

Some aspects of statistical inference from
non-experimental data
(Notes for AMS Working Party)

A. P. Dawid
Department of Statistical Science
University College London

August 25, 2006

1 Preamble

Most of the traditional tools of statistics (significance testing, confidence intervals, *etc.*) have been developed to aid the interpretation of well-conducted experimental studies, and centre around issues of separating “signal” from “noise” in the light of necessarily finite sample size. However, while these problems obviously do not disappear when we move to non-experimental studies, the main issues facing the interpretation of data from such studies are different, revolving around the potentially disturbing effects of *bias*, rather than noise.

2 Prediction

Any study or data-analysis is useful only insofar as it helps us to make predictions and decisions for the future. It is helpful to distinguish between two possible foci of such a prediction:

Internal validity Does the study support inferences about further individuals similar to those studied?

External validity Are its results relevant in the wider world?

One can have internal validity without external validity, but not *vice versa*. Consequently we start by focusing on the assessment of internal validity — since when that fails we can make no useful predictions at all.

3 Decision problem

For clarity I consider a simple problem, of comparing two treatments, aspirin ($T = 1$) and chalk tablets ($T = 0$) for headaches. We are interested in the response, Y : how long, in minutes, does it take the headache to disappear?

We suppose we have a database S of patients who have been given one of the two treatments, and their responses Y recorded. We may also have measured further covariates — attributes of the patients existing prior to treatment — such as age, sex, medical history, . . .

The data may have been collected under experimental or non-experimental conditions. Our knowledge of the processes that led to these particular patients being included, and to their treatment allocations, may be full, partial, or absent. For concreteness we will use the term “doctor” for the agent responsible for these decisions, although in some contexts this might be the patients themselves, who self-select, or “Nature”, that selects for them.

Now consider a new patient P . Because we are currently only considering “internal validity”, we assume that P has been drawn from the same population, and meets the same inclusion criteria, as the patients in the database. (This state of affairs will frequently be only a “thought experiment”, rather than a realistic situation — but nonetheless helpful to focus understanding). Our decision problem is to determine which treatment to give to P .

Whichever treatment we decide on, we will remain uncertain about the ensuing value of Y . We need to quantify that uncertainty. This will ideally be in the form of a probability distribution (or some appropriate summary, *e.g.* its mean) for Y , for either contemplated treatment. Knowing these distributions is not enough to determine the optimal treatment choice — which in particular will also depend on utilities (valuations) of the various outcomes — but it is as much as we can get from statistical analysis. We would thus like to use our data, if possible, to help us assess these distributions.

4 Allocation bias

Let S_1 denote the subset of patients who received aspirin. It would seem *prima facie* that we might form an estimate of the distribution of Y (or of its mean, or . . .) from the data in S_1 , and regard this as an estimate of the distribution (or . . .) of Y for our new patient P , *if* we were to assign him aspirin. However to justify this we need to argue that P is “like” — the technical terms is “exchangeable with”¹ — the patients in S_1 . This might be a reasonable assumption if *e.g.* the allocation of treatments to the patients in S was randomized, so that S_1 is just a random subset of S .

However in a non-randomized study the set of patients allocated to receive aspirin might well NOT behave like a random subsample of S . For example, the

¹A group of individuals (here, S_1 augmented by P) is exchangeable (relative to a certain state of knowledge) if joint uncertainty about their values is unaffected by shuffling the order in which we take them.

doctor might have assigned aspirin to those patients he thought had stronger constitutions than normal. In that case we could regard P as exchangeable with S_1 only if we have reason to think that P too has a stronger constitution than normal (assuming that this might of itself affect recovery time).

Of course all the above considerations apply equally to the chalk tablets: to justify using the data on S_0 to assess uncertainty about P 's response to chalk, we would now have to argue that he is exchangeable with the patients in S_0 (e.g., like them he has a *weaker* constitution than normal). Only when we can regard P as exchangeable with the patients in BOTH subgroups, S_1 and S_0 , can we use the database to assess uncertainty about the response of P to either treatment, and so make an informed choice between them. But clearly in that case the two groups of patients S_1 and S_0 , each being exchangeable with P , have to be exchangeable with each other: in the database, we must be “comparing like with like”. The above decision-theoretic analysis justifies the need for this (perhaps self-evident) requirement.

However, in anything other than a completely randomized study it can be hard to argue that this exchangeability property holds. We now consider some ways around this difficulty.

5 Pre-stratification

Suppose we think that response to treatment might vary with the sex of the patient. We might then conduct an experiment *stratified by sex* (perhaps in proportions different from those arising naturally in the study population, and possibly different in the two treatment groups). Then we obviously can not just lump together all the patients receiving aspirin, regardless of sex. Here sex is a covariate that we are *obliged* to take into account.

If we consider a new patient P known to be male, we would like to argue that he is exchangeable with the corresponding stratum (*i.e.*, the men) in the study, and thus use the data on men only, under each treatment, to assess P 's distributions for Y , for each treatment.²

Again, to justify using these estimates to help us in treating P , we need to be able to argue that P (known to be male) is simultaneously exchangeable with the men who got aspirin, and exchangeable with the men who got chalk — which requires that these two groups be exchangeable with one another: the property of *conditional (within-stratum) exchangeability*. This could usually be accepted if we have randomized patients to treatments within each stratum, but otherwise it could again be hard to justify. And any attempt at such justification must of necessity involve considerations beyond the internal evidence of the database.

Even when we can assume within-stratum exchangeability, a potential problem arises when we don't know the stratum of the new patient P — unlikely

²Or, if there are too few in the relevant stratum — as frequently happens when stratifying by many factors simultaneously — we can call on a host of statistical models and methods which aim to “borrow strength” from data in other strata to infer appropriate treatment-specific distributions within this one.

when stratification is by sex, but possible if, *e.g.*, we have pre-stratified by some hard-to-measure genetic factor. However if we further know or can estimate the probabilities of P belonging to each stratum, we can use these to construct a suitable mixture of the estimated within-stratum distributions for use in such a case.

6 Post-stratification and other adjustments

Suppose first we have conducted a fully randomized experiment. When we look at the data we see sex-specific differences in response. If these can be believed, then we should again conduct an appropriately stratified analysis, and use that for prediction for a new patient P of known sex. Or we might find that the response varies with some quantitative characteristic, *e.g.* number of cigarettes smoked, and run a regression/analysis of covariance/... on that to construct appropriate prediction formulae to use for P . (However, we couldn't use these if we didn't know P 's smoking status: then the original analysis ignoring smoking would still be appropriate). The reason it is OK to operate in this way in a randomized experiment is that, when exchangeability holds overall, so too will conditional exchangeability (given any pre-existing covariates).

Similarly, if we have conducted a randomized pre-stratified experiment, then we will have conditional exchangeability so long as we include the stratification variables among the conditioning covariates.

There are problems here of statistical significance: the apparent message of the data, whereby we “explain” much of the variation of Y in terms of a whole collection of covariates, may in fact just reflect noise. This is particularly problematic if the covariates identified as making a difference are themselves the result of an intensive “data-mining” exercise. But serious though such problems can be, they can “in principle” be solved by making the experiment large enough.

7 Non-experimental studies

How though are we to proceed when — as is typical in observational studies — we can not fully justify any exchangeability assumptions? It is common in such a case for a large variety of covariates to be measured, in the hope that exchangeability might hold conditionally on all of them. But quantity is no substitute for quality, nor hope for reasoned argument. When might we expect this ploy to succeed?

7.1 Allocation and application

When assessing conditional exchangeability, it is helpful to distinguish between the effects of *treatment allocation* and those of *treatment application*. Even after allocation, we can imagine overriding the doctor, and assigning treatments to the patients different from those he prescribed. Allocation bias is due to the differential effects on response of those pre-existing characteristics of the

patients (*e.g.* how strong is their constitution?) which govern how the doctor allocates patients to treatments. This can lead to systematic differences between allocation groups, even when we override the doctor and give all the patients the same treatment.

We introduce a further (typically multivariate) *allocation variable* Z , defined as some characteristic or set of characteristics of the patient that, for the patients in S , governed treatment allocation. We allow non-deterministic dependence, as *e.g.* when, after taking Z into account, the actual allocation further depends on a randomizing device. In particular, in a fully randomized experiment Z is effectively absent; while in a randomized pre-stratified experiment Z is constant in any stratum.

7.1.1 Measured Z

Suppose first that Z is known and measured. Then we will have exchangeability conditional on Z . Thus we should be OK using (suitable) prediction formulae that include Z as an explanatory variable.

7.1.2 Unmeasured Z

When Z is known but unmeasured we have a real problem. We would find it difficult to argue for exchangeability, whether unconditional or conditional on any measured covariates. This is the situation of “confounding”, and a missing variable such as Z is often termed an “unmeasured confounder” (although if measured it would be an *unconfounder*, not a confounder!).

7.1.3 Unknown Z

In most observational studies we have very little idea as to how the treatments were allocated, so can not even identify Z . One might then conduct sensitivity analyses to see how much difference it makes to vary the specification of Z . One reason for measuring and adjusting for a multitude of covariates is the hope that between them they might determine treatment allocation, and so serve the function of allocation variable. But this needs to be argued independently — it can not be deduced by analysis of the data. And whenever it is simply not the case that the allocation process depended only on measured covariates, no amount of statistical modelling and adjustment can yield valid prediction formulae.

7.1.4 Propensity score

One method of accounting for allocation bias is by use of the *propensity score*. Suppose we know, for each patient in S , the probability that he was going to be assigned to aspirin. Then this “propensity score” PS will serve as allocation variable Z^3 , and we will have conditional exchangeability given PS . We

³It is in fact the minimal such variable.

can thus proceed as above, adjusting appropriately for PS . In particular, because it is a simple 1-dimensional quantity, we can typically form reasonably large (post-)strata by grouping together patients whose propensity scores are all close to some value (“propensity score matching”) and performing within-group comparisons.

The problem is that we do not usually know PS . Methods have been proposed that involve estimating PS by *e.g.* running some form of regression on the data in S , with actual treatment allocation as the dependent variable, and appropriate available covariates as the predictor variables. But even if we ignore the (often substantial) problem that sampling variability may lead to a poor estimate of PS , this again can succeed only when allocation truly is determined (up to irrelevant randomization) by the covariates included. If we can not justify that assumption, we can not rely on propensity matching methods.

Even when we *can* justify propensity matching, it is not clear that its two-stage approach, of first estimating PS and then adjusting for it, is superior to the one-step approach of direct adjustment for the covariates. It is true that the adjustment step, involving only one covariate rather than many, is likely to be more reliable; but the first step, of estimating PS , itself involves processing many covariates, thereby reintroducing the supposedly eliminated sampling uncertainty. In certain specialized contexts, such as individual choice behaviour in Economics, there may be theory (*e.g.* “individuals act so as to maximize their expected utility”) that restricts the form of PS , up to a small number of unknown parameters which might then be well estimated. But this is unlikely to be the case for the medical problems we are concerned with.

8 Other experiments

It is always better to plan an experiment in such a way that we can trust its results, rather than to try and salvage something from a badly conducted study. So well-designed randomized experiments will always be the gold standard. But often there are practical impediments to conducting these. Then we should consider what is the best experiment we can do in the circumstances.

8.1 Instrumental variables

We might wish to control which treatment a patient takes, but be unable to do so. For example, we might recommend a treatment, using proper randomization, but be unable to enforce compliance.

There are two ways we could analyse the data from such an experiment:

Intention to treat Regard the *recommendation* as a treatment in itself, ignoring which tablet the patient actually swallows. If we consider “treating” the new patient P in the same way, and can assume he will exhibit the same kind of (non-)compliance behaviour as the patients in the database, then we can use the data directly to predict the effect of *recommending* aspirin/chalk.

Instrumental variable If we are truly interested in the effect of making the patient take the treatment, we can call on some theory that sometimes allows us to say something about this from such “partial compliance” data. This theory was originally developed by econometricians, under specialized assumptions that will often not apply in medical contexts (although this fact is often unappreciated). There is however some relevant work that weakens those assumptions — although this may only allow us to formulate relatively uninformative inequalities for the desired effects, even when the experiment is huge.

Mendelian randomization is a currently hot topic in this area. Here the “treatment” is possession of some phenotype, and the place of the randomized “treatment recommendation” is taken by possession of some genotype (randomly assigned at birth). This approach would appear to have substantial potential, but again the analysis is predicated on some strong assumptions, which have to be properly understood and justified in any application.

8.2 Internal comparison

Sometimes we can compare two treatments on the same subject. For example, for headaches we might conduct a cross-over trial, in which, say, two distinct headache episodes of the same individual are treated, one with chalk and one with aspirin (the order being randomized). Or we might compare two different eczema treatments, one on each arm of the same individual (the assignment to arm being randomized). There are now new problems that need to be addressed, such as the effects of carry-over and the passage of time in a cross-over trial. But such internal comparisons are typically much more accurate than comparisons across distinct patients.

For present purposes, an important advantage of an internal comparison is that it is hard (though not impossible) for it to be affected by allocation bias in such a way as to seem, falsely, to favour one treatment rather than another.

9 External validity

So far we have only considered conditions that might justify using data to help make a decision for a new patient regarded as essentially similar to those in the database. Henceforth we assume that these are satisfied. But even then it can be hard to argue for the relevance of the data beyond that often artificial context.

So now consider a new patient P in some *different* context. In particular, we do not assume that P is exchangeable with the patients in S .

We might hope that, so long as our analysis is conditional on appropriate covariates, we would still have conditional exchangeability of P with the patients

in the database. If we can argue for that⁴ then, exactly as before, we can use the data to help assess and compare P 's response to each treatment.

In other cases we might not be able or willing to make this argument for the available covariates, but only relative to still further, unobserved, covariates. With additional assumptions we can still make some progress. Thus suppose we concentrate on the expectation (mean) of Y , and assume that, conditional on some *unobserved* covariate U , the expectation of Y , if the patient is given treatment t , has the following *additive* form:⁵

$$E_t(Y | U) = a_t + b(U). \tag{1}$$

It follows that

$$E_1(Y | U) - E_0(Y | U) = \alpha \tag{2}$$

(where $\alpha = a_1 - a_0$): for any individual, his mean for Y when given aspirin exceeds that for when given chalk by the constant amount α , independently of his value for U .

Then, even when P is not exchangeable with the patients in S , his mean difference α will be the same as for them — and so can be estimated from their data.

The above argument can also be conducted after conditioning on further observed covariates (which may be necessary to ensure internal validity).

How are we to argue in favour of (1)? — well, that is another story. . . . But we should not even begin to tell that story until we have already been able to argue for *internal* validity.

10 Unstable treatments

When we use the same mathematical symbol for a quantity, we want it to mean the same thing. So when we have referred to giving treatment $T = 1$, that should mean the same for the patients in S and the new patient P . But there is more to giving/taking a treatment than its biochemical activity. There are also all the surrounding social contexts and emotional pressures that can affect the body's response.

When the study is placebo-controlled and double-blind, we can expect such extraneous factors to be the same for both treatments — they can then be considered as unobserved pre-existing covariates, and so included in U above. So if we can still assume internal validity and (1), we can again estimate the treatment difference α from the data on S , and apply that to the new patient P (assuming that for him, too, the extraneous factors influencing response to treatment are the same for both putative treatments).

However in the absence of blinding, we can not usually make such an argument. If we believe that treatment response may have been affected by *e.g.* the

⁴Note: Even in the internal context, the basis for such an argument had to extend beyond the evidence in the database. And even when we can make such an internal argument, still further arguments will be needed to extend this property to the external context of interest.

⁵We can also consider variations, such as multiplicative models.

doctor's expectations, the best inferences possible are to new situations where the same expectations are attached to the treatments. And this is typically not what we want to know...

11 Issues untouched

These are of course manifold. One important one I am aware of is the whole issue of making causal inferences from case-control and similar studies.

I should appreciate suggestions from the Working Party on other vital statistical issues that I ought to think about.

Acknowledgment

I am grateful to Vanessa Didelez for helpful discussions.

Philip Dawid
August 25, 2006