

The “rule and merit” method: An alternative to tests of statistical significance

Eric Marder
Eric Marder Associates, Inc.

Although tests of statistical significance have been the subject of much criticism, they continue to be a major tool in the analysis and presentation of experimental findings. This article presents a probability problem to demonstrate a basic inconsistency in the way statistical significance tests are used when deciding among actions, and proposes a simple alternative method for dealing with the uncertainty inherent in the outcome of experiments.

Introduction

Through the years, many investigators from different disciplines have called attention to conceptual problems with tests of statistical significance (Berkson 1942; Rozeboom 1960; Kaiser 1960; Grant 1962; Clark 1963; Carver 1978; Falk 1986; Harris 1991). Much of this history is covered in three recent articles (Falk & Greenbaum 1995; Goodman 1999a; Johnson 1999). Concurrently, various remedies have been proposed, but these, in turn, have been the subject of criticism. Confidence intervals, the reporting of p-values, and three-tail tests perpetuate the logical problems of significance tests in different form. Focusing on effect and relevance sidesteps the need to address the role of chance. Bayesian approaches incorporate subjective elements. And relying on replications places excessive and impractical demands on the investigation. On balance, neither the criticisms nor the proposed remedies have made a dent in long-established habits of thought. Most courses and textbooks of statistics routinely present tests of significance as established truth, and tests of significance continue to be a pervasive tool of experimental research. It is difficult to find empirical studies in marketing research, the social sciences, psychology, biology, medicine, and some physical sciences that do not report tests of significance, and the use of these tests continues to be defended (Fleiss 1986; Chow 1988; Frick 1996; Harris 1997; Abelson 1997).

For the most part, the criticisms of significance testing have revolved around two issues: the failure to take prior

knowledge into account in drawing substantive conclusions (Goodman 1999b) and the erroneous belief that a “significant” result justifies rejecting the null hypothesis, stemming from a confusion of $P(R|H_0)$, the conditional probability of a result, given the null hypothesis, with $P(H_0|R)$ the conditional probability of the null hypothesis, given the result (Falk & Greenbaum 1995).

This article questions significance testing from still another perspective, charging it with throwing away information in the special case of studies conducted to aid some end-user (a “client”) in making a decision. The critique is mounted without prejudice to the debate between the frequentists and the Bayesians. It accepts at face value the frequentist logic of tests of significance and uses that logic itself to identify flaws in these tests for the purpose at hand. In particular, it presents a special probability problem to demonstrate that a test of statistical significance can be not merely inappropriate, but just plain wrong. In its place, it proposes the “rule and merit” method, a simple alternative for dealing with the uncertainty inherent in the outcome of empirical experiments.

Statistical significance

Take the simplest of cases: a test to evaluate the effectiveness of two treatments, two product formulations, two drugs, or say two ads. The study is a rigorous experiment. 200 respondents received Ad A. Another, randomly equivalent group of 200 respondents, received Ad B. In group A, 57% of the respondents subsequently bought the test brand; in group B, 58% bought it. What have we learned about the effectiveness of Ads A and B?

What conclusions are appropriate? Does it matter which ad is run?

In addressing these questions, researchers usually resort to a test of statistical significance. An experienced researcher will recognize by inspection that the one percent difference cannot possibly be statistically significant — not at the 99% level, not at the 95% level, not even at the 80% level. If required to perform the computation formally, he may use the normal distribution as a reasonable approximation of the binomial, and obtain

$$z = \frac{X_A - X_B}{\sqrt{\frac{P_A Q_A}{n_A} + \frac{P_B Q_B}{n_B}}} = .20$$

which is far below the 1.96 necessary for significance at the 95% level, and even far below the 1.29 necessary for significance at the 80% level. Accordingly, he accepts the null hypothesis and reports that the result was “not statistically significant at the 80% level”.

When asked what this means, he offers this definition: “If there had been no difference whatever between the ads — indeed if both groups had been exposed to the same ad — we would have gotten, by chance alone, a difference equal to or larger than the one observed, more often than 20% of the time.”

In this reply, the researcher has carefully phrased his statement to render it technically correct. But he has employed a semantic sleight of hand. Instead of addressing the question on his client’s mind, he has deftly changed the subject, and has answered a different question — one that happens to be irrelevant because the client has no interest in what would have happened if there had been no difference between the ads. What she wants to know is, having gotten what she got, in particular 57% buying after exposure to Ad A and 58% buying after exposure to Ad B, what, if anything, has she learned and what should she do. Accordingly, she demands clarification.

In an effort to explain his original, correct but involuted statement in simpler terms, the researcher may resort to saying, “We found no difference between the ads. Toss a coin. Use either one.” If he explicitly offers this “explanation” or implicitly allows his less sophisticated client

to infer it, he has crossed a delicate but critical line from being merely irrelevant to being downright wrong. The objective fact is that the two ads did not produce the same result. 58 is not the same as 57. And though the difference is very small, it is not zero, and contains some information about the ads, information the client may decide to set aside, but not before it has been properly reported to her.

How then, should the result of this test be reported? The researcher must be willing to be both categorical and responsible. To begin with, he must report the result at face value: “Ad A produced 57% buying. Ad B produced 58%. The test showed that Ad B is better. Use Ad B.” This is the categorical part of the answer. But he must also be responsible and add this caution: “Of course, it is possible that my conclusion is wrong. This is always so, no matter how large the result may happen to be. For this reason, I routinely add a warning. In this particular case, there is a 42% probability that the conclusion I have just reported to you is wrong. You may not like this, but if you used the other ad, your probability of being wrong would be 58%, which is worse. Whether the result you obtained was or was not significant at any pre-specified level is irrelevant. There is nothing intrinsically noble about the significance levels ordinarily used. Ninety-nine percent, 95%, or 90% are no better than any other levels you could have arbitrarily chosen instead, for example 93.8%, or 88.5%, or 62.7%. If you have information from other sources, you may certainly use Ad A on those other grounds, but if you want to base your decision on the results of this test, you must use Ad B.”

It may appear that this amounts to eliminating statistics. To some extent, this is so. The statistics, however, are not eliminated altogether. They are merely assigned their proper function. As long as circumstances require us to choose between two actions (A or B), and no other data are available, and the decision cannot be deferred, statistics are indeed irrelevant. The only thing that matters is which option received the larger number. But if we don’t have to act immediately and want to decide whether to defer acting until additional data become available, statistics can help by providing an estimate of the risk that the conclusion of the current study is wrong. Before going on to discuss how this risk should be computed, we examine

the fundamental logic of the computation. This logic is quite simple, but may be troubling to someone who learned statistics from one of the many textbooks that present tests of statistical significance as established “science”. Perhaps the issue can be thrown into sharp relief by analyzing a dramatic probability problem.

A probability problem

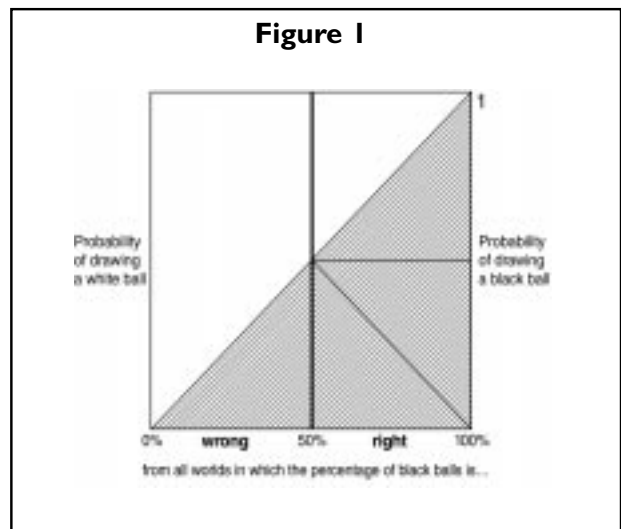
An urn contains a very large number, say 10^{10} , of black balls and white balls. You know nothing about the urn or the balls in it. Assume that the urn is as likely to contain any particular percentage of black balls as any other. You must estimate whether the urn contains more black balls or more white balls. In a medical context, this may be a matter of life and death. If your estimate is right, you live; if it is wrong, you die. In the absence of further information, you are reduced to tossing a coin, and your probability of coming out alive is 50%. But you are allowed to draw a random sample from the urn before making your estimate. Unfortunately, you are limited to a small sample size, in fact to $n=1$. You draw and obtain a black ball. The question is whether this information changes your probability of surviving. More specifically, does it change this probability: (a) not at all, (b) by an infinitesimal amount, or (c) by a substantial amount?

In attempting to deal with this question, you can use a test of statistical significance or you can use what I will call the “rule”. If you use a test of statistical significance, you formulate the null hypothesis that the percentage of black balls in the urn is equal to the percentage of white balls, $H_0: \mu_b = \mu_w = 50\%$. The test of significance requires that you determine the probability of having obtained by chance alone a result as different from 50% in either direction as the one you obtained. This probability is 100%. The result cannot therefore be “statistically significant” under any circumstances, no matter what level of significance you use, and you must conclude that there is “no difference” between black and white. The sample of one has not improved your chances. You may toss a coin. Your probability of surviving remains 50%.

Alternatively, you can use the rule. This consists of making a categorical decision. If the ball that was drawn was black, you estimate that the urn contains more black

balls. If it was white, you estimate that the urn contains more white balls. If you use this rule, your probability of surviving increases to 75%. In this case, the test of significance has not merely been irrelevant, but actually wrong. This result is so counter-intuitive that people usually refuse to believe it. But there is a simple proof for it.

Imagine an infinite number of urns (worlds), each containing an infinite number of balls. Every point on the x-axis of Figure I represents one such world. The left-most point represents the world in which all balls are white. The right-most point represents the world in which all balls are black. The points in between represent all possible worlds, ranging continuously from 0% to 100% black balls. The height of the diagonal line represents the relative probability of obtaining a black ball from a world, and the shaded area represents all the black balls in all the worlds. You know that the single black ball that was drawn must have come from one of the worlds on the x-axis, but you don't know from which one. If you estimate in accordance with the “rule”, you will be right if the ball came from any of the worlds to the right of the 50% point; and you will be wrong if it came from any of the worlds to the left of the 50% point. But, as the diagram indicates, the area under the diagonal to the right of the 50% point is three times as large as the area to the left of it. Accordingly, the odds are 3:1 that your estimate was right rather than wrong. Knowing the colour of this one ball has improved your chances of surviving from 50% to 75%.



This analysis does not actually use a different logic than that used in a test of significance. But it produces a different conclusion because it generalizes this logic. Instead of considering only one of an infinite number of hypothetical possibilities (the null case), it considers all hypothetical possibilities simultaneously. Instead of focusing only on the probability of having obtained the observed result from a true value of zero, it considers the probability of having obtained the observed result from any theoretically possible true value.

Computing the probability that the rule yields the right answer

We now return to the experiment concerning the effectiveness of Ads A and B, in which 57% bought the brand after exposure to Ad A and 58% bought it after exposure to Ad B. Applying the rule, we conclude that Ad B is the better ad. What is the probability that this conclusion is right? The computation of this probability, which is called the “merit” of the conclusion, resembles a one-tail test of significance. Its meaning, however, is entirely different. For the sake of clarity, we perform the computation before presenting its rationale.

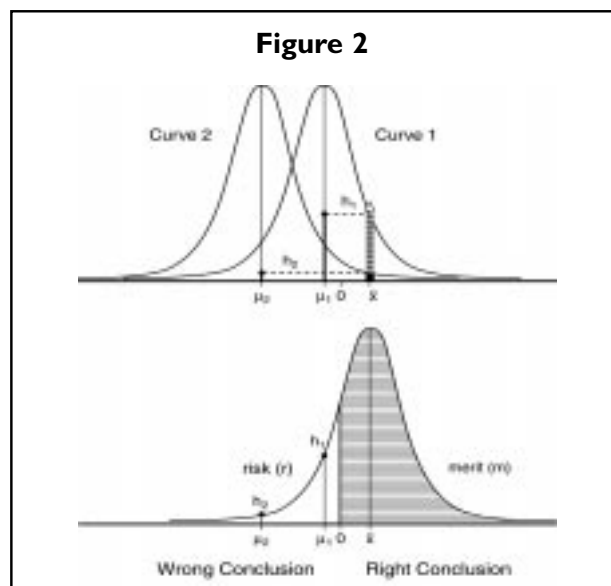
The difference in buying associated with the two ads is $58\% - 57\% = 1\%$ (.01). The standard error of this difference is

$$\sqrt{\frac{(.58)(.42)}{200} + \frac{(.57)(.43)}{200}} = .049$$

Dividing the observed difference by its standard error, gives a $z = .01/.049 = .20$. Looking up $z = .20$ in a table of the cumulative normal distribution, we find a probability of .579. This is the merit of the conclusion, $m = 58\%$.

The logic on which this computation is based is analogous to the logic we used in solving the problem of the black balls and white balls. Let the true difference between the ads be μ_1 , and let the observed difference be \bar{x} . The top half of Figure 2 shows the sampling distribution of \bar{x} which is a normal distribution with mean μ_1 and standard deviation $\sigma_{\bar{x}}$. The height of the normal ordinate of this curve (curve 1) at \bar{x} , is h_1 , and Δ is an infinitesimal interval on the x-axis. The probability (p_1) of having obtained an observed difference in the range

$\bar{x} \pm \frac{\Delta}{2}$, when the true difference is μ_1 , is therefore $p_1 = h_1 \Delta$. We transpose this ordinate and plot it over μ_1 . We next consider the case when the true difference is μ_2 , represented by curve 2. The probability (p_2) of having obtained an observed difference in the range $\bar{x} \pm \frac{\Delta}{2}$ when the true difference is μ_2 , is $p_2 = h_2 \Delta$. Again we transpose the ordinate. This time we plot it over μ_2 . If we repeat this process for all points on the axis, we generate the curve shown on the bottom half of Figure 2. This is just the sampling distribution of \bar{x} , transposed. In a test of the null hypothesis, this curve would have been built around 0. It is now built around the observed mean \bar{x} — the same familiar curve in a different place with a different meaning. The area under the curve is still equal to 1, but it now represents the probability of having obtained the observed difference \bar{x} from any hypothetical true difference. The vertical line at 0 divides this area into two segments. The segment to the left of 0 represents the probability of having obtained the observed value \bar{x} from some true value $\mu_i < 0$. If that was the case, the reported conclusion that Ad B produced a larger effect than Ad A was wrong. This probability is called the “risk” (r). The segment to the right of 0 represents the probability of having obtained the observed value \bar{x} from some true value $\mu_j > 0$. If that was the case, the reported conclusion was right. This probability is called the “merit” (m) of the conclusion, $m = 1 - r$, and is shaded on the bottom half of Figure 2.



Generalizations and conclusion

To avoid digressions, we have limited our discussion to experiments designed to guide action. The rule of choosing the treatment that got the biggest score is obviously relevant in those cases. But what about studies that don't appear to fit this description? Suppose a study is undertaken solely for the purpose of increasing "understanding" of a subject. On closer scrutiny, even such a study must lead to action. Minimally, it must affect what we believe about the subject. But beliefs can be relevant only if they have consequences, either immediate or eventual. They may, for example, lead us to accept one theory over another, or to continue an investigation, or to abandon it. Limiting the analysis to experiments designed to guide action may, therefore, not have been all that much of a limitation. From a broad enough perspective, most, if not all, experiments meet this standard, though the alternatives are not generally equal-cost alternatives.

For simplicity's sake, we have so far dealt exclusively with equal-cost alternatives (e.g., Ad A vs. Ad B). It is easy, however, to extend the analysis to cases in which the costs associated with the different treatments are unequal. If Ad A had been a black-and-white ad and B a colour ad, some new "zero" point could have been computed. We might have determined, for example, that it would be necessary for Ad B (the colour ad) to outperform Ad A (the black-and-white ad) by three percentage points to offset Ad B's higher cost. The vertical line which was set at 0 would have been shifted to +3, and we would have proceeded precisely as before. In particular, we would have added 3% to Ad A's 57%. Ad A's 60% would then have beaten Ad B's 58%.

One particular misuse of significance tests is widely recognized: the confusion between statistical and substantive significance. Suppose an experiment demonstrates that taking a medication for 20 years increases life expectancy by .1 years. The medication has only minor side effects, in particular, occasional nausea. The sample is large, so the finding is statistically significant at the 99.9% level. This statistical significance, however, is irrelevant until we have considered the substantive significance of the finding which necessarily involves cost-benefit considerations. How much does the end-user of the information (the client) value a .1 year increase in life

expectancy? Quite apart from financial aspects, what inconvenience and discomfort is she willing to accept to achieve it?

From the rule and merit perspective, we begin by requiring the client to specify her "break-even" point. A particular client may be willing to tolerate the medication's side effects in exchange for an increase in life-expectancy of 1.5 years. Given this information, the zero-point is set at 1.5, and we proceed as before. Our recommendation now depends solely on whether the experimental test-control difference turns out to be larger or smaller than 1.5. If larger, we recommend that she take the medication; if smaller, that she not take it. In either case, the conclusion is accompanied by a report of its merit.

Since different individuals must be expected to have different break-even points, a report may consist of three columns. Column 1 lists break-even points, (.5, 1, 1.5, 2, 2.5 ...); column 2 lists the yes-no recommendation (take the medication or don't take it), and column 3 shows the merit of the recommendation. An end-user can enter the table with her own break-even point and obtain the recommendation and its merit.

More generally, both in-going desires and in-going beliefs can be accommodated. If the client has a strong desire to use Ad A, she may be unwilling to switch, unless Ad B turns out to be at least five share points better than Ad A. In that case, the break-even point can be shifted to 5, and we proceed as before. Alternatively, she may have a strong belief that Ad A is better and may be unwilling to reconsider the issue unless the merit of the contrary conclusion is at least 99%. In that case, the merit of the conclusion will determine whether she accepts the conclusion.

The rule and merit method can be generalized to studies of more than two treatments. Assume that there are K treatments, and that treatment Z received the largest score. If we recommend treatment Z , our recommendation will be right if that treatment is actually better than each of the alternatives. If m_{zi} is the probability that treatment Z is better than treatment i , then m_z (the probability that treatment Z is better than all of the treatments) is:

$$m_z = \prod_{i=1}^{k-1} m_{zi}$$

If, for example, the ad study had included a third ad, and we had observed 57%, 60%, and 63% buying for Ads A, B and C, respectively, we would have reported that Ad C was the best ad, and that the merit of that conclusion was $m_z = m_{zA} m_{zB} = (73\%)(89\%) = 65\%$.

Merit has been defined as the probability of being right when reporting that the treatment that received the largest score in a test is the best treatment. Superficially, this may remind one of the “p-value” of a result. But the p-value in a test of significance and merit are quite different. A p-value is the probability of having obtained an observed or larger result by chance under the null hypothesis. Merit is the probability that the rule used to generate the conclusion has yielded the right conclusion.

We have called attention to the fact that the computation of merit is numerically similar to a one-tail test of significance. But this similarity is numerical rather than substantive. A one-tail test of significance is just what it is called, a test of significance. It tests a null hypothesis at some pre-designated significance level. Its output is an on-off statement, reporting that the observed result is either “significant” or “not significant” at some specified significance level. When we compute the merit of a conclusion, on the other hand, the arbitrary significance level is gone. The absurd idea that a result that has a t of 1.97 is “significant”, while one that has a t that is a trifle smaller, say $t = 1.95$, is “not significant”, is gone. In its place, we have a simple rule, accompanied by a simple statistical statement.

Unlike a test of significance, merit has no bearing on the action to be taken, except when it is possible to defer that action until additional data become available, which happens rarely in practice. If, based on the outcome of a five-year trial, you must decide which of two drugs to take, the only rational course of action is to choose the drug that produced the larger number of cures. If, based on the results of a three-month ad test, you must decide which of two ads to run, the only rational course of action is to run the ad that produced more sales. You need neither a statistical test nor a computation of merit. The rule tells you what to do. But you can look at merit to measure the extent to which you are at risk when you act on the findings of the study at hand.

The conceptual difference between a test of statistical significance and the rule and merit method can be summarized as follows. Both methods offer a strategy for dealing with the uncertainty that is inevitably inherent in empirical experiments. But they pose and answer different questions. A test of statistical significance computes the probability of having obtained by chance alone a result comparable to that obtained in the experiment, and compares that probability to a yardstick known as the “level of significance”, often arbitrarily set at 95%. The rule and merit method takes the directional outcome of the test at face value, reports it as such, and computes the probability that the reported conclusion is right.

References

- Abelson, R.P. (1997) On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 12–4.
- Berkson, J. (1942) Tests of significance considered as evidence. *Journal of the American Statistical Association*, 37: 325–35.
- Carver, R.P. (1978) The case against statistical significance testing. *Harvard Educational Review*, 48: 378–99.
- Chow, S.L. (1988) Significance test or effect size? *Psychological Bulletin*: 105–10.
- Clark, C.A. (1963) Hypothesis testing in relation to statistical methodology. *Review of Educational Research*, 33: 455–73.
- Falk, R. (1986) Misconceptions of statistical significance. *Journal of Structural Learning*, 9: 83–96.
- Falk, R. & C.W. Greenbaum (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, 5: 75–98.
- Fleiss, J.L. (1986) Significance tests have a role in epidemiologic research: Reactions to A. M. Walker. *Different Views*: 559–60.
- Frick, R.W. (1996) The appropriate use of null hypothesis testing. *Psychological Methods*: 379–90.

-
- Goodman, S.N. (1999a) Toward evidence-based medical statistics. I: The p value fallacy. *Annals of Internal Medicine*, 130: 995–1004.
- (1999b) Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130: 1005–13.
- Grant, D.A. (1962) Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 69: 54–61.
- Harris, M.J. (1991) Significance tests are not enough: The role of effect-size estimation in theory corroboration. *Theory & Psychology*, 1: 375–82.
- Harris, R.J. (1997) Significance tests have their place. *Psychological Science*: 8–11.
- Johnson, D.H. (1999) The insignificance of statistical significance testing. *Journal of Wildlife Management*: 763–72.
- Kaiser, H.F. (1960) Directional statistical decision. *Psychological Review*, 67: 160–7.
- Rozeboom, W.W. (1960) The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57: 416–28.

About the author

Eric Marder is chairman of Eric Marder Associates, Inc., the firm he founded in 1960.