On the Relevance of Experimental Design and Statistical Analysis, in our Work

Nathaniel

April 2007

Purpose

When I started working on this DS and PD subject it was a return to some laboratory work after a lot of non-experimentally based work.
Some of the things with which we work are quite simple in this way: easily verifiable assumptions of linearity, together with no or easily described dynamic state, make for a system where a small number of measurements can define the system sufficiently for our needs.

Other things that we work with are very different: reversible and irreversible aging of specimens, large variation in supposedly similar specimens, high non-linearity and statefulness, many plausible significant parameters, and several different effects captured in one measurement, combine to make it very hard to establish within the constraints of available time and specimens a reliable model of the significant parameters, the relation of controlled and observed variables, and the variance in the specimens or measurements. Particularly in looking to model the observed phenomena in terms of component physical mechanisms, a good idea of this elusive model is important.

A good book [1] that I read just a couple of years ago started me thinking about the usefulness of randomising the order of measurement sequences. This was particularly relevant to some measurements where I was feeling a little worried about the fine interpretation of a trend of DS results at low frequency and varied voltage; the voltage level had been progressively raised during the measurement. I had already considered reversing the time-order of measurements to check whether some other parameter having an effect. It was amusing to me how the very idea of ever conducting an experiment with so obvious a potential for erroneous conclusions, was looked on as obviously bad technique — any other factor that changes with time, such as aging, could lead to the belief that the intended variable of voltage was having the effect when it wasn't.

Another point mentioned in that book was the use of 'factorial design' and the advantages of an experiment where one varies several factors. More recently, the possible usefulness of factorial design of experiments, perhaps also with Analysis of Variations of the results, struck me as something worthwhile considering in the investigation of PD. I have therefore been doing some background reading in the last month. Some of the main points are

mentioned in the following sections.

The books [1] and [2] as well as a few works by Richard Feynman[1] give plenty of interesting thought about the aims and effective methods of coming to know something about nature. I have indulged in all of these, and will sometime leave the two cited books at work for anyone else who's interested to read them. [3] is available on the web as a rather basic (and early) introduction to randomisation and estimation of variance.

The remainder of this 'article' is a rushed and not very detailed account some points from the references and from my thoughts from a little lab work with PD, finishing with thoughts on how we might improve our methods. It is particularly relevant to PD since there lies the greatest range of random variation, poor repeatability, and need for good design to minimise time. Any accusations of poor methods should be taken as levelled mainly against myself ...

Summary of main points

This is the section for anyone who has got this far but isn't going to read all the other pages. The first few points are a little trite, but I certainly can need to remember them sometimes. The other points are the main ones from the rest of the article.

> Keep the overall aim in sight, and regularly review whether what is being done to achieve this still seems the best way. Obvious, but easily forgotten!
>
> Check on existing knowledge in the field, that might suggest good ranges for investigation or previously unconsidered problems to be overcome. Whether such prior knowledge is more help in improving design and saving time, than it is a hindrance in providing experimenter bias and discouraging exploration, is of course case-dependent: I suspect it's often helpful in our field to think a little first, freshly, but soon turn to some references to get other ideas.
>
> Look into whether a similar type of experiment is well known in another discipline, along with a large body of knowledge about good ways of performing the experiment and analysing its results.
>
> Consider the significance of results. How likely is it that a result that is taken to confirm the truth of an hypothesis could have come about by chance? How likely is it that the hypothesis was right but was rejected by bad luck in the experiment's results? What has been done in the experiment to *allow* these assessments of probability to be made (randomising)?

[1] Feynman's rather well-known lecture about pseudo-science and scientific integrity is available at (among other sources) `http://wwwcdf.pd.infn.it/~loreti/science.html` — I find it amusing how someone so brilliant and clear-thinking can use silly arguments such as that (in this case) the decrease in ability at schools demonstrates that modern educational theory is not working! Still, it's a good read, and it is the origin of the much used and highly amusing term 'cargo-cult science'. Other works such as *The Character of Physical Law* are worth reading.

Rather than going quickly into long sessions of experiment (or simulation), consider carefully the design of the experiment, in order to give as efficiently as possible the information about variance in results, effects of various factors, and possibly higher-order effects between different factors.

Deciding criteria *in advance* for accepting or rejecting an hypothesis, is no bad idea: it is all too easy to invent excuses after the fact, to explain those tedious deviations from the neat, theory-saving expected result. Any already verified problems in apparatus should be taken into account in the criteria; any apparent problems that the results seem to indicate, should themselves be confirmed, by experiment designed for the purpose, rather than just being used to brush away an undesirable result.

Take advantage, where relevant (several areas in my case), of the idea of a 'factorial design'. This can include variation even of factors not expected to have influence, or expected to have a known influence that can be compensated: such assumptions are good to confirm, and if they are confirmed as correct then a better estimate of variance is achieved by the extra measurements. It may also or instead omit random combinations of factors, in order to save time at the cost of less well-defined results.

Avoid the very easy temptation to avoid digging deeper when it begins to look as if all those beautiful curves were due only to the measuring equipment warming up, or the sample degrading over time, or something else other than the investigated factor: take advantage of randomisation of the order of measurement points in order to give a known, low probability of trends being seen due to other factors. Remember that knowledge of how things really are is (we hope) the key aim, even if it means more work and an acceptance that the world isn't as simple as might seem nice.

It is tempting to make statements about 'not much more', 'not significantly', 'small' etc. changes. It is better to start thinking more about the actual ranges of values we believe we mean, and why we believe that, then to make a more specific (and therefore more easily falsified, hence the desire to avoid it!) claim that actually means something. But doubtless I do this sort of thing all the time.

Comparison of our work and other fields

There are many other areas of experimentation that are more vague than ours, on account of a subject area in which the conditions cannot be controlled or the dependent variables measured as easily as they usually can for us. Think for example of the eternal discussion of how bad in the long term it might or might not be for us to eat salt, or whether commonly experienced leveles of power-frequency EM fields increase the risk of cancer, or almost any matter in educational psychology!

In these subjects, there is a much greater emphasis placed on statistical theory and formal methods for setting up a hypothesis and testing it. All this boring stuff from school-days, about Chi-squared, Poisson (not a PDE), t-distributions, F-distributions,

null-hypotheses, etc. comes into play, and, being no longer so pointless as it seemed then, gets really interesting! As the experimental results become more shakey (vague — small effects, small samples), so there becomes a greater need of precise, formal analysis of the statistical significance of results, as well as a greater need for care that an expensive experiment is optimised to give as much useful information as possible, and a greater need for avoidance of experimenter bias by clearly established criteria before the results are known.

Of course, there are plenty of cases closer to our interests where these problems exist, such as assessment of failures of engineering components (cables, circuit breakers, etc.) where timescales are too large for realistic testing in the lab, histories aren't well known, specimens are removed from service sometimes before failure, so failure times aren't really known, etc. Even these are quite easy compared to working with long-term effects in people, but here at least the matter of explicit use of statistics is important.

In the physical sciences and engineering, there is often much less explicit consideration given to experimental design and statistical analysis of experimental results. It is not unusual that plenty of specimens are available, variation between them is small, conditions can be tightly controlled, and many quite accurate measurements can be made. With such good conditions, together with some very well verified fundamental principles and assumptions, very accurate descriptions can often be made of physical relations, so that even fractional percent deviations between a model and an experiment must be accounted for. There is not then much reason to worry about how likely, for example, a result was to have come from pure chance. Common sense and some helpful clues from others' example and one's own experience and thoughts, lead the investigator to think about possible external factors, and to try to test for their presence.[2] We can feel lucky that we have in general an area in which experiment can relatively easily lead to knowledge in which one can have a good degree of confidence.

In some of our work – particularly that with PD – there is a good case for more formal design and analysis than I have noticed being used. In any of our work, the well-known problems of experimenter bias and general human desire to avoid upsetting one's pet little theories, is lurking, and should be borne in mind!
When there are weak effects, and considerable sampling noise, there's a considerable chance of a particular result suggestive of a relation between some parameters, being due only to chance, due to the small sample of points taken. This is where the hypothesis-testing language comes in, to require a definition of a null, and one or more alternative, hypotheses, to be tested. As the sampling noise could make the 'type 1' or 'type 2' errors (respectively, a sample from a population that's actually described by the null hypothesis appears to fit an alternative hypothesis, or vice versa) occur, a 'significance level' is chosen, stating the probability of a positive result (alternative hypothesis accepted) not being due purely to sampling noise on a null-conforming population. What to choose for the significance level is obviously crucial, as it is a trade-off between the two types

[2] An interesting, extreme case, often cited, is that of Blondlot who apparently imagined, and published works about, 'N-rays' until a sceptic secretly disabled the 'generator' in order to show that Blondlot still claimed to detect the rays even without the supposed source being present `http://skepdic.com/blondlot.html`

of error, either of which may have considerable cost. It is a little surprising that it so very often seems to be chosen as 90% or 95%. I always found this rather annoyingly arbitrary, surely inappropriate in many cases: I was amused at the comment in [1] that these common levels are often used without thought, and that the level can reasonably be expected to vary greatly depending on the situation.[3] This really gets the important point, that one is always weighing the importance of either sort of possible error, and that different circumstances call for vastly different significance levels.

Variance and Randomisation

Already mentioned in the introduction is the point in [1] of the importance of randomisation for the avoidance of apparent dependencies of an observed variable on a controlled variable, that may simply be due to some extraneous factor, possibly one that is stimulated by the controlled variable, e.g. aging by time for which voltage has been applied.

Assessing the variance between measurements, specimens etc. is important in order to know how reliable one's results from a measurement are as a predictor of a repeated measurement. By randomising, for example, the time-order of measurements, or the combinations of factors used, the probability of a result having arisen by chance can be easily determined.

Factorial Design and Analysis of Variance

The basic description of 'full factorial design' is simple: take N factors (independent variables) to be studied, select two or more different values to use for each, then measure the observed variable or variables with every combination of these factors' values. When several factors and levels are involved, the number of measurements of course gets rather large in this way!

By including repitition of points in a factorial design, the variance in results can be estimated even in the presence of higher order interactions betewen variables.[4] Rather than trying to make a perfect repitition of a point, a further factor can be added, believed to have either no effect or a known and therefore compensatable effect. If there is a significant trend depending on this extra factor, then one has been able to learn that the supposedly unimportant factor probably *is* in fact important; if, however, the assumption of this factor's insignificance is supported by the results, there are extra measurements for determination of the variance.

Going the other way, a fractional factorial design may instead be used, to save time. Here, some points are omitted; every factor is modified, but not with every combination of all

[3] From [1]: "Books on statistics are likely to dismiss this question with the statement that usually 5 per cent or sometimes 1 per cent is taken as the level. A more realistic statement would seem to be that a level somewhere between 40 per cent and 0.0001 per cent will be appropriate for most cases, the exact value being the result of considerable thought on the part of the investigator..."

[4] The higher order interactions are where instead of a pure first-order relation such as 'changing input factor n by this much causes this much change in output factor m, regardless of other factors', the change in output caused by the input varies depending on the value of some other factor(s).

the other factors' values. Higher order effects between variables then cannot always be distinguished from other first order effects (this is called 'confounding'), so the usefulness of fractional factorial methods depends on the truth of the assumption that the only one lot of these effects exists.

The classic analysis of factorial experiments is Analysis of Variations (ANOVA). The ideas are nicely described in [2] and an overview is given in the websites (wikipedia). The mean of all measurements is taken, then means of different rows or columns along which a particular factor is constant are taken and compared for each factor.
Multivariate ANOVA (MANOVA) is for the case where one is interested in more than one measured variable; as well as the features of ANOVA, interactions between measured variables are considered (it is this which makes it more thorough than just ANOVA for each measured variable).

Relevance to our Experimental and Simulation work

General Ideas

The general ideas (not specificially factorial design and its analysis) are good to keep in mind all the time! These are, for example: taking steps to reduce the effect of the experimenter's desire for things to fit the 'hoped for' pattern; esimating variance; considering how much more likely the observed results are under one hypothesis than under another; randomising measurements in order to have a basis for claims of how likely a trend is to have appeared by chance; picking test-specimens to different treatments at random after making them, rather than knowing that 'these ones are the special ones' while making them; and keeping in mind the extent of errors in results, variation in specimens, and required accuracy in the work in which the results are to be used.

The classic process of formally setting hypotheses and significance levels before the experiment is not very important to us; it is more appropriate to weak effects (so weak that we would never bother to investigate them in the time that we would rather spend on the stronger effects) or situations where important decisions, perhaps with different parties having to compromise, have to be made depending on some result. It is good, however, to think a little of what hypotheses are being considered, the risks of either type of error, the experimental accuracy and specimen variation and therefore an approximate idea of what range of results would be acceptable for accepting each hypothesis; a temptation to accept a different range after the results are known should be treated with suspicion! Consideration of significance is particularly relevant when starting a type of work that one hasn't done before, where for example the probability of chance generation of results, due perhaps to much greater variance in test specimens, may be wrongly estimated if not explicitly considered.

Factorial Design

Wondering about the effects of recent excitation on PD measurements made me start to query whether the above mentioned methods developed for designing and analysing

experiments would be of significant use, i.e., worth the time expended.

The problem is such that there are several input variables (time since last excitation, frequency of last excitation, amplitude of last excitation, materials, cavity size, temperature) and several interesting output variables (largest PD charge, mean charge, distribution in phase), and even the direction of a relation is in many cases not at all reliably known in advance. The work is also time-consuming, with long measurement times at low frequency, a need for many cycles to get enough PDs to make a good pattern, a need of repeated measurements to assess random variation, and possible longer-term changes occurring that suggest a need of randomisation of the order of measurements.

My conclusion here is again that anything approaching the rigour with which some other disciplines work on the analysis of factorial experiments would be absurd in our case, but there is a good case for using *some* features of factorial experiment in our PD (and possibly some other) work, and for trying simple ANOVA as an aid to spotting patterns when many factors are involved.

One sort of PD work is the detailed investigation of the particular form of a relation between one or two controlled variables and one or two observed variables; for example, applied V and f and resultant total and peak PD charges. Apart from other necessary matters such as repitition with several specimens, and randomisation, one is likely to want some repitition of measurements even on each specimen, to assess variance in measurement. If there *are* other factors that could be varied, such as the dead-time if this is believed to be too short in all cases to be losing significant PD, it would be good that repeated points should be made with moderately changed dead-time (or other factor).

Another sort of PD invesigation is when there are many factors that *might* more or less influence the result, as in the investigation of sensitivity to settings of the measurement system, or effects of recent excitation of a specimen. A set of measurements that varies each of these to only a few different values, e.g. 2 or 3, together with some basic ANOVA consideration of the results, would likely be a good idea compared to the rather haphazard practice of discovering the importance of one factor at a time while puzzling over strange experimental results. Still, since we are only really going to focus on quite strong effects, and just mention weaker effects,[5] there's not much point taking a lot of time looking at every variation! A quick script to take an input that describes the factors in a set of measurements, and outputs ANOVA results for some statistic (e.g. Q_{PD} or C'') would be a sensible start to speeding up the problem of looking for weak trends in many factors. []

[5]For example, there's not much chance of the PD modelling being at all soon so sophisticated that we are satisfied of excellent modelling of effects of V, f and cavity size, and decide to take the time to perfect the model's description of the effects of the number of cycles of excitation that happened one minute before the measurement!

References

[1] E. Bright Wilson Jr. *An introduction to scientific research.* McGraw-Hill Book Company, Inc., 1952.

[2] R. A. Fisher. *The design of experiments.* Oliver and Boyd, London, Sixth edition, reprinted 1953.

[3] R. Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513, 1926. `http://digital.library.adelaide.edu.au/coll/special//fisher/48.pdf`.

[4] Wikipedia. Factorial experiment, March 2007. `http://en.wikipedia.org/wiki/Factorial_experiment`.

[5] Wikipedia. Anova: Analysis of variations, March 2007. `http://en.wikipedia.org/wiki/ANOVA`.

[6] Wikipedia. Manova: Multivariate analysis of variations, March 2007. `http://en.wikipedia.org/wiki/MANOVA`.

[7] Six Sigma (a manufacturing-based website). Design of experiments, March 2007. `http://www.isixsigma.com/tt/doe/`.