Significance testing with incompletely randomised cases cannot possibly work

Stephen Gorard
Durham University
s.a.c.gorard@durham.ac.uk

Abstract

This brief paper illustrates why the use of significance testing cannot possibly work with incompletely randomised cases. The first section reminds readers of the logical argument of "denying the consequence", and the fallacy of trying to affirm the consequence, of a set of premises. The second section extends the argument of the denying the consequence to the weaker situation where there is uncertainty, and the third shows that this weaker situation is the 'logical' basis for the practice of significance testing when analysing data. The fourth section looks at how the same argument becomes a fallacy when conducting significance tests with incompletely randomised or non-random cases. The final section summarises the implications for analysts, and for their future analyses and reporting.

The logic of denying the consequence

The logical argument of denying the consequence or argument *modus tollendo Tollens* involves two premises - a proposition or assumption (H), and a factual step or statement (~~D~~):

IF H is true THEN D is true
D is not true
Therefore H is not true

This argument in three steps is necessarily valid in logical terms, and can be proved by a variety of means. One is via a truth table. Table 1 shows all four possible combinations of the truth of two binary variables H and D. If we accept the premise "If H is true then D is true" D must be true whenever H is true. This means that the third row of the table (scored through) is not possible. D cannot be false when H is true. The other rows are unaffected – because the premise says nothing about D when H is not true, and so D can be true or false when H is false. If we also accept the second premise that D is not true, then the second and fourth rows of the table, when D is true, are not possible (in italics). This leaves just one row in bold, showing that both D and H are false. Therefore, given the first premise, ~~D~~ implies ~~H~~.

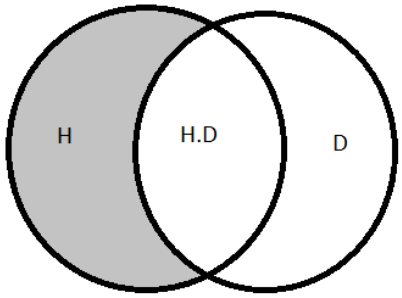
Table 1 – Truth table for denying the consequence

H	D
False	**False**
False	*True*
~~True~~	~~False~~
True	*True*

Because whenever H is true D must also be true, it follows that if D is not true H cannot be true. However, this premise does not imply that D cannot be true when H is not true. As a Venn diagram the situation would look like Figure 1. The left circle represents where H is true, and the right circle where D is true. The two circles have an intersection where both H and D are true. Under the premise "If H is true then D is true", the left hand shaded area H.~~D~~ must be empty. If D is not true so that all of D is also empty, the intersection H.D must be empty as well because it lies within circle D. But otherwise D can be true independently of H (in the right hand segment of circle D, or ~~H~~.D). The logical argument of

denying the consequence itself, as portrayed in Figure 1, does not say anything about the relative size of the areas H.D and ~~H~~.D.

Figure 1 – Denying the consequence



This logical argument is in clear contrast to the logical error of affirming the consequence, which looks like this:

IF H is true THEN D is true
D is true
Therefore H is true

Here the conclusion is invalid. Again, this can demonstrated through a variety of means including truth tables (Table 2). The third row is still not possible because of the first premise, and is scored through. The second premise "D is true" means that the first row of the table is now impossible (where D is false), and is in italics. This leaves two rows (both in bold). In one (the second row) H is false and in the other (the fourth row) H is true. So, when D is true, H can be either true or false and there is no way of deciding between the two, on the basis of D. Affirming the consequence just does not work.

Table 2 – Truth table for affirming the consequence

H	D
False	*False*
False	**True**
~~True~~	~~False~~
True	**True**

This can also be seen via Figure 1. As before, given the premise "If H is true then D is true", H cannot be true without D also being true, and so the shaded area must be empty. Here though, D is true and so the rest of the diagram must be unshaded. Therefore, D can be true without H also being true (the right hand portion of circle D), and so the conclusion "H is true" is not necessarily justified. The conclusion "Therefore H may be true, or it may be false" is perhaps a better summary of the outcome of this argument. This fallacy of affirming the consequence is well-known – as in the age-old example of a fallacious argument "All Greeks are men (*sic*), this is a man, and so this is a Greek". Despite this, affirming the consequence remains a very common error in argument and mistaken scientific reasoning.

Adding uncertainty to the argument

Both of these examples of arguments – one valid and one invalid – are based on binary logical entities that can be definitively declared either true or false. A proposition or assumption H sets up the situation, and then a factual statement about D is used to help decide what this would tell us about the assumption. In real-life such definitive assurances of truth and falsehood are rare, and statements of fact are more often linked to some level of uncertainty about their truth. If the factual step (D) of *modus Tollens* is converted to an uncertain claim, then the first argument becomes:

IF H is true THEN D is true

D is probably not true
Therefore H is probably not true

This is no longer such a clear argument. Looking at Table 1, the third row is still not possible because if H is true then D must also be true. But the other three rows remain possible, even if some are less likely than others. If D is probably not true, then the second and fourth rows are less likely and the first row is more likely. But the argument itself conveys no information about exactly how likely each row is. This is where a Venn diagram might be clearer. The situation is again as in Figure 1. The uncertainty introduced can be envisaged as concerning a likely occurrence of the three different outcomes (or sectors in the diagram). The circles H and D, including their interaction segment, are therefore seen as sets containing possible occurrences. Given the first premise of the argument, the shaded area must be empty (as also shown by the third row of Table 1). There can be no possible occurrences of H and not D. The second premise "D is probably not true" means that there are few possible occurrences in all of circle D. There will, however, still be some such occurrences else D would be known to be false (as it is in the original and valid version of the argument). If all of these possible occurrences are in the right hand area (~~H~~.D) then H is indeed false. However, if even one possible outcome is in the area H.D then H may still be true. As with the fallacy of affirming the consequence, a more accurate portrayal of the conclusion of this uncertainty argument would be "Therefore H may be true, or not". Once the factual part of the argument becomes uncertain the conclusion is no longer logically entailed. And the probability of H being true depends on the precise size (number of possible occurrences) of H.D in relation to D.

There is no way of telling from the uncertainty version of the argument itself how likely or unlikely it is that H is true. The argument says nothing about the relative size of the areas H.D and ~~H~~.D. If H.D is empty, then H is not true (all of circle H would be shaded). If H.D is only a small proportion of D (i.e. its possible occurrences are few in relation to the overall number of occurrences) then H will be unlikely to be true, but still could be (Figure 2). If ~~H~~.D is empty then H will be true, while if ~~H~~.D is only a small proportion of D then H is more likely is to be true but could still be false (Figure 3).

Figure 2 – Denying the consequence, small intersection

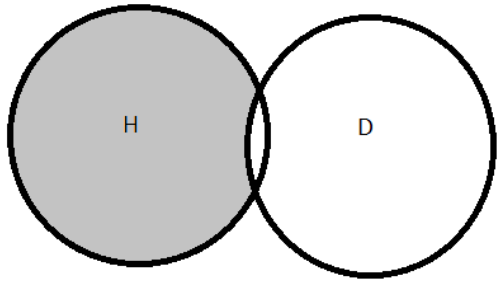
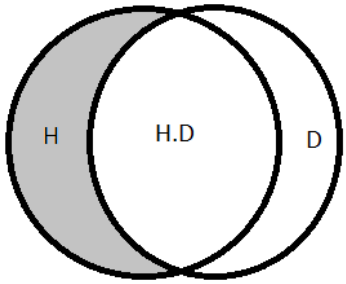


Figure 3 – Denying the consequence, large intersection



Without knowing what proportion of D is also H (i.e. the area of H.D divided by area of D, or n of occurrences in H.D divided by N of occurrences in D), knowing that D is probably empty does not allow us to decide how likely it is that H.D itself is empty - or in other words how likely it is that the original assumption H is false.

The 'logic' of significance testing with fully random cases

The last point is important because it is this uncertain version of the argument that forms the apparent 'logical' basis for significance testing. In a significance test, the first premise (H) is the assumption or hypothesis on which the ensuing probabilities are calculated. This assumption is that any finding or factual claim from the data (D) has been created solely by the randomisation of the cases (sample) from which D is measured. The finding is treated as a fluke arising from the nature of the cases chosen in one 'trial' or study. This assumption is commonly referred to as the null hypothesis (or more strictly the nil null hypothesis), and here it is represented by H. If H is true, it is possible to assess the probability of the factual findings (or more extreme ones) arising by chance. This is the p-value produced by a significance test. The logic of significance testing can therefore be portrayed as:

> IF the findings are solely random (H) THEN the probability of findings at least as extreme as those found is p(D|H)
> p(D|H) is a small probability
> Therefore H|D is probably not true

Here H is that the findings D are solely due to randomisation of the cases. The probability p(D|H) is the probability of a finding at least as extreme as the one found, given that H is assumed to be true. The probability p(H|D) is the probability of H actually being true once the finding has been obtained. It is not logical to try and derive p(H|D) directly from p(D|H). The probability of H can be large when the probability of D is low and *vice versa*. In the same way, if a coin toss is known to be unbiased and the result due solely to chance then it is possible to calculate the probability of getting three heads in a row. But merely getting three heads in a row with a coin that is not known to be unbiased does not say anything at all about whether the coin toss is actually unbiased in real life. The two probabilities are completely different. When analysts use p(D|H) which is the p-value from their significance tests as an estimate of p(H|D) they are committing a clear logical error (as clear as in the affirming the consequence example). This does not make p(D|H) completely uninformative, but it does mean that it can be very misleading.

Converting pD given that H must be true (the p value) into pH given D is not an obvious or simple procedure. Crucially it requires knowledge of further values, such as pD without pH or *a priori*, that are usually not available in real-life. Bayes' Theorem for converting one to the other is:

$$p(H|D) = \frac{p(D|H) \,.\, pH}{pD}$$

This theorem confirms that the uncertain form of the *modus Tollens* argument used for significance tests does not produce a valid conclusion. If p(D|H) is small it does not follow that p(H|D) is small. Whether this is true depends on the unconditional or *a priori* probabilities pH and pD. If pH/pD is very large then p(H|D) is much bigger than p(D|H). If pH/pD is very small then p(H|D) is much smaller than p(D|H). But a significance test only provides the p-value, or p(D|H). It does not provide either pD or PH, and so it cannot be used to assess the probability of the hypothesis being true, or p(H|D). That assessment depends on the proportion of Figure 1 that is in H.D compared to the overall scale of circle D. This proportion cannot be computed from the p-value. The p-value from a significance test is instead an estimate of the overall scale of circle D (how many valid possible occurrences remain, even though pD is small). It says nothing about the relative size of H.D. The p-value can be very small, while the probability of H can remain large, and *vice versa*. It bears repeating that the probability of the finding being true given H is not the probability that H is true; nor can it converted to the probability that H is true without a lot more information than is usually available when conducting a significance test. If the intention when using a significance test is to assess the probability of H being true based solely on the findings from any study, then this approach simply does not work on logical grounds.

The 'logic' of significance testing with non-random cases

However, the real-life situation is almost always much worse than this. Many analysts conduct significance tests based on data obtained from cases that have not been randomised. Significance test results are regularly reported based on convenience samples, heavily incomplete samples, and for population data. Analysts should not do this, because the algorithms used by significance tests are based on full randomisation. However, calculators and computers do not know the source of any data entered, and so will run the significance test anyway if told to by an unthinking analyst. If pressed on their reasons, analysts reporting such tests may suggest that these can still provide some valuable insight into the level of uncertainty in any result. To see why this idea of a significance test providing any useful information for data drawn from non-random cases must be wrong, we can add or make explicit a further premise to the *modus Tollens* argument. Here H again represents the assumption that any findings are solely due to chance, and R is a new premise stating that the findings were obtained from cases that were fully randomised (as they must be according to the mathematics on which the p-value is based). Randomised could mean a sample chosen randomly from a known population, or where the population was randomly allocated to two or more groups. Cases can be randomised (R) and the initial hypothesis can then be true (there is no pattern in D) or false (there is a substantive finding not created by randomisation alone). H and R are independent of each other to that extent.

The logical version of the argument, illustrated without uncertainty at first in order to make the point as simply as possible, would be:

> IF H is true AND R is true THEN D is true
> D is not true
> Therefore H is not true OR R is not true

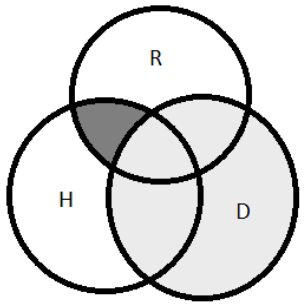
The possible results are shown in Table 3. There are now eight possible combinations of the three binary variables H, R and D. The first step of the argument assumes that D is true when both H and R are true. Therefore, the seventh row of Table 3 is not possible (and is scored through). All other combinations are possible because either H or R are false or because D is true. The second step of the argument is that D is not true, and so the four rows where D is true are also not possible, and appear in italics. This leaves three combinations, and in all three (in bold) at least one of H or R must be false.

Table 3 – Truth table for denying the consequence with an extra premise

H	R	D
False	**False**	**False**
False	*False*	*True*
False	**True**	**False**
False	*True*	*True*
True	**False**	**False**
True	*False*	*True*
~~True~~	~~True~~	~~False~~
True	*True*	*True*

This is illustrated further in Figure 4. Now there are three circles – one each for the premises H and R and one for D. Where H is true and R is true then D is true, and therefore the segment of the Venn diagram where H is true and R is true but D is not has been darker shaded, and must be empty. If D is not true then all of the right hand circle would also be empty and is lighter shaded. However, either H or R could still be true, or neither could be true. Put another way, at least one of H or R must be false (because if both are true then D is true), under these conditions. But we already know that R is actually false because we are attempting a significance test with non-random cases. Therefore, the statement that D is not true yields no useful information about H or about R. R is already known to be false. And as a consequence, H could still be true or false, and the 'test' provides no information at all about which.

Figure 4 – Denying the consequence, with non-random cases



Adding uncertainty makes the situation more complex but the argument still yields no useful information about that uncertainty or anything else. The argument would be:

> IF we imagine that the cases are randomised (R), and that the findings are solely random (H)
> THEN the probability of findings as or more extreme as those found is p(D|(R.H))
> p(D|(R.H)) is a small probability
> Therefore (R.H)|D is probably not true

This argument is completely pointless. Even if we accept the logic of significance testing, the conclusion that one or both of the initial assumptions is unlikely to be true does not help at all. If we have run a significance test based on non-random cases then we already know that R is false. If R is known to be false then the probability of R AND H being true given the data D must always be zero. The data D is irrelevant, and nothing can be learnt from it about the probability of H through this spurious process. So although it is possible to play the game of imagining that non-randomised cases are actually random and so compute a p-value, this can never ever yield any useful conclusion or indeed any information at all (and this is so even if we accept that the logic of significance tests would have made sense with random cases). Under these conditions running a significance test is completely absurd. It is a pseudo-analysis that looks as if something logical or scientific is being done but is really a kind of ritual incantation, or conjuring trick for the unwary.

Conclusion

Running significance tests with randomised cases, and using p(D|H) as some kind of estimate for p(H|D) is not logical without prior knowledge of the unconditional probabilities pH and pD, and is very misleading in practice. But running significance tests with non-random cases is even worse. It is a purely empty ritual, and will always mislead analysts and readers who accept p-values as somehow informative about the level of uncertainty in H. Non-random samples include samples designed to be random but which are incomplete because some cases are missing (non-response), drop out or do not provide key data. This means that, even if they worked with ideal data, significance tests should not be used with real datasets where any data is missing (or recorded in error). This covers just about every real-life use of significance tests. Nor should any technique predicated on the same assumptions as significance testing (such as power calculations or confidence intervals) be used. Logically they will always tend to mislead, and in practice with non-random samples (i.e. nearly all real-life datasets) they will always mislead.