STATISTICAL METHODS

When differences are equivalent

Graham McBride

Traditional statistical tests may not provide satisfactory answers to questions of environmental impact because they may not be asking the right questions. A new procedure – equivalence tests – may do so. Equivalence tests recognise that while differences always occur (e.g., between upstream and downstream sites), they may be small enough to be considered "equivalent". The required calculations are no more difficult than those used in traditional tests.

MOST DISCOURSES on statistics start from the general and then go to the particular. This article inverts that order, in the hope of making the material more digestible! A number of calculations are referred to as we go along. These have been performed using standard functions in a Microsoft Excel spreadsheet. A one-page summary of the calculation procedure is available from the author.

Gold mining impacts

My ecologist colleague John Quinn and his co-workers examined benthic (streambed) invertebrate communities upstream and downstream of alluvial gold mining operations on six streams on the South Island's West Coast (Quinn *et al.* 1992). The effect of the mining was to increase the cloudiness of the stream water and the accumulation of bed sediment.

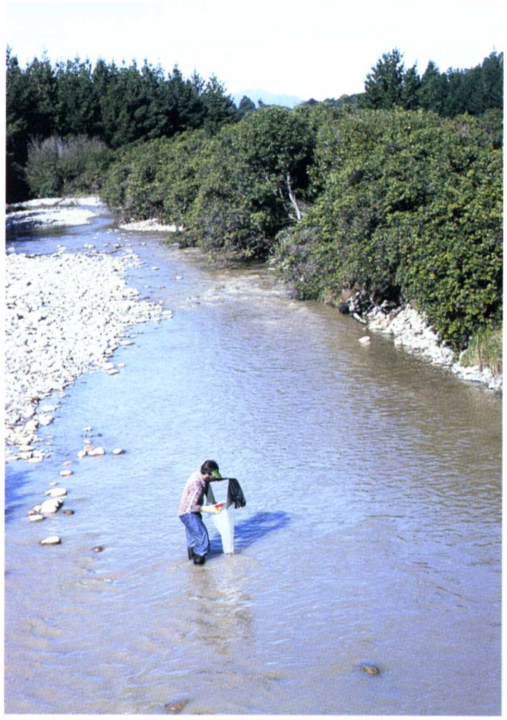
In each stream, the survey team collected seven upstream replicate samples of invertebrates and seven downstream replicate samples. Each sample was taken from a bed area of 0.1 m^2 at a site in a "run". A run is a stretch of steadily flowing, unbroken water, intermediate in depth and velocity between a "riffle" and a "pool". The upstream and downstream sites on each stream were chosen so that they were similar in character.

From their surveys the ecologists were able to calculate, among other things, the "taxonomic richness" at each site. This is the number of invertebrate species recorded in the 0.1 m^2 measured at each site, averaged over the seven replicates. As shown on the figure (below) a reduction in the average taxonomic richness from upstream to downstream was measured in all streams, although in Waimea Creek the reduction was very small. Of course the samples are only a small fraction of the benthic invertebrates of the stream, so we can never know whether these results truly represent the mining impact. We need to be aware of this uncertainty when addressing the important question: in which streams is the change in taxonomic richness really different and in which streams is it not?

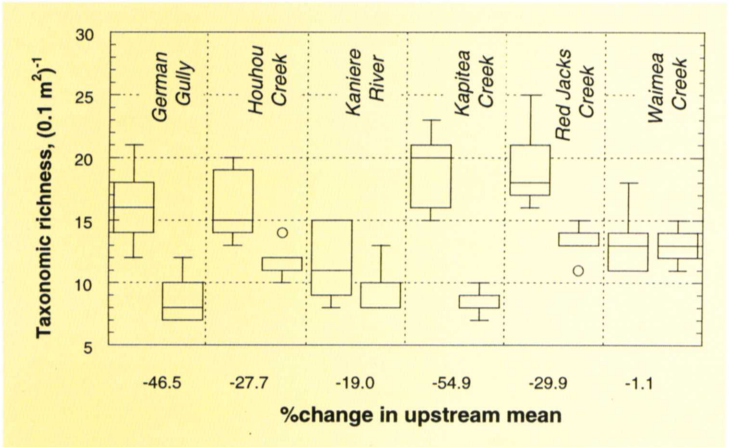
This is where statistical "hypothesis testing" comes in – because it deals with the uncertainty.

Traditional null hypothesis test

An hypothesis is a proposition that we make as a starting point for our investigations. A traditional null hypothesis test examines the hypothesis that there is no difference at all (not to the zillionth decimal place!) – hence "null" – between the things we are comparing. The hypothesis is tested by calculating the probability of getting results at least as different as those measured, merely by chance, *if this hypothesis were true.* Statistical theory enables us to calculate this probability quite simply. If the probability is small (usually taken as less than 5%) we reject the hypothesis and say that we have found a "statistically significant" difference.



above: *John Quinn sampling invertebrates in Kapitea Creek, downstream of the mining operation (13 April 1989). (Photo: Rob Davies-Colley)*



left: *Boxplots of taxonomic richness data for six streams (from Quinn* et al. *1992). The first of each pair of boxplots is for the upstream site. The boxes contain half the data and the line through each box is that site's median value. The crosses are the site mean values and the circles are outliers. The boxes and whiskers indicate the spread and skewness of the data.*

Outcomes of difference and equivalence tests for upstream/downstream benthic invertebrate taxonomic richness data

German Gully	Houhou	Kaniere	Kapitea	Red Jacks	Waimea
*Null hypothesis test (*denotes "statistically significant" result)***					
*	*	-	*	*	-
*Tested hypothesis is "equivalence" (i.e., true difference is less than 20% of upstream value) ***					
Inequiv.	Equiv.	Equiv.	Inequiv.	Equiv.	Equiv.
*Tested hypothesis is "inequivalence" (i.e., true difference is greater than 20% of upstream value)***					
Inequiv.	Inequiv.	Inequiv.	Inequiv.	Inequiv.	Equiv.
Bayes' probability that then change in taxonomic richness is within 20% of the upstream value (%)					
0.3	14.0	53.3	0.01	7.7	97.1

**The significance level for all comparisons is $\alpha = 5\%$. For further discussion on this and on the calculation methods (including their application to the Waimea Creek data), refer to the explanatory sheet available from the author.

Using that test on John's data shows that the differences in upstream/downstream taxonomic richness at four of these six sites are statistically significant. The exceptions are Kaniere River and Waimea Creek (see the Table above). This finding would often be interpreted to mean that a "real difference" has been detected between upstream and downstream sites on the other four streams.

No impact at all?

But why would you believe that the null hypothesis could actually be true – that there is no difference, none at all? Surely a mining operation could be expected to have *some* impact? And environmental factors alone could be expected to cause differences between sites, even if there were no mining operation.

So, is the right question being asked? Are the differences in invertebrate animal communities detected on the four streams ecologically significant? The answer is "not necessarily". It can be shown that finding a statistically significant difference becomes ever more likely with increased numbers of replicates, and that this is because we are testing a null hypothesis. That is, the detectable difference tends to become smaller with a larger number of samples. So a "statistically significant" difference is not necessarily "ecologically significant". Perhaps the wrong question is being asked.

Similar questions arose in drugs-testing some years ago. In this area of research it is now largely agreed that testing a null hypothesis is not appropriate. Why should one believe that two drugs could have exactly the same effect? It has become common practice for drugs-testing (e.g., Chow and Liu 1992) – but for practically no other field – to test whether or not a difference might be within or beyond some prescribed interval, rather than futilely imagining that it might be exactly zero. The size of the interval is set by the drugs licensing agencies and is set small enough to provide appropriate health protection to patients. Testing for differences falling within a given interval is generally known as "equivalence testing".

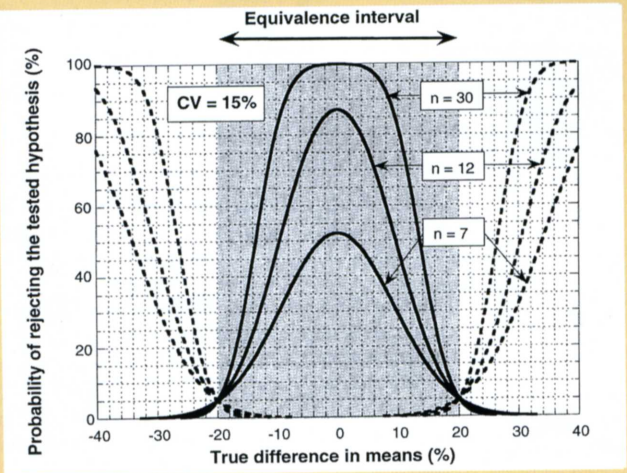
For benthic invertebrate data, the equivalence interval should correspond to differences judged by ecologists or regulators to be "ecologically significant". If we conclude that the true difference is within that interval we would say that the sites are "equivalent"; if not, we would say that they are "inequivalent", and so infer that there has been an impact. This language avoids the use of the term "different": after all, everything is different to some degree. A major advantage of this approach is that increased sampling can strengthen the support for a true hypothesis (see the "Power curves" panel). In contrast, support for a null hypothesis is weakened by increased sampling.

Acknowledgements

This work was funded by the Foundation for Research, Science and Technology (contract C01616). John Quinn kindly made his West Coast stream data available. Colleagues (Judi Hewitt, Kevin Collier, Niall Broekhuizen, Rob Davies-Colley, Cathy Kilroy and John Quinn) reviewed the text.

Power curves: strengthening or weakening support for an hypothesis

THE FIGURE BELOW (from McBride, in press) shows "power curves" for equivalence tests on data like the West Coast streams' taxonomic richness. (These data had a coefficient of variation (CV) of about 15%.)



Power curves show the probability of rejecting the tested hypothesis for a range of the true (but unknown) differences in mean taxonomic richness for various numbers (n) of replicates. The solid lines refer to tests of the inequivalence hypothesis (that the true difference is outside the equivalence interval) and the dashed lines refer to tests of the equivalence hypothesis (that the true difference is within the equivalence interval). The equivalence interval is ±20% of the upstream mean taxonomic richness. The significance level is $\alpha = 5\%$, so that each curve passes through the 5% rejection probability at the edges of the equivalence interval.

The figure shows that the further the true difference lies within the equivalence interval, the larger the probability of rejecting the inequivalence hypothesis, and the smaller the probability of rejecting the equivalence hypothesis. The converse applies if the true difference lies beyond the interval. Furthermore, the window of uncertainty surrounding the edges of the equivalence interval becomes smaller as the number of replicates increases. For example, with only seven replicates the power to reject the inequivalence hypothesis (and so infer equivalence) only exceeds 50% when the true difference is less than 3%. For 10 replicates this power is attained for a true difference of about 10%, and for 30 replicates it is attained at a true difference of about 14%. Once power exceeds 50% an hypothesis will be rejected.

Performing the equivalence tests

First we have to state the size of the equivalence interval. In general John reckons that a change of 20% from the upstream average taxonomic richness is environmentally significant. (He also looks at other information of course, such as loss of keystone species – species which are known to be critical to the structure of the stream community.)

Armed with this insight we can perform equivalence tests on his data. But first we must face a new, and important, question. That is, which of the two possible hypotheses should we test: (a) that the differences are equivalent, or (b) that they are not? Quite different answers can result, particularly if our measured difference is close to the edge of the equivalence interval. This is because of our demand for a "small" probability of making an error if the tested hypothesis is true. We are in effect saying: I will only reject my hypothesis if strong evidence is produced against it. That of course means that weak evidence won't count against it.

For example, the measured change in taxonomic richness in the Kaniere River – 19% of the upstream value – is very close to the critical value of 20%. If we test the hypothesis of inequivalence [case (b)], we might expect to have a hard time rejecting it for the Kaniere data, even though the measured difference was a little less than 20%. Rejection should be much easier for the Waimea where the measured difference was only 1%.

But let's first follow the standard practice in the environmental sciences of assuming no impact. Therefore we test the equivalence hypothesis [case (a)]. That is, we assume that any differences are small enough (i.e., less than 20% of the upstream value) for us to consider the upstream and downstream sites on the streams to be equivalent in their taxonomic richness.

The result is that we reject the hypothesis for two of the six streams (see Table) and conclude that only German Gully and Kapitea Creek have inequivalent taxonomic richness upstream and downstream of the mining operation. The other four are "equivalent", and it could be inferred that there is unlikely to have been an impact of any note on them.

But should we have made the equivalence assumption? If we want to emphasise environmental protection shouldn't we first assume that the sites are inequivalent, and only lose faith in that assumption if there is strong evidence against it (as in drugs testing)? If we do that, we find that only one of the streams, Waimea Creek, has equivalence of taxonomic richness between upstream and downstream sites (see Table).

So, if we take the latter (precautionary) approach we conclude that five of the streams are impacted by the mining operations. But if we take the opposite tack (testing the equivalence hypothesis, so minimising the chance of "crying wolf" – claiming an impact when it is not ecologically significant) we conclude that only two streams are impacted.

Which is right? Well, the statistician can't say! It's all a question of what burden of proof is adopted. For example, in criminal proceedings there would be many more convictions if juries were instructed to assume the defendant guilty, unless found innocent "beyond reasonable doubt". For our stream data one could be tempted to adopt the intermediary results given by the null hypothesis test (because they imply that four streams are affected). However, it is merely coincidental that these results are intermediary.

Yet another way

All of these techniques work by a procedure which first assumes an hypothesis (e.g., equivalence) to be true, and then asks the question: "what then is the probability of getting data at least as extreme as this, just by chance?" That probability is used as a weight of evidence against the hypothesis. But the more direct question, and some would say the more interesting and relevant question, inverts this to ask: "what is the probability that the taxonomic richness at the upstream and downstream sites are equivalent, given the actual data we have obtained?" To answer this question one has to use "Bayesian" statistical methods. These only work if the investigator (or regulator) is prepared to state the degree of belief held in the hypothesis before the data were collected. The procedure then updates this belief in the light of the actual data obtained.

The good news is that this prior belief can take the form: "I don't know". Doing that, we can use a Bayesian equivalence test procedure to calculate the probability that the true difference in taxonomic richness is less than 20% of the upstream value. The results are also shown on the Table. The probabilities shown can also be thought of as a weight of evidence for or against equivalence, and hence constitute a test.

Conclusion

The above discusses a number of ways of testing whether upstream and downstream sites are equivalent. How are we to interpret the results? Well, it's up to you! If you want my opinion, I'd go for the Bayesian results because they are giving a direct answer to the question asked. And I'd say that the effects of mining operations on the taxonomic richness of benthic invertebrates were negligible on one stream, marginal on another, likely on two and definite in the other two. ■

Graham McBride is based at NIWA in Hamilton.

Further information

For more information about equivalence tests and for copies of the one-page summary of the calculations mentioned in the article, please contact the author: Graham McBride, NIWA, PO Box 11-115, Hamilton (ph. 07 856 1726, fax. 07 856 0151, email: g.mcbride@niwa.cri.nz).

Further reading

Berger, J.O. and Sellke, T. 1987. Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association 82*: 112-139 (with discussion).

Chow, S.-C. and Liu, J.-P. 1992. *Design and Analysis of Bioavailability and Bioequivalence Studies.* New York: Marcel Dekker.

Dixon, P.M. 1998. Assessing effect and no effect with equivalence tests. *In*: Newman, M. and Strojan, C. (eds). *Risk Assessment: Logic and Measurement.* Chelsea, MI: Ann Arbor Press.

McBride, G.B. 1997. Potential for use of equivalence testing in environmental science and management: a review. *NIWA Science and Technology Series No. 45*, 37 p.

McBride, G.B. in press. Equivalence tests can enhance environmental science and management. *Australian & New Zealand Journal of Statistics 4*(1).

McBride, G.B., Loftis, J.C. and Adkins, N.C. 1993. What do significance tests really tell us about the environment? *Environmental Management 17*(4): 423-432 (errata in 18: 317).

McDonald, L.L. and Erickson, W.P. 1994. Testing for bioequivalence in field studies: has a disturbed site been adequately reclaimed? *In*: Fletcher D.J. and Manly, B.F.J. (eds). *Statistics in Ecology and Environmental Monitoring.* Otago Conference series No. 2, University of Otago, Dunedin: 183-197.

Phillips, K.F. 1990. Power of two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics 18*(2): 137-144.

Quinn, J.M., Davies-Colley, R.J., Hickey, C.W., Vickers, M.L. and Ryan, P.A. 1992. Effects of clay discharges on streams. 2. Benthic invertebrates. *Hydrobiologia 248*: 235-247.

Rodda, B.E. and Davis, R.L. 1980. Determining the probability of an important difference in bioavailability. *Clinical Pharmacology and Therapeutics 28*: 252-257.

Schuirmann, D.J. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics 15*: 657-680.