Notes on Testing Statistical and Causal Theories
JOHN WORRALL

1.Testing Statistical Theories

Not all theoretical claims on which evidence clearly bears are like the deterministic theories we come across in, for example, Newtonian physics (and some, though not all, other branches of physics). Many claims – especially ones of direct everyday relevance – are ***stochastic or statistical*** in nature. You are told that you have a 'better chance' of avoiding heart attacks and strokes if you take regular statins; that you are at 'higher risk' of developing lung cancer if you are a regular smoker of cigarettes; that the 'probability' that there will be major flooding in London has been decreased by X% by the Thames Barrier scheme, etc etc.

Newton's theory is deterministic in the following sense: it tells you that if some system starts in configuration S at some initial time t, then it will (for sure) be in configuration S' at some specified later time t'. If, to take a simple example, a single body of mass 2kg is initially at rest but is then subjected to a constant total force of 6 Newtons, then it will accelerate at a rate of 3 meters per sec per sec, from which we can in turn work out how far from the starting point it will (definitely) be after 1, 2, 3 …seconds. (This is because Newton's second law is of the form, *for any body, anywhere in the universe,* if its mass is m and the total force acting on it is F, then its acceleration will (always) be F/m. The law is usually expressed F = ma.) The outcome is determinate. If there were a single 2 kg body that did not accelerate at 3 m per second per second when subjected to a total force of 6 Newtons, then Newton's theory would be observationally refuted.

Obviously, this is not true of the statistical hypotheses that I listed. We all know of people who have smoked heavily for many decades but never develop lung cancer (though – unless they are run down by a bus or whatever - they are actually extraordinarily likely to develop *some* smoking related illness, of which of course lung cancer is only one). Also, sadly there are cases of people who have never smoked a cigarette in their lives but who develop lung cancer. But neither type of person, of course, refutes the claim that you are at much greater risk of lung cancer if you smoke cigarettes. These claims have a probabilistic or stochastic rather than deterministic character (with an added 'causality' element that we will consider shortly in part **3**).

Can probabilistic (or stochastic) claims or theories nonetheless be tested? Statistical testing (albeit always stressing the underlying ideas rather than the technicalities) will come up in somewhat more detail later in the course, but the most basic idea of a (classical) test of a statistical hypothesis is as follows.

Suppose, to take the traditional simplest example, you are concerned to test the hypothesis that some particular coin is a 'fair' one: that there is 'just as much chance' of it coming up heads on any given toss, as there is of it coming up tails. This of course translates in quantitative terms into the claim that, with this coin, Probability(heads) = ½.

Assuming that you can't test the coin physically to check that it is completely symmetric about the plane that is parallel to either face and passes through its centre, then the only relevant evidence is to be obtained by actually ***tossing the coin a number of times and recording the outcomes***. Suppose that we decide to test the hypothesis by tossing the coin 10 times.

The problem is that – unlike in the deterministic case - *any* outcome of this test, right through from 0 heads out of 10 to 10 heads out of 10, is compatible with the hypothesis that the coin is fair. At any rate when two events are *independent* (as they are in the case of individual tosses of a coin – the outcome on one toss has no influence on later ones), the probability that both events will occur is obtained by just multiplying the two individual probabilities together. So if we are interested in the probability of two successive tosses both yielding a head, then the probability (both heads) = prob(head on first toss) x prob (head on 2nd toss) = ½ x ½ = ¼ . So the probability that all 10 heads in the test we envisaged *on the hypothesis that the coin is fair* is certainly not zero, but instead ½ times itself 10 times, i.e. $(1/2)^{10}$ – this is a very small number (or, alternatively, one in a very large number), but it is nothing compared, for example, to the probability of any individual ticket winning the UK lottery (around 1 in 14 million) and after all there is a lottery-winner most weeks. In fact it is easy to prove that not only is 10 heads out of 10 compatible with a fair coin, the probability that if the trial of tossing it 10 times is repeated again and again, then the probability approaches 1 (that is, it becomes 'probabilistically certain') that at least one such repetition will yield 10 heads. And if such an outcome is (probabilistically) bound to occur sometime why can't that sometime be now (i.e. the first time we do the trial consisting of 10 tosses)?

But equally *any other* probability for heads (i.e. any hypothesis to the effect that the coin is biased one way or the other to some degree) is compatible with 10 heads out of 10 in such a trial. So how are we ever going to get the statistical evidence to distinguish different claims about the (un)/fairness of the coin? The proposal (first made by the eminent statistician Sir Ronald Fisher) was that we are never going to learn anything of a probabilistic nature unless we adopt a *rejection rule*. That is, unless we make the decision to reject a hypothesis if the evidence that we actually observe has a sufficiently low probability of happening on the assumption that the hypothesis is true. So the idea is that we will reject the hypothesis that the coin is fair if we observe either an extremely high or an extremely low number of heads in our series of 10 tosses. If all 10 turned out to be heads, for example, then even though – as we saw – this is certainly compatible with the hypothesis that the coin is fair, we will say in effect it is just too improbable that it would occur on that hypothesis and hence we reject the hypothesis.

Of course the question immediately arises of what 'extreme' means. Any decision here will be conventional – but Fisher's argument was, as we saw, that we would never learn anything in this area unless we make some such decision. The usual convention is the so-called 95% significance test. In the coin case (or in any similar case) we work out the set of possible extreme outcomes (some very low in heads, some very high in heads) which cover in total 5% of the probability. This set of events is the 'significance region'; and if the actual outcome falls in that region the hypothesis is rejected. I have not done the maths but I would guess that in our case the significance region will contain the outcomes of 0 and 1 and 9 and 10 heads. (The details don't matter so long as you get the general idea.) So if any of those possible outcomes occurs we reject the hypothesis that the coin is fair, but if anything between 2 and 8 heads occurs then we fail to reject (and therefore 'accept') the hypothesis.

Before seeing how, if at all, this ties in with our general ideas about testing, let's see how this idea pans out in more realistic cases – of course we are never concerned in real science with the issue of whether some particular coin is fair! Suppose we are testing the idea that a particular proposed new medical treatment is more effective than some currently accepted treatment. Very few treatments in medicine are 100% effective and so we are dealing here with a statistical/stochastic hypothesis - we expect that some, but not all, of those given the new treatment will recover (assuming recovery is the appropriate 'outcome measure') and similarly some, but not all of those given the currently accepted treatment will recover. How then do we test the claim of greater effectiveness? The standard Fisherian idea is to set up the 'null hypothesis' this says that there is in fact *no difference* between the two treatments and hence that any patient has the same probability of recovering whether s/he is given the new experimental treatment or the old currently accepted one. Just like the hypothesis of the fair dice, this dictates a set of probabilities for observed *differences* in recovery rates recorded in the two groups that is symmetrical about zero (remember this is zero *difference* between the two group outcomes).

The idea, again, is that the null hypothesis is rejected if an outcome occurs that falls in the 'significance region' consisting of those extreme outcomes (big differences in recovery rates between the two groups one way or the other) which together account for just 5% of the total probability.

There are lots of issues about the details here, but the general idea seems to fit quite well with our ideas about testing and how tests results might supply evidence in favour of some theoretical claim. Any positive difference between the recovery rates in the experimental group and the control group would be explained by the theory that the experimental treatment is more effective;

but we only count such a positive result as giving real support if the outcome is very unlikely to occur if the greater effectiveness hypothesis is false (here identified with the ‘null hypothesis’ being true so that there is in fact no difference and so recovery is just as likely whichever treatment you are given).

2. The difference between a statistical and a (non-deterministic) causal hypothesis

What do we mean when we say that ‘smoking causes cancer’ or ‘taking tannins causes fewer heartattacks’? Clearly *not*, as we noted, that everyone who smokes will develop cancer or that no one who takes statins will suffer a heartattack. The claims, as noted, have something to do with increased (or decreased) probability: there is more chance of developing lung cancer if you smoke, less chance of having a heart attack if you take statins. But these claims about probability cannot *completely* capture what we mean by the causal claims. One reason is this.

If two variables X and Y (say X is smoking at least 5 cigarettes a day and Y is developing lung cancer) are probabilistically associated then they, are if you like, probabilistically associated ‘*both ways’.* The way to express a probabilistic connection is via a *conditional probability* What, for example, is the probability of a card drawn at random from a well shuffled pack being a heart, *given* that it is a red card? The answer obviously is ½ (since there are 26 red cards of which 13 are hearts). Thus being a heart and being red are of course probabilistically related: there is more chance of a card drawn at random being a heart if it is red, than there is just of its being a heart: P(heart) = ¼ but as we just saw P(heart/red) = ½. So that is how we express conditional probabilities: P(X/Y) is the probability that event X will occur given that event Y has occurred. Part of the meaning of the claim that smoking causes lung cancer is then that P(lung cancer/smoke) ≠ P(lung cancer) and in fact P(lung cancer/smoke) > P(lung cancer) (‘>’ just means ‘is greater than’).

But this can only be part of the story. It is easy to prove from the principles of probability that if P(X/Y) > P(X) then equally P(Y/X) > P(Y). Probabilistic association is *symmetric*. No need to prove this formally because it is easy to see intuitively: if we took a bunch of cigarette smokers from the general population we would expect to find a greater proportion among them of people who develop lung cancer than we would in the general population; but equally (in fact the two descriptions are entirely equivalent) if we took a bunch of people from the general population who had developed lung cancer we would expect to find among them a greater proportion of cigarette smokers than we would find in the general population. P(lung cancer/smoke) > P(lung cancer) but equally P(smoke/lung cancer) > P(smoke) (where P(smoke) etc should be thought of as’ what’s the chance that someone drawn at random from the general population will be a smoker?’)

However, although probabilistic dependence (to use the technical term) is symmetric, *causal dependence* is clearly *not*. Smoking causes lung cancer but lung cancer does not cause smoking. So there must be something more to this particular causal claim (and this is true in general).

The second reason why probabilistic relations cannot fully capture what we mean by causal claims of this kind is the possibility of committing the *post hoc ergo propter hoc* fallacy. The causal claim that smoking causes lung cancer certainly entails that P(lung cancer/smoke) > P(lung cancer). But it will also surely be true, for example, that P(lung cancer/ there are more than 3 ashtrays in your house) > P(lung cancer) and yet obviously owning ashtrays does not cause lung cancer. Instead, lung cancer and owning ashtrays are two effects of the real cause here namely smoking. The probabilistic dependence holds in both these cases (genuine cause and effect *and* two effects of a ‘common cause’) and hence cannot distinguish between them. Of course, given the sorts of things we all know – ash tray ownership just isn’t the sort of thing that could possibly cause cancer – in this case there would be no temptation to commit the post hoc ergo propter hoc fallacy (revealed as the fallacy of inferring that an observed probabilistic connection must be a causal connection). But we can imagine Martians, built on entirely different physiological principles, observing earthlings, noting that there were many more cases of lung cancer amongst those who own ashtrays and inferring that it would a good idea if ash tray ownership were banned on Earth (assuming they were nice little green ‘men’ who had our interests at heart).

This is often called in applied statistics courses the problem of distinguishing causal connections from ‘mere correlations’. Ashtrays and cancer is my example but another one often cited is the correlation in the Danish population between having storks’ nests on your house’s roof and having a large family: within that population P(large family/storks nests on your roof) > P(large family). (This is apparently a fact about Denmark.) But clearly although having those nests on your house’s roof thus ‘increases your probability of having, or being from, a large family’ it does not do so in a causal way and hence this statistic gives no support to the old tale about babies being delivered by storks! Rather the situation is that the bigger your house, the larger your roof and the larger your roof the more chance of storks nesting on it, and independently, the larger your family the bigger house you will want at least (and so, given favourable economic circumstances, the larger house you will have).

Notice then that the problem of distinguishing genuine causes of this probabilistic kind from ‘mere correlations’ has a definite pragmatic pay-off: interventions will only succeed if you have identified a *causal* connection. Anti-smoking campaigns around the world have undoubtedly reduced the rates of lung cancer; however a campaign to reduce lung cancer rates by banning ashtray ownership would have had no effect except to produce more dirty carpets! Similarly if the Danish government wanted to go on a birth control campaign, it would not be a smart move to do it by sweeping all the storks’ nests from peoples’ houses!

3. Testing for causes (or testing causal hypotheses)

Suppose the Martians I mentioned earlier really arrived at the theory that ashtray ownership causes lung cancer. What tests could we run to provide evidence (assuming of course that the tests came out in the way we expected *and* that we had found some Martian-English translators!) that they were wrong? Surely we would reason as follows: well certainly there are more lung cancer victims amongst the ashtray-owners, but we believe that this is accidental – this means that if we were to divide the whole population into smokers and non-smokers and look at ashtray ownership and lung cancer within the two groups *separately*, then we would find that there were identical (or at least very similar) rates of lung cancer in the smokers who owned ashtrays and those who didn’t; and also identical rates of lung cancer in the non-smokers who owned ashtrays and those who didn’t. That is, although P(lung cancer/own ashtrays) > P(lung cancer), if you further ‘conditionalise on’ or ‘*control for’* smoking then this increase will go away: that is, P(lung cancer/own ashtrays & smoke) = P(lung cancer/smoke). Once you have ‘controlled for smoking’ ashtray-ownership has no further effect on lung cancer rates.

Again, there are a lot of details that could be discussed but the essential point is that this fits in very well with our general ideas about evidence, or at any rate telling, evidence only arising from probing (in Popper’s term ‘severe’) tests. The observed connection between ashtray ownership and developing lung cancer would certainly be explained if it were true that ashtray ownership causes lung cancer. But that observation is no real evidence for that causal claim, because it is *not* improbable that that observation would still result if the causal claim were false (that is some other rival causal claim were true). Really good evidence for a theory is not only explained by the theory but also refutes (or at any rate tells against) plausible rivals. So, since the rival theory that it is smoking that causes lung cancer would also explain the relationship between ashtray ownership and lung cancer (given what we know about the process of smoking), a severe test of the ashtray hypothesis must distinguish it from the smoking hypothesis. That is, the severe test, the only one that would give real evidence for the ashtray hypothesis, is the one that ‘controls for smoking’ by looking at smokers and non-smokers separately and sees if ashtray ownership makes any difference in those two sub-groups separately. And that is the test this hypothesis fails.

But turning from the negative (no one thinks ashtray ownership causes cancer) to the positive (we all do pretty much accept that the connection between smoking and cancer is causal and hence that reductions in the smoking rates have and will lead to reductions in lung cancer rates): how do we know – that is, what is the telling evidence that – smoking causes lung cancer? Well clearly part of the evidence is that there are more cases of lung cancer amongst smokers than there are amongst non-smokers. But Bradford Hill and Doll in the UK (and a US team independently) did not of course stop at this – as we just agreed in the ashtray case, this is not really telling evidence unless other rival explanations are at the same time ruled out by it. And indeed Bradford Hill and Doll considered all sorts of other explanations for the observed correlation between smoking and cancer aside from the causal link: perhaps, just to take one alternative they considered (one that was regarded as very plausible at the time), it was air pollution that really caused increased cancer rates and there happened to be more smokers in the polluted areas than in non-polluted ones. In response to this suggestion they 'controlled for air pollution', that is, they separated out the sub-population of those living in polluted areas (cities and large towns) and the sub-population in non-polluted areas (rural dwellers) and looked at smoking and cancer rates in the two sub-populations: what they found was that in this case (and unlike in the ashtray case) the probabilistic connection did *not* disappear – there were more lung cancer cases amongst smokers *both* in the high pollution areas *and* in the low pollution ones: P(cancer/live in a high pollution area & smoke) > P(cancer/Live in high pollution area) and P(cancer/live in low pollution area & smoke) > P(cancer/live in low pollution area)*; hence the observed evidence would not only be explained if the claim that smoking causes cancer were true, but also it is inconsistent with this very plausible alternative. The method of eliminating alternative explanations in this way by 'controlling' the data for them (that is, by conditionalising on the alternative possible cause) is *the* method of epidemiology. It fits very nicely into our general description of a severe test: each plausible rival supplies a possible way in which the evidence might be explained even though the theory whose evidential considerations are under review were false.

Of course, by the nature of science, evidence can be strong, even very strong but never totally safe from future overthrow. It is strong to the extent that rival explanations have been eliminated. Although we have by now indeed eliminated lots and lots of plausible alternative explanations of the smoking and lung cancer data aside from the causal link and hence have very strong evidence for that link, it is always of course *possible* that tomorrow someone will come up with another factor which reveals that smoking and developing lung cancer are in fact common effects of some underlying cause – that is, an alternative explanation that actually leads to the rejection of the smoking causes lung cancer hypothesis. Indeed Fisher himself (and later the noted psychologist Hans Eysenck) formulated the hypothesis that there is some gene (or more accurately genetic disposition) that *both* increases the probability that someone will smoke *and* (independently) increases the probability that someone develops lung cancer. If so, there would be no point in trying to persuade smokers to give up. Of course, if some particular genetic make-up G could be specified that was the alleged 'common cause' here, then this claim could be tested. And tested via our now well-worn route: divide the population into subpopulations – those with genetic make-up G and those without G and then check to see whether or not smoking continues to raise the probability of developing cancer in both sub-populations; if it does not, then this would be very telling evidence in favour of the genetic hypothesis, if the relationship between smoking and cancer continues to hold in both of these sub-populations this would, on the contrary, be further telling evidence in favour of the smoking causes lung cancer theory.

Sadly neither Fisher nor Eysenck had any idea as to what this alleged genetic predisposition might be and hence could provide no way of specifying G. This means that their suggestion remains *untestable* since there is no way of creating the relevant split into sub-populations to check if the smoking-cancer link remains. Hence it was (and remains) an entirely unscientific, non-evidence based mere logical possibility. (This did not stop the Tobacco Industry vigorously promoting it, of course. Though even they now seem to have pretty well given up.)

* You might find it more natural to substitute probabilities that contrast what happens if you do or don't smoke; that is to read these as:
P(cancer/live in a high pollution area & smoke) > P(cancer/Live in high pollution area & don't smoke) and P(cancer/live in low pollution area & smoke) > P(cancer/live in low pollution area & don't smoke). This would render more naturally the finding that smoking still mattered even when you take air pollution into account: whatever the pollution in the area you live, you are still more likely to develop lung cancer if you smoke *than if you do not*. These formulations may look different

from the ones in the main text, but the two formulations can in fact readily be proved to be logically equivalent. That is P(cancer/high pollution area & smoke) > P(cancer/high pollution area) if and only if P(cancer/high pollution & smoke) > P(cancer/high pollution & don't smoke).