Why animal toxicology often fails to inform the finder-of-fact.

Animal tests are used to identify and characterise toxins. When presented in court, positive results are evidence in favour of foreseeable harm to humans. Duty of care standards and causation arguments follow.

The standard statistical method for interpreting the results of animal experiments is the "*t-test*". Much reliance is placed on this test when the results are marginal – precisely the results which drive emerging risks research.

The t-test is adopted in order to serve a precautionary purpose suited to regulation and risk aversion, not to serve the standards of evidence required of a reasonable person. Experts who rely on the use of marginal toxicity t-testing to support arguments on foreseeability, generic causation, specific causation and duty of care standards may be inadvertently importing a standard of analysis which is not openly informative at common law.

Marginal results create uncertainty in the management of emerging liability risks. In theory, risk managers must adopt the same standards that will be used in court. By tradition this is the balance of probabilities test, but expert evidence based on a precautionary analysis may hold sway. The systems of research, financial responsibility, regulation and common law fact finding clash.

An alternative "*probability of difference*" test is proposed for use in the common law and financial responsibility settings. Unlike the t-test, this is transparent, compatible with the common law and, stable. It is also very easy to use and is intuitive. The use of the t-test in animal experiments is conspicuous in its unjustifiable assumption of "means". This is explained in detail in this paper.

Introduction

Animal tests are a valuable source of information on the potential for toxic effects in humans. Absence of toxicity at realistic doses, and a hundred times higher, provides reassurance. On the other hand, the presence of toxicity elicits a precautionary response and the search for the 'safe' exposure level. Much depends on being able to show that groups of animals, treated differently, show different rates of disease. The convention is to apply the t-test to the data.

This paper explains the basis of the t-test and then explores the incompatibility between the t-test and the finding of fact at common law. Having concluded that in marginal cases the animal test statistics are falsely conceived and *most likely wrong*, I propose an alternative known as the "*probability of difference*" test. The t-test approach creates the opportunity for injustice by biasing views on causation, foreseeability and duty of care standards. Confusion over whether the courts will adopt traditional reason for fact finding, or instead of this a precautionary standard makes it particularly difficult to reserving for emerging liability risks. A precautionary, i.e. risk averse, standard is a policy choice. It is chosen for regulatory pruposes.

Rejecting the null hypothesis

It is with this potential for a precautionary response in mind that leads to the conventional choice of statistical methods. Known as the "*t-test*" the idea is to measure how likely it is that the results from two groups of test animals are NOT different. That is, how likely is the null hypothesis? If the probability of the null hypothesis is less than 5%, then the convention has it that the results from the two groups are different. If one is the control group and the other is the experiment group then a "p" value of less than 5% indicates that the experiment group is detectably different from the control group. If the only known difference between the groups is the presence of a proposed toxin then the difference is perhaps readily assumed to be because of the toxin.

Is this also good enough to inform fact finding at common law?

A five percent probability of the null hypothesis being correct implies a 95% probability that it is incorrect. That is, it is 95% certain that the experiment group had a different rate of injury than the control group. Surely 95% would satisfy the balance of probabilities test required for evidence at the common law?

A closer look at the t-test as used in animal experiments suggests reasons to doubt that it can be used to settle issues of fact at common law.

A typical experiment

Four or five equal groups of closely related rats are housed and handled in the same way for two years. One group receives daily a harmless substance and the other four groups each receive a different daily fixed dose of agent X. Doses cover a wide range of values so as to provide a thorough test of toxicity.

For example, there are 90 rats per group.

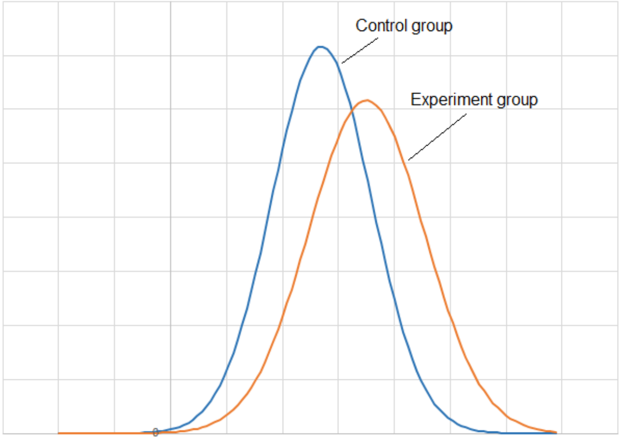
After 2 years the rats are examined in great detail. 4 rats in the control group develop a specific condition. In the lowest dose group there are 5, in the next 3, in the next 4 and in last also 5. Does agent X increase the risk of this condition? Medical science would say the p value for the null hypothesis is 0.0014. That is, there is 0.14% chance that NO is the right answer. By symmetry, there is a >99.8% chance that the answer is YES, agent X is associated with an increased risk.

But what if the control group had developed 5? What if it had developed 3? How likely is it that 4 is the right answer? If the control group had by chance developed 5 diagnoses then agent x would be considered for its protective effect.

In practice, the control experiment is done only once, and the specific condition is either present or absent. If the experiment was then done 100 times the value observed on this first occasion could turn out to be the mean value, or it could be an outlier and the true mean was 18. If the true control mean was 18, does that mean that agent X is strongly protective? The problem is, that the t-test as used in animal experiments is based on the assumption that the observed value in a once-only experiment, IS the mean value.

Comparison of true means

When experiments are repeated many times the values often turn out to be distributed normally about the mean, with a width characterised by the standard deviation (SD). Suppose we have two experiments repeated many times and then compared. The raw data might look as follows:



Given that the illustrated difference is marginal a reasonable question to ask is: would it be possible, on at least some separate occasions, for the number of disease cases in the control group to be higher than the number of disease cases in the experiment group? The answer is clearly, yes. Just because on one occasion the control group disease rate was below the experiment group disease rate doesn't mean that it would be the next time the experiment was done. It may even be an unusual event.

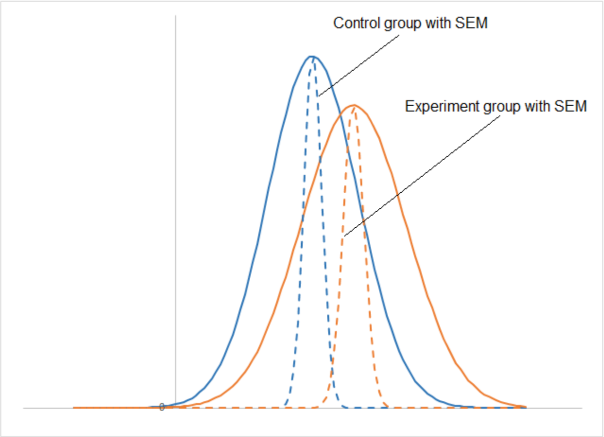
Given the curves in the figure above, the probability is, that if you did this many times, the control group disease rate would on average be lower than the experiment group disease rate. But if you only made one measurement pair there is a very good chance that the control group rate is higher than the experiment group rate. So, what I'm suggesting is that there is a potential for considerable doubt about the interpretation of any one separate pair of observations with a marginal difference.

For non-marginal differences uncertainty is unlikely to arise. For example, 4 (±2) controls have the disease but 32 (±5) of the experiment group do. For this, you don't need a t-test. [The error values provided in brackets are taken from a binomial theory analysis.] To make this plain, there is a 1% chance that the control group would generate as many as 9 diagnoses and a 1% chance that the experiment group would generate 23. Note, 9 is clearly well below 23. The groups are different.

The t-test

How does it work?

In data analysis one valid comparison is to measure the difference between measured means. For true means, the comparison relies on the mean and the standard error of the mean (SEM). SEM is SD divided by the square root of the number of animals in each group. This illustrated as follows:



When comparing the means, with their now very much reduced spread as generated by SEM, there is very little chance now that the experiment group mean is equal to or below the control group mean. The experiment has a very good chance of rejecting the null hypothesis. This comparison is the t-test. The null hypothesis is rejected if the overlap between the SEM curves is less than 5%.

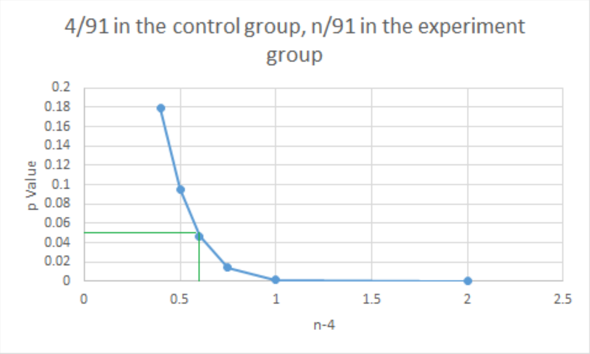
Presumption

But the approach shown here presumes that the observed number of diagnoses is the same as the arithmetic mean that would be found from a large number of identical experiments. There may be good reasons to question this, as illustrated above. The binomial distribution provides a quantitative analysis which may be used to assess this doubt. See below.

Instability

It is also a cause of concern, that the t-test is exceptionally sensitive to small changes.

In the following graph the number of diagnoses in the experiment group (i.e. n) is varied by sub-unit amounts, and the 5% threshold is shown in green. The comparison goes from obviously insignificant to exceptionally precise in less than one change of diagnosis. This means that slightly different conditions could transform the outcome from ‘no problem’ to ‘certainly toxic’.



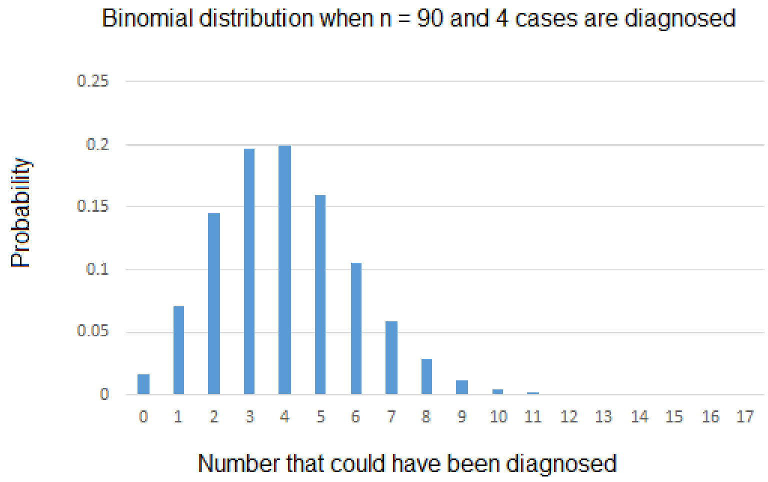
Such instability would not seem to provide a sound basis for advising a common law court. The instability is generated by the very steep curvature in the two SEM Gaussians where they meet.

The binomial distribution

How likely is it that a one-off observation really is a true mean?

To answer this we need to use the binomial distribution.

In the following diagram the bars represent the probability that a given number of animals will be diagnosed. In this example, there are 90 animals in the group and 4 were actually diagnosed with the disease. As you can see, 4 is only marginally more likely than 3. It is also plain 2 and 5 are about equally likely but clearly less likely than 4. And so on.

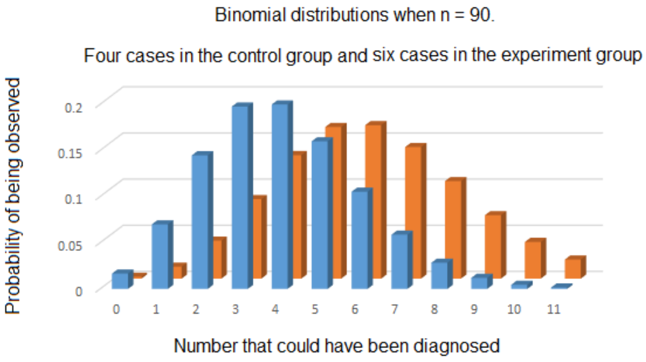


The mean of this distribution function is 4 and the standard deviation is 1.95.

Note that there is only a 19.9% chance that if the experiment was repeated, the observed number of diagnoses would be 4.

80% of the time you would get a different "mean" to use in the t-test.

The t-test mean presumption problem is illustrated further by comparing the binomial distributions for 4 diagnoses in the control group and 6 diagnoses in the experiment group.



In this example there is a 27% chance that the experiment group result would be at or below 4 and a 21% chance that the control number would be six or more. A pair-wise comparison of SEM curves doesn't change this at all; very precisely wrong is still wrong.

So what do we do?

What we need is a method which compares the full range of values that could reasonably be expected to represent the control group and the experiment group. The binomial distribution represents all the possible "means", weighted according to probability.

The question to ask is, *what is the probability that the two observed diagnosis rates are different*? If the probability of difference is greater than 50% then this would satisfy the common law standard test of fact.

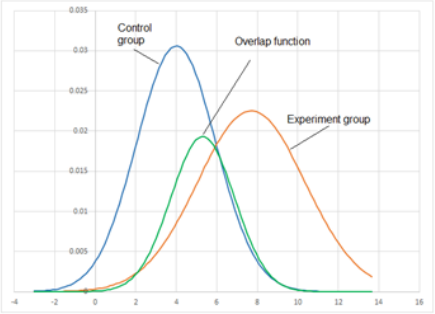
The *Radar*[1] service has developed just such a method. It is referred to here as the "probability of difference" test. It is available as an excel spreadsheet.

In pictures

The maths used here is long established, but applying it to animal experimentation seems to be an original idea.

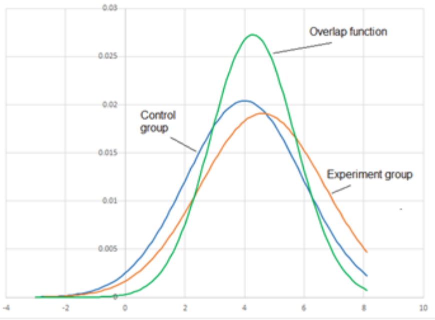
In both graphs the number of diagnosed control animals is 4 out of 90, with a probability distribution illustrated in blue. In orange is the distribution obtained when the test of fact shows a significant difference.

On the left, the test of fact is the "probability of difference test" with the difference set at 50%. On the right, the test of fact is the "t-test" set at p = 0.05. The significant mean for the probability of difference test is ≥ 7.7. The significant mean for the t-test is ≥ 4.6.



The point at which a common law compatible test establishes the fact of difference.

The overlap function is generated by the "probability of difference" tool, developed by *Radar*.



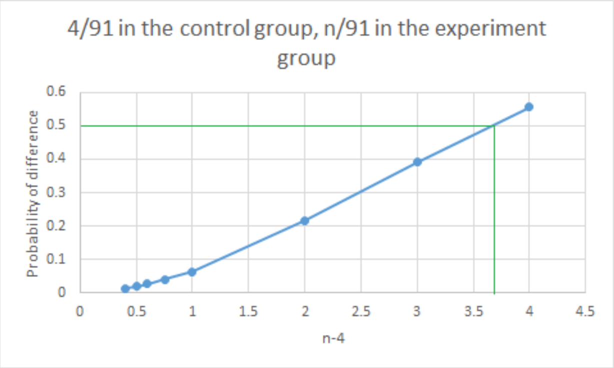
The point at which the t-test establishes the fact of difference.

In both graphs, the area of the green distribution represents the probability that the blue and orange curves are measuring the same thing. On the left, the value is 50%, the common law test of fact. On the right it is >97%. That is, the control group and the experiment group are very probably not different when the t-test, with all its assumptions, is satisfied. A measured value of 4.6 is probably measuring the same thing as a measured value of 4.0.

Sensitivity of the 'probability of difference' test – the effect of small changes?

The *probability of difference* test has a linear trajectory either side of the 50% threshold. The effect of minor variations can be considered. Opinion about the effect of minor variations is readily informed.

[1] The *Radar* service provides information to liability insurers. The aim is to inform their judgement as to scientific issues and how these might influence their business. www.reliabilityoxford.co.uk . Subscribers are provided with an Excel tool that does the work for them.



Summary of the animal stats argument

Animal test scientists need to know if exposed animals are at higher or lower risk of disease than those in the control group. The statistical test they use assumes that the number of animals observed with the diagnosis is the mean value that would be obtained if the experiment was repeated many times. *This assumption is probably wrong* in typical experiments. It does however satisfy the precautionary ethic of regulatory science in that minor differences are found to be significant and there is an opportunity to justify an intervention. The test is highly unstable to tiny variations in experiment.

By comparison, the *probability of difference* test makes no assumptions, meets the standard test of fact used at common law, is very stable to variation and is suited to sensitivity analysis and opinion formation.

While I don't propose that the probability of difference test is applied across the board I do propose that in marginal cases when experts state that causation was demonstrated in animal experiments they be cross-examined on the statistical tests they use. Regulatory science should continue to use the tests it is familiar with for regulatory purposes but expert witnesses should not presume that these will be informative at common law.

One problem for liability risk managers is that the precautionary approach is habitually adopted by expert witnesses unless they are specifically instructed to adopt the balance of probabilities test. Even so, and for its own reasons, the court may decide on a case by case basis if it prefers traditional reason or, precaution. Those reserving for emerging risks are required to adopt the balance of probabilities test when assessing the probability of loss, but when those setting the pace are free to choose precaution or traditional reason at will, this is not a race that can be won.

Dr Andrew Auty