Michael Lavine and Joseph Horowitz

We would like to offer some comments on the definition and interpretation of P-values. To set the stage, consider an over-simplified multiple comparisons situation in which we test 100 hypotheses H_j, for $j = 1, \ldots, 100$. The jth hypothesis yields a z-score Z_j. Suppose it turns out that $|Z_{22}| = 3$ is the largest absolute value of the 100 z-scores.

A P-value is, to paraphrase the ASA statement, the probability under a specified model that a statistic would be more extreme than its observed value. Thus, to have P-values, we need models and statistics; each P-value pertains to a particular (model, statistic) pair. We focus our comments on just two of the many pairs we might consider:

Pair A ($H_{0,A}$: $\mu_{22} = 0$, statistic$_A$: Z_{22}) and

Pair B ($H_{0,B}$: $\mu_1 = \cdots = \mu_{100} = 0$, statistic$_B$: Z_{J^*}) where J^* is the maximizer of $|Z_j|$.

Each of these pairs has a P-value. In our experiment, under the obvious Normality assumption, $P_A = 2(1 - \Phi(3)) \approx .0027$, even if attention was focussed on Pair A only after the data were collected, whereas P_B cannot be calculated without further assumptions about the joint distribution of $Z_1, \ldots, Z_{100}$.

With this background, we make the following observations.

1. In multiple-comparison settings one often encounters the question Q_1: *Should P-values be adjusted?*, which sounds like a technical question about statistics, to be answered by statistical theory. But because P_A ($\approx .0027$) is already a valid P-value (for A) without adjustment, Q_1 puts the emphasis in the wrong place. Often a more useful question is Q_2: *Which pair, and therefore which P-value, should we care about, A or B?*, a question about the investigation, to be answered in collaboration with the investigator in the context of background knowledge.

   One might observe that A does not accurately represent the way the data were collected. That may be true, but there is nothing in the definition of P-value to say that H_0 must reflect the experimental design. One might argue that $Z_1, \ldots, Z_{100}$ ought to be modelled jointly, not separately. That may be true, but there is nothing in the definition of P-value to say that the model in the (model, statistic) pair must accurately reflect the distribution of the data. One might note that if we report P_A, some people will interpret it as a P-value for B. That may be true, but is the result of a misunderstanding and does not mean that P_A is not a valid P-value for A.

   As statisticians, we can point out the differences between A and B; we can help build models for the joint distribution of $Z_1, \ldots, Z_{100}$; we can explain the different distributions of Z_{22} and Z_{J^*}; and we can help researchers think about whether they should care about A or B. But where the ASA's

statement says *"[c]onducting multiple analyses of the data and reporting only those with certain p-values . . . renders the reported p-values essentially uninterpretable,"* we would say instead that results should be reported so that they are useful to readers interested in A, B, or any other hypothesis that might be of interest, and so that they help readers distinguish and decide between A and B.

2. The ASA's statement says *"Cherry-picking promising findings . . . leads to a spurious excess of statistically significant results."* But there are at least two points of view regarding spurious excess.

   (a) There are 100 individual hypotheses similar to $H_{0,A}$; they are $\mu_1 = 0, \ldots, \mu_{100} = 0$. If all 100 hypotheses are true, then about five of them will yield P-values less than about .05. There is no excess of small P-values or declarations of significance.

   (b) There is a single hypothesis $H_{0,B}$. If it is true and we calculate the 100 P-values pertaining to the 100 individual hypotheses then there is a large probability that one or more of them will be less than .05. There is an excess of small P-values and declarations of significance.

   It seems, to us, that the purported excess of small P-values in (b) is due to treating individual P-values of type A as though they are of type B. Whether there is truly a spurious excess depends on whether we care about pairs like A or like B.

3. Whatever is the joint distribution of $Z_1, \ldots, Z_{100}$, $P_B \geq P_A$. In fact, P_B is greater than or equal to each of the 100 P-values in (a) above. Assuming independence of $Z_1, \ldots, Z_{100}$ gives $P_B = 1-(.9973)^{100} \approx .24$ which, under the usual interpretation, means that the data are compatible with $H_{0,B}$. The same data also yield $P_A \approx .0027$, which means that the data are not compatible with $H_{0,A}$. But because $H_{0,B} \subset H_{0,A}$ — i.e. $H_{0,B} \Rightarrow H_{0,A}$ — those two inferences about compatibility are incompatible. That's a general phenomenon of P-values pointed out by Schervish (1996): if a parameter space Θ can be partitioned into null and alternative hypotheses in two ways such that, say, $H_0 \subset H_0'$, so, necessarily, $H_a' \equiv H_0'^c \subset H_a \equiv H_0^c$, then the P-value for H_0 may be larger than the P-value for H_0', even though logic dictates that the data must be at least as compatible with H_0' as with H_0. The incompatibility is inherent to P-values and cannot be resolved. Thus, P-values cannot be interpreted formally as evidence measures or, at least, the mapping between P-values and "evidence" varies according to circumstance. The ASA statement's Principle 1: *"P-values can indicate how incompatible the data are with a specified statistical model"* can be interpreted only informally, at best.

Bibliography

Schervish, Mark J. (1996). “P values: What they are and what they are not”. In: *The American Statistician* 50, pp. 203–206.