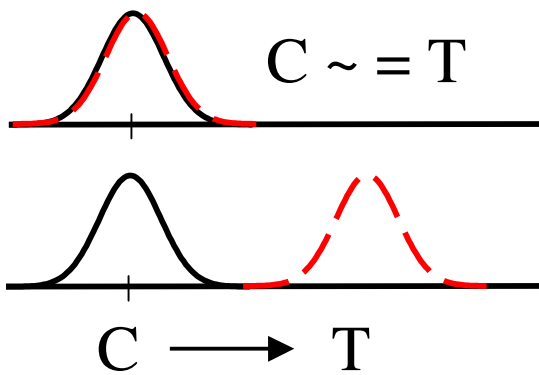


Valid Inference from Data

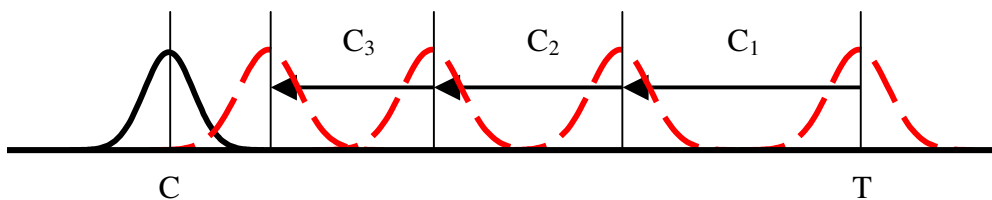
The first point is that research data should be made available. If someone is making a claim, then to assess the claim it makes sense that the data be available. A National Academy of Science panel agrees and their thoughts are given in a book, *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Congress has acknowledged that if data is being used to make a regulatory decision, it should be made available. It is just “fair play” and “good science” that data be available.

The second point is that multiple testing is an important issue. What is multiple testing? It is just asking many questions using the same data set. If you ask 20 questions and statistically test at 0.05, then you expect one “statistically significant” result by chance alone. About 65% of the time, you will see one or more “statistically significant” results. So getting a statistically significant result from a complex study where many questions are asked is not unusual at all. False positives are common in complex experiments where many questions are asked. Ioannidis, JAMA, 2005, estimated that 80% of the claims coming from observational studies failed to replicate. Not adjusting for multiple testing is a big part of the problem. More recent data supports the 80% false claims rate. Experimenters should carefully note how many questions are under consideration and they should adjust their analysis to reflect the number of questions asked. If they do not adjust for multiple testing, then their p-values (and confidence limits) are too small (narrow) and are likely meaningless



The third point relates to questions of confounding bias, i.e. that the observed result could be in fact related to other factors than the one under investigation. In a randomized clinical trial, through randomization, the effects of this type of bias are largely, but not completely, removed. If treatment has an effect it will move the distribution of the treated patients away from the control patients. A p-value can be used to assess the likelihood that two samples came from the same parent distribution. If the two distributions move apart and do not overlap, it is argued that the treatment had an effect.

Observational studies are much more subtle. Two populations will most typically differ in the property of interest, but that difference could well be due to one or more confounding variables. For example, two high schools might differ in the speed with which their students can run the 100m dash, but the difference would be largely explainable if one school had only male students and the other only female students. Gender is the confounding variable. In complex observational studies, confounders can be used to statistically adjust (correct) for bias.

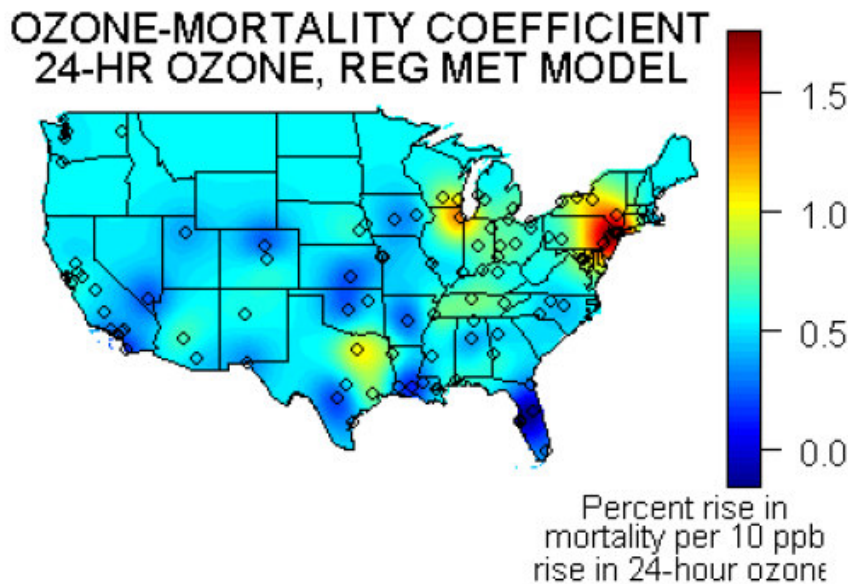


We see here that T-group mean moves toward the C-group mean as confounders  $C_1$ ,  $C_2$ , and  $C_3$  are identified and their effects removed. Suppose that an unidentified confounder is left in so that the mean of the C-group and the T-group, corrected for identified confounders, still differ. Increasing the sample size will increase power to detect an effect, but the effect is due to the bias that has not been removed.



The T-group will be declared different from the C-group, unfairly.

There are more complex problems with statistical analysis. Consider the claim that ozone is related to death rate.



Very small changes in ozone levels have been associated with increased death rate. But this average overall effect is actually very uneven geographically. (Figure from R.L. Smith, UNC, Statistics) If there is an effect, it is much more pronounced in New York/New Jersey. (The effect in the Chicago area is largely the result from eight very hot days.) The effect, if it is real, is quite small and could be due to some other confounding variable. In any case, ozone does not seem to be a general problem; if it has an effect, the effect appears to be much more local.

So what is the status of these getting-valid-inference-from-data problems? Data availability is a solvable problem: funding agencies can mandate sharing of data, and journals, by refusing to publish papers relying on unavailable data, can support funding agency policies. Statistical tools are available for handling multiplicity. While in any specific instance there may be subtleties (related to "independence" among multiple hypotheses) about how best to apply them, statistical methods are readily available. Observational studies continue to stress available statistical theory and methodology. When there are many potential confounders, it is virtually certain that some of these will be unbalanced. (This can be true even in randomized trials.) Model-based approaches seem promising but are largely untested.

In summary, data used in decision-making should be available to anyone. Multiple testing questions are subtle; so anyone evaluating a claim should carefully look into how many questions were at issue and was there any correction for multiple testing. It would be very embarrassing if

observational environmental studies had a false claim rate of the 80% that is seen in observational medical studies. Observational studies are subject to bias; experts and courts have used the guidance that a risk ratio has to be larger than 2.0 before it is considered admissible for consideration as real.

Stan Young, [young@niss.org](mailto:young@niss.org), 919 685 9328, is the Assistant Director of Bioinformatics at the National Institute of Statistical Sciences. He is also an adjunct professor of statistics at three major research universities. He is the co-author of the book, Resampling-Based Multiple Testing, with Peter Westfall. Dr. Young works on the application of statistics to applied problems in medicine, biology, genetics, chemistry, toxicology, etc.