest Journal 3: 7-9.

Harker, F. R., Maindonald, J. H. & Jackson, P. J. 1996. Penetrometer measurements of apple and kiwifruit texture: operator and instrument differences. Journal of the American Society for Horticultural Science 121: 927-936.

Voisey 1977. Examination of operational aspects of fruit pressure tests. Canadian Institute of Food Science & Technology 10: 284-294.

Planning of Research

Beveridge 3rd edn 1957. The Art of Scientific Investigation. William Heinemann Ltd., London.

Sagan, C. 1997. The Demon-Haunted World. Science as a Candle in the Dark. Headline Book Publishing, London.

Data from Progabide (for Epilepsy) Trial

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data. Biometrics 46: 657-671

Randomised Controlled Trials

Amberson, A. B. Jr, McMahon, B. T., and Pinner, M. 1931. A clinical trial of sanocrysin in pulmonary tuberculosis. American Review of Tuberculosis 24: 401-435.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R.,Rennie, D., Schulz, K. F., Simel, D., and Stroup, D. F. 1996. Improving the Quality of Reporting of Randomised Controlled Trials: the CONSORT Statement. Journal of the American Medical Association 276: 637 - 639. [See also the web page http://www.ama-assn.org/public/journals/jama/jlist.htm]

ICH 1996. Guideline for Good Clinical Practice. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Available from http://www.pharmweb.net/pwmirror/pw9/ifpma/ich1.html

ICH 1998. Statistical Principles for Clinical Trials. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Available from http://www.pharmweb.net/pwmirror/pw9/ifpma/ich1.html

Dickmann, W.J., Davis, M.E. et al. 1953. Does the administration of diethylstilbesterol during pregnancy have therapeutic value. American Journal of Obstetrics and Gynecology 66: 1062-1081.

Piantadosi, Steven. 1997. Clinical Trials: A Methodologic Perspective. Wiley 1997.

Diehl, H. S., Baker, A. B. and Cowan, D. W. 1938. Cold Vaccines: an evaluation of a controlled study. Journal of the American Medical Association 111: 1168-1173.

Fleiss, J. L. 1986. The Design and Analysis of Clinical Experiments. Wiley, New York.

Gehan, E. A. and Lemak, N. A. 1994. Statistics in Medical Research. Plenum Medical Book Company, New York.

JAMA 1997. Clinical trial investigators talk about getting data. Journal of the American Medical Association 277: pp.1833-1836.

Louis, P. C. A. 1835. Recherches sur les effets de la saignée dans quelques maladies inflammatoires; et sur l'action de l'emetique et des vesicatoires dans la pneumomonie. J. B. Ballière, Paris. Translated by C. G. Putnam. Hillary Gray, Boston, 1836.

Meinert, E. L. & Tonascia, S. 1986. Clinical Trials. Design, Conduct and Analysis. Oxford University Press, New York.

Senn, S. 2000. Consensus and controversy in pharmaceutical statistics. The Statistician 49, Part II: 135-176.

Shapiro, S. H. and Louis, T. A. 1983. Clinical Trials. Issues and Approaches. Marcel Dekker 1983.

Therapeutic Goods Association 1991. Guidelines for Good Clinical Research Practice (GRCP) in Australia.