ate Editor, Academic Emergency Medicine

Key words. hunger; emergency department; workload; emergency physicians; public health.

Reference

1. Kersey MA, Beran MS, McGovern PG, Biros MH, Lurie N. The prevalence and effects of hunger in an emergency department patient population. Acad Emerg Med. 1999; 6:1109–14.

p < 0.05: Threshold for Decerebrate Genuflection

The **p** in p-values is a convenient statistical shorthand for **p**robability, measured on a continuum from 0 to 1. In classic *statistical* hypothesis testing, p-values are interpreted under the null hypothesis of no difference between the data sets being compared. The value assigned to p is the probability that, under conditions defined by the null hypothesis, a difference between the data sets at least as large as the one observed could have occurred by chance. The smaller the p-value, the lower the probability that such a finding represents a random event and, correspondingly, the stronger the argument for rejection of the null hypothesis.[1,2]

Utility of p-Values

p-values diminish the guesswork entailed in deciding whether observed findings are numerically stable. For example, if one-year survival of patients in a clinical trial randomized to treatment A (RxA) is 50%, and that of patients randomized to RxB is 75%, should we conclude that the two treatments are associated with different outcomes? Clearly, the question can't be answered without further information. If there were four patients randomized to each arm of the trial (N = 8 total), of whom two receiving RxA, and three receiving RxB, survived, we would immediately recognize that these proportions could easily have occurred by chance. From this we would conclude that the findings are therefore too numerically unstable to support any inferences about real differences between the two interventions. In contrast, if there were 400 patients randomized to each arm of the trial (N = 800 total), with the same 25% absolute difference in survival, we could confidently conclude that the findings are quite stable, and that this observed difference is unlikely to be a chance event. In neither instance would p-values be needed to assist in interpretation of these data. However, as the number of patients randomized is allowed to converge from the opposite extremes of 800 and 8 toward some middle ground, we would rapidly encounter the limits of our mathematical intuition. In this intermediate zone, where most data reside, the p-value serves as a statistical guidepost, providing a sense of whether the observed differences in mortality are likely to represent a treatment effect or chance event.

Role of p-Values in Classic Hypothesis Testing

p-values are an integral part of classic hypothesis testing, which proceeds in the following stepwise fashion.[1–3]

First, we choose a null hypothesis, stated as a proposition contrary to that which we hope to find. Thus, if we wish to identify a difference in a variable of interest between two groups, say age, the null hypothesis asserts that there is no age difference between the groups. That which we hope to prove—a difference in age—is known as the *alternative* hypothesis. The formal logic underlying this construct is proof by contradiction.[3] If a valid argument logically contradicts a stated premise, we conclude that the premise must be false. If only two outcomes are possible, rejection of the null implies acceptance of its alternative. Therefore, either we reject the null—and by definition accept the alternative hypothesis—or we fail to reject the null hypothesis, and therefore cannot embrace its logical alternative.

The second step in hypothesis testing entails selection of an appropriate test statistic. For detecting differences between groups with respect to a continuous variable such as age, the t-test is typically used.[4] If the data are non-normally distributed small samples, a nonparametric test such as the Wilcoxon-Mann-Whitney would be a more appropriate choice.[5]

Third, the probability distribution of the chosen test statistic is examined under the null hypothesis. For parametric tests, such as the t-test, the distribution is based upon a theoretical mathematical model requiring that the data configuration meet certain underlying assumptions.[4,6] For nonparametric tests, the distribution of the statistic is derived from combinations/permutations of the data itself without theoretical constraints.[1,5] With the widespread availability of p-values as an intrinsic component of the output from virtually all statistical software packages, this third step in classic hypothesis testing has become invisible. However, it remains conceptually important, since failure to choose a test statistic appropriate for the data set under examination can substantially distort results.

Fourth, we decide whether the p-value should be one- or two-tailed.[7] Although the two-tailed value is almost always chosen, it is entirely appropriate to use a one-tailed p-value if the investigator is concerned only with a unilateral difference.[3,7] For example, among patients with a disease carrying a 100% mortality, one would be interested only in improved survival among those randomized to the treatment arm of a clinical trial (since treatment cannot push the mortality beyond the 100% ceiling that already exists in the control group). The advantage of a one-tailed p is that fewer patients are required in the sample size calculation than for the traditional two-sided p.[2]

Fifth, one chooses an α level for p. This level becomes the threshold criterion below which the null hypothesis is rejected.[3] By convention, this value of α has been arbitrarily set at $p < 0.05$ for most of this century.[1] Allegedly based upon an offhand comment made by Sir Ronald Fisher (of the Fisher's exact test), this arbitrary cut-point has, for no apparent or rational reason, very nearly achieved the status of an immutable constant.[1]

Based upon the foregoing five steps, we arrive finally at one of two conclusions: Either 1) the p-value is below the chosen α level, the null hypothesis is rejected, the alternative hypothesis is accepted, and statistical significance is declared; or 2) the p-value exceeds or equals α, the null hypothesis is conceded, the alternative hypothesis is unsupported and the results are deemed statistically nonsignificant.[1,2,3,7]

Whatever the historical derivation of $p < 0.05$ might have been, blind adherence to a single, arbitrary, context-insensitive cut-point as the sole criterion for identification of statistical significance prompted Feinstein to characterize $p < 0.05$ as "the threshold for decerebrate genuflection" (Feinstein AR, personal communication, 1990). The problem with $p < 0.05$ lies not in the value itself, but rather in dichotomizing any distribution of continuous data. Referred to in diagnostic testing as the "single-cutoff trap,"[8] reduction of continuous data to binary categories not only markedly decreases the information contained in the native data set, but may also misrepresent its content.[8]

INTERPRETATION OF P-VALUES

Much of the difficulty in interpreting p-values arises from two sources. The first, which is closely linked to statistical α and β error, represents a failure to distinguish between clinical (or quantitative) importance and statistical significance. Because the p-value not only is determined by the clinical magnitude of a finding, but is also strongly influenced by sample size,[1,3,7] a marginal quantitative difference accompanied by a sufficiently large sample size can generate a statistically significant p-value that is clinically meaningless. Conversely, given a quantitatively important difference, a small or "underpowered" sample size can produce a statistically nonsignificant p-value.[1,3,7]

The second source of confusion in interpretation of p-values derives from a failure to recognize that these probabilities are not valid omnibus measures of the *strength* of evidence supporting a finding. This is because p-values, constrained as they are by classic hypothesis testing, can look only in the direction of the null hypothesis. Thus, the complement of the p-value $(1 - p)$ will not necessarily represent the probability that the alternative hypothesis is true or false. This is because the degree of truth or falsity of the alternative hypothesis depends not only upon the p-value associated with the null hypothesis, but also upon the independent prior probability of that specific alternative hypothesis's being true.[3]

MULTIPLE COMPARISONS

Multiple comparisons represent an important additional source of statistical false positives or α errors deserving separate consideration.[3,6,7] This problem is less a consequence of the nature of probability than another artifact of establishing an all-purpose threshold p-value for declaration of statistical significance. Each time a comparison is undertaken and deemed statistically significant at the $p < \alpha$ (0.05) level, there remains an α% (5%) probability that this represents a false-positive finding. Because the likelihood of obtaining a false-positive finding increases with each additional comparison, some systematic adjustment proportional to the number of tests performed is needed. Otherwise, obtaining a "statistically significant" p-value will simply be a matter of performing a sufficient number of statistical tests. Unlimited exploratory, data-driven significance testing of a data set, without adjustment for multiple comparisons, is commonly referred to as data "dredging."[1–3] Because of the high probability of chance identification of spurious findings, unless prospective validation of results on an independent cohort is planned, such an undertaking has limited scientific merit.

Although a large number of unplanned (*post-hoc*) tests seem more prone to generation of false-positive findings than a small number of planned (*a-priori*) ones, there is no general agreement on how one should decrement the value of α (p-value) to adjust for multiple comparisons. The simplest strategy, and one of the most commonly used, is the Bonferroni correction.[2,6] The threshold value for statistical significance, α (conventionally $p < 0.05$), is simply divided by the number of comparisons performed. The dividend then becomes the new threshold for declaring any single observation to be statistically significant. Thus, for five comparisons, the conventional threshold for statistical significance drops to $0.05/5 = 0.01$, for ten comparisons to 0.005, etc.

Many argue that this strategy is too conservative and that overadjustment for α error unduly increases the likelihood of missing findings that may be important (increased false-negative results or β error).[1,2,6] An alternative, less conservative approach can be devised from estimation of the probability of false positives, p(FP), under conditions where p is set at $\alpha = 0.05$. Thus, $p(FP) = (1 - 0.95^n)$, where n is the number of comparisons. Skipping the algebra, the generalized formula for adjusting α_{adj} for n comparisons = $\{[\alpha^2]/[1 - (1 - \alpha)^n]\}$, in which α is traditionally set at 0.05.[4]

For small numbers of comparisons, both adjustments provide similar results. However, as the number of tests increases, the two adjustments diverge slightly, e.g., for 20 tests, the Bonferroni-adjusted p-value would be 0.0025, vs 0.004 for the alternative method.[2,6]

Significant Ambiguity

"Significance" is a term containing such ambiguity that some authors have suggested it be expunged from the lexicon of both science and statistics.[9] Unfortunately, it is deeply entrenched in medicine, and for the foreseeable future, we appear to be stuck with it. The simplest way to give some meaning back to the term "significant" is to use it only with such conditional modifiers as "clinical" or "statistical."[1,3]

As noted earlier, p-values provide assistance only in determination of *statistical* significance. Used judiciously, i.e., nondichotomously to avoid the single-cutoff trap,[8] and with appropriate attention to sample size[3] and multiple comparisons,[2,6] a probabilistic assessment of the role of chance in producing results can be obtained. However, without a concomitant and independent determination of the quantitative or *clinical* significance of a given finding, statistical significance, considered in isolation, is at best ambiguous, and at worst misleading.[1]

An Alternative to p-Values and Significance Testing

Many biomedical journals now require authors to express their findings using interval estimation in preference to significance testing, i.e., confidence intervals (CIs) rather than (or in addition to) p-values.[10] This requirement is based upon the superior informational content and configuration of CIs, i.e., provision not only of statistical information but, more importantly, quantitative/clinical information in an economical, explicit, and precise format.

Defined loosely, *n%* CIs (conventionally 95% confidence intervals) indicate that we can be about *n%* certain that the "true" result of a methodologically valid investigation lies within the limits that bound this interval.[10] Several of the specific advantages of CIs over p-values include the following[1,3]:

As noted above, one of the problems with freestanding p-values is an inability to determine the extent to which their magnitude is driven by observed quantitative differences vs sample size.[1,4] The CI, unlike the p-value, is able to disentangle these two constituent parts by displaying the observed quantitative difference as the point estimate.[10] This estimate—which represents the best, mathematically unbiased index of the difference between two groups under comparison—stands alone and can be viewed separately from any statistical information that might otherwise obscure it. The sample size, rather than being buried within the p-value, is inversely proportional to the width of the CI, and can be determined by direct inspection of the distance between its boundaries.[1]

The relationship of confidence limits to the null point (either zero for means and proportions, or unity for risks and ratios) provides additional information not available from p-values. The focus, however, should not be on whether the CI embraces the null point, since that would mean falling into the same single-cutoff trap of $p < 0.05$,[8] but rather on the "tilt" of the interval.[3,10] Using the null point as a fulcrum, examination of the direction and extent to which the interval is "leaning" can provide a nearly graphical display of study findings.[10] For example, a mean difference of 23 mm between two groups of patients on a visual analog scale (VAS), favoring RxB over RxA, accompanied by a 95% CI of (−2 mm to 47 mm) suggests that, with a few more patients, there probably would have been a statistically significant difference to accompany the quantitatively significant difference of 23 mm. From this alone, we might reasonably conclude that, pending a larger study or further information in the form of a meta-analysis, RxB ought to be administered to patients in preference to RxA. In contrast, if we were using classic hypothesis testing, and only p-values were displayed, we might easily reach the "β-erroneous" conclusion that RxA and RxB are equianalgesic, if only "$p > 0.05$" is reported.[3,10]

Similarly, if the relative risk of admission associated with a very expensive intervention RxA, compared with an inexpensive standard of care RxB, is reported as 0.98 (95% CI = 0.97 to 0.99), we might decide that the cost of RxA was not worth the very modest reduction in admission rate —in spite of its exclusion of the null point of 1 for relative risks. With a sample large enough to drive the p-value for a relative risk down to, say $p = 0.001$ (which is entirely consistent with an N of sufficient size to generate such a precise CI), examination of this p-value in isolation might well lead us to conclude (this time "α-erroneously") that the investment of resources in RxA would be cost-effective.

Because interval estimation and significance testing operate under the same constraints of probability, adjustment of CIs for multiple comparisons, analogous to that of p-values, seems appropriate.[10] Although there is no consensus on a methodology for this, direct application of the Bonferroni concept to CI has the appeal of being straightforward.[2,6] Similar to the proportionate reduction in p-values according to α/n for n comparisons, increasing the precision of the CI proportionate to $[1 - (\alpha/n)]\%$ is a reasonable, though unproved, strategy for reducing false-positive findings. Thus, if an investigator performs five unplanned comparisons on a data set, the traditional $(1 - \alpha)\%$ or 95% confidence intervals typically used for displaying findings should perhaps be recalculated as $[1 - (\alpha/5)]\% = (1 - [0.01]) = 99\%$ CI.

Finally, it is worth remembering that CIs, though substantially more durable than p-values, are not immune to abuse. Indeed, a 95% CI represents nothing more than a level of "confidence" derived directly from the complement of $p < 0.05$ through subtraction, i.e., $(1 - 0.05) = 0.95 = 95\%$. Therefore, to require that a CI exclude the null as a prerequisite to careful inspection of its point estimate, confidence limits, and width is not "significantly different" from "decerebrate genuflection" at a threshold value of $p < 0.05$.—E. JOHN GALLAGHER, MD, *Department of Emergency Medicine, Albert Einstein College of Medicine, Bronx, NY*

Key words. p-value; p < 0.05; significance; statistics; probability; hypothesis testing.

References

1. Feinstein AR. Clinical Epideminology. The Architecture of Clinical Research. Philadelphia: W. B. Saunders, 1985, pp 130–69.

2. Lang TA, Secic M. How to Report Statistics in Medicine. Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia: American College of Physicians, 1997, pp 65–92.

3. Bailar JC III, Mosteller F. Medical Uses of Statistics (2nd ed). Boston: NEJM Books, 1992, pp 181–200.

4. Glantz SA. Primer of Biostatistics (3rd ed). New York: McGraw-Hill, 1992, pp 67–104.

5. Krauth J. Distribution-free Statistics: An Application-oriented Approach. Amsterdam, The Netherlands: Elsevier, pp 48–57.

6. Dawson-Saunders B, Trappe RG. Basic and Clinical Biostatistics (2nd ed). Norwalk, CT: Appleton & Lange, 1994, pp 95–142.

7. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. Biostatistics in Clinical Medicine (2nd ed). New York: Macmillan, 1987, pp 151–67.

8. Young MJ, Fried LS, Eisenberg JM, et al. The single-cutoff trap: implications for Bayesian analysis of stress electrocardiograms. Med Decis Making. 1989; 9:176–80.

9. Significance of significant [editorial]. N Engl J Med. 1968; 278:1232–3.

10. Gardner MJ, Altman DG. Statistics with Confidence. London: BMJ Publishing, pp 3–26.