tea-Tests

by Peter Killeen, Arizona State University

Few subjects in psychology elicit greater fear among students, and greater ambivalence among faculty, than statistical inference. It's difficult enough to get the calculations straight; then deciding what you can infer from the printout is like reading tea leaves in a room full of tasseographers: Whatever one concludes, another will gainsay; and a deep sense of hocus pocus pervades the whole affair. Psychologists can avoid tea leaves, but, alas, they can't avoid *t*-tests.

Last semester a student dropped by my office to discuss her data. She had failed to replicate a reliable result from the literature. Despite a healthy effect size, the small number of subjects kept her *p*-value above the magic point-oh-five. I explained that her results actually lent some support to the original claim, as her relatively large effect was in the correct direction. She asked "How much support?" and I held my fingers apart a little bit.

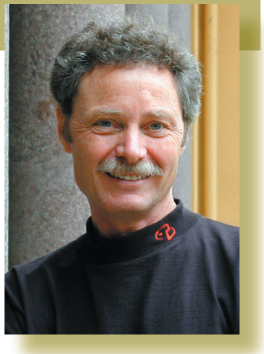
Forthwith she set about collecting reams of data and came back with a highly significant *p*-value, pleased to be able to reject the null hypothesis. I explained that she couldn't do that. All she was permitted to do was to act surprised at the deviation of the data from what was expected under the null hypothesis. For her $p < .01$, in fact, she was entitled to act quite surprised. She acted quite surprised.

"If I can't use statistics to draw conclusions about my hypothesis" she sniffed between stifled sobs, "then why do you teach all those statistics classes"? I explained that placebos can be very effective; but only if we believe in them. Now, as a behaviorist, I know that it takes rats only a couple of trials learn to avoid situations of pain or frustration. Students are smarter. She left.

Nickerson (2000) provides a none-too-brief breviary of the many ways in which null hypothesis statistical tests (NHST) are misunderstood. They are misunderstood both because they involve inverse inference, a problematic endeavor, and also because they are often mischaracterized in widely used texts (Cohen, 1994). Nickerson's authoritative sixty pages may lead readers to suspect that there is something fundamentally wrong with an inferential system that text-book writers can't get right. If the probability is much less than .05 that NHST will ever permit conclusions concerning hypotheses, shouldn't we do more than act surprised? Shouldn't we reject NHST?

Inverse Inference Given a fair coin, what is the probability of 8 heads in ten flips? That's direct inference. It is straightforward to compute because its downhill, from a stipulated population parameter ($p(H) = .5$) to a sample statistic. But now consider a coin that landed heads in 8 out of 10 flips. What is the probability that it is fair? That's inverse inference, and it is complicated because it is uphill, from a measured statistic to a population parameter. Our answer must depend in part on whether the coin came from our pocket, or from that of a guy trying to make a bar-room bet with us; just how we flip it, and so on, and on.

Such considerations are called *priors*, or *conditionals*, or *givens*. If I tell you that it's a fair coin to start, those priors are all taken care of, assumed, "given" by assertion. In the real world, however, such assumptions eventually need justification, and that justification involves inverse inference: You need to go up before you can go down.



Peter Killeen

Fisher-Neyman-Pearson statistics—what most of us use most of the time—provide optimal estimates of the probability of observing some statistic given an assumption, hypothesis, or parameter (unbiased coins, null hypothesis, $p = .5$). Call those probability estimates $p(d|a)$, with *d* the data, and the conditional $|a$ the given assumptions. Predicting *d* given *a*, $p(d|a)$, is simple direct inference. Going the inverse direction, to the probability of an assumption given the data, $p(a|d)$, is possible, but only with yet more assumptions. Bayes showed that we can make the conversion if we have estimates of the baserates: the prior probability of the assumption (that the coin was fair to start) and the prior probability of the data (the probability of 8 heads in 10 flips of coins in general).

Neither of these priors is easy to establish. But without them, as Fisher warned, "Such a test of significance does not authorize us to make any statement about the hypothesis in question in terms of mathematical probability" (Fisher, 1959, p. 35). That's why my student couldn't legitimately reject the null hypothesis, given her data. I patiently explained this to her when she finally stuck her head back in. She suggested that the prior probability of her hypothesis was close to 1.0, but when I winced, she left again, before I could explain that she also needed the prior probability of her data. I haven't seen her lately.

Prior Priors Based on the work of Reverend Bayes and Pierre-Simon Laplace (who attended a Benedictine priory school), modern Bayesians have attempted to provide the necessary prior probability distributions (see, e.g., Lee and Wagenmakers, 2005). Because we often have little or no information about the prior probability of a hypothesis, the problem of how to express ignorance mathematically must be solved. Bayesians have designed machinery that incorporates all the information that we do have about the priors, and are otherwise mute.

But critics read their lips, arguing that there is no way that they can be dumb enough. No sooner had Laplace harnessed Bayes' theorem for scientific analysis than George Boole cautioned: "When the defect of data is supplied by hypothesis [about the prior probabilities], the solution will, in general, vary with the nature of the hypothesis assumed; … I hope that a question, second to none other in the theory of probabilities in importance, will receive the careful attention it deserves" (Boole, 1854, as cited by Fisher, 1936, p. 248).

Despite its importance, and despite the careful attention it has received, there is no agreement on the answer. Many have attempted to untie this Gordian knot; most famously Fisher with his patient but inconclusive work on "fiducial probabilities". Many have cut the knot, but still couldn't get the old cart to move. Others have turned their backs on the antiquity, fast in its temple, and found a different wagon to ride.

A Different Wagon We can evaluate research claims much more directly by giving up any attempt to determine parameters or to reject hypotheses. "Sure," you smile, "give up our goals and achieving them is no longer a problem. Isn't the whole purpose of research to either prove things—or, after Popper, to disprove them?" Whose goals? Rejecting null hypotheses has been a sirens' call that has seduced too many scientists, to their delusion and their field's discomfiture.

Think of great advances in science, and few cases of NHST come to mind. Pasteur did not reject the null hypothesis that life can start spontaneously: He found maggots when the lids were off and not when they were on. Could he have done a *t*-test, would it have strengthened his claim? Darwin did not reject the null hypothesis of speciation without variation and selection; nor is it clear he ever could have. Watson and Crick did not reject the Null Helix Hypothesis. Skinner did not train dogs to jump through hoops significantly more often than chance.

Medical trials measure relative risk reduction; if negligible, the procedure is not pursued, whether or not the improvement is significant. When medical researchers take traditional statistical inference too seriously, they are as chagrined by its results as we (Ioannidis, 2005). "Proof" originally meant a test that provides evidence concerning a claim. All that data provide is evidence. There are better ways to use that evidence than in doomed attempts to prove or disprove hypotheses. We can use it to predict whether our results will replicate.

Replication Statisticians and scientists alike embrace replicability as an inferential goal. As Cohen (1994) said, "Given the problems of statistical induction, we must finally rely, as have the older sciences, on replicability" (p. 1001). Predicting replicability is easier than asserting or denying the truth of hypotheses. Getting on board this wagon is also easy, because we need merely rebadge some of the basic statistics that we already know. There are two steps to the process: Determine the sampling distribution of replicates, and then define what we want "replication" to mean.

1. Consider the left bell curve in the top of Figure 1. It is the sampling distribution of a statistic under the null hypothesis. Randomly select 5 Lipton® tea bags from a box, and weigh each. Do the same with 5 Salada® tea bags. Plot the difference in average weights on the *x*-axis. Repeat this many times and the histogram will look like the top right curve. Such sampling distributions form the basis of most inferential tests we use. The mean of a sample, *M*, can be predicted from (or serve as an estimate of) the mean of the population μ. The variance of the sample, s^2, can be used to estimate the variance of the population, σ^2. Sampling distributions are often normally distributed with mean m and variance σ^2/n. They are used to predict how often a statistic such as a mean, a difference of means $(M_E - M_C)$ or an effect size (mean difference divided by standard deviation: $d_1 = (M_E - M_C)/s)$, will take a particular value.

NHST typically sets the expected value of these statistics to zero (e.g. H_0: $d_1 = (\mu_E - \mu_C)/\sigma = 0$), or no real difference in weight of tea bags in our experiment, as in the left curve in Figure 1. If experimental and control samples were chosen from the same population this has to be true, because the population has a single mean, μ. But our measurement or experiment may have de facto created two populations with different means. Or we may find that one dimension of our variable, here the brand of tea, is correlated with another, such as bag weight. Such additional information may warrant the assertion that there is a "real" difference between groups. If the statistic we derived from our samples is sufficiently deviant from zero—if it falls into the dark right tail of the sampling distribution in Figure 1––we conclude that the data are surprisingly ("significantly") deviant from what is expected given the Null.

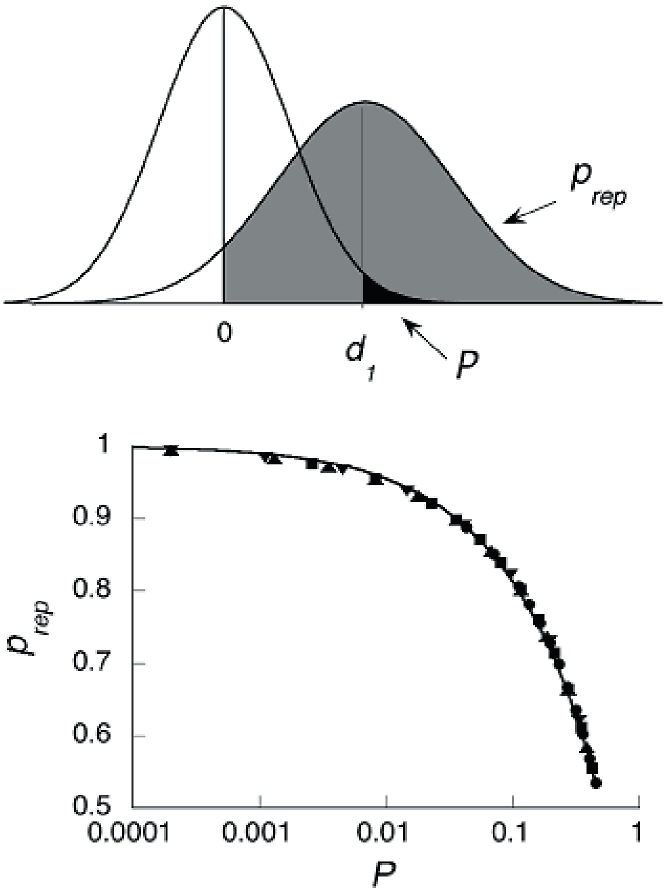


Figure 1. The left curve at top is the sampling distribution for a statistic such as a mean or effect size *(d)* under the null hypothesis. The traditional *p*-value is the area to the right of the obtained statistic, d_1, shown in black. Shift this curve to it's most likely position (the observed statistic) and double its variance (to account for the sampling error in the original plus that in the replicate) to create the distribution expected for replications. The probability of finding an effect of the same sign (p_{rep}) is given by the shaded area. The curve at bottom shows that as power or effect size change, *p* and p_{rep} change in complement. The figure is reproduced from Killeen (2005a).

Given more, we could conclude more. We utilized data from the experiment to estimate the standard deviation of the sampling distribution; why not also use it to estimate the mean? In for a penny, in for a pound: Slide the distribution in Figure 1 to the right, to center it over the measured effect size, d_1. We can't know that that is precisely where it belongs; d_1 will deviate randomly around the true (population) effect size δ. Its deviation is *sampling error*. Any attempted replication will also deviate from δ by its sampling error. And the replication statistic d_2 will deviate from the first value d_1 by the sum of those two errors. In the long run (many original experiments, many replications), the distribution of replication attempts will be centered on δ with a variance twice that of the observed data. This is shown by the bell curve on the right of Figure 1, centered over our best current estimate of δ, d_1 . That curve is the "posterior (after the first measurement) predictive distribution". It is our best guess of where, and with what probability, the statistics from replications of our experiment will fall.

Just as our evaluation of the fairness of the coin depends on whose pocket it came from, our evaluation of a scientific claim will depend on everything we know about it. But everyone will know different things about any phenomenon, and as soon as that subjective prior knowledge enters the picture, probabilities themselves become subjective—a function of both the data and who is answering the question. This is, after all, why people bet on horses—and on anything else that moves—despite how many data are already public: Each believes that their own subjective priors are better than the opponents'. But we may give data a fair shake by assuming that we know nothing about the phenomenon a priori, to let the data speak for themselves. This means using *uninformative* priors that wash out of our answer as soon as we have collected a few data.

2. The second step is to decide on what we mean by replicate. How close do we have to come? Most often the claim that wants testing is that a manipulation had an effect, or that a relationship exists between two variables. Suppose an original experiment found an effect size of 0.5, which might have arisen from a difference of 1.0 between mean scores of samples whose standard deviation averaged 2.0. The investigator claimed that her manipulation was effective. A replication finding an effect size of 0.4 would provide support for that claim. Indeed, a replication finding an effect size of 0.2, supports the original claim *even if it does not achieve traditional statistical significance*. It provides weak evidence in favor of the claim. Only effect sizes the opposite direction are evidence against the claim. Meta-analyses may show us that each additional experiment gives us additional confidence in the true effect size being significantly greater than zero, even if the constituent experiments did not achieve significance.

Let us therefore take *replication* to mean measurement of an effect *in the same direction* as the original. The probability of this happening is given by the gray area under the replicate sampling distribution to the right of 0, most easily found in a table as the area from $-\infty$ to $z = k\,M/\sigma$, with $k = 1/\sqrt{2}$.

There is obviously a close relation between this area, which I call the probability of replication p_{rep}, and traditional p values based on the same equation with $k = -1$. As the effect size or the number of observations varies, p and p_{rep} vary in complementary fashion, as is shown at the bottom of Figure 1. In particular, whenever a p value has been calculated, one can immediately infer p_{rep} by (a) calculating the z-score corresponding to $1 - p$, (b) dividing it by the square root of 2, and (c) finding the probability associated with this new z-score: $p_{rep} = N\,[N^{-1}(1 - p) \,/\, \sqrt{2}\,]$; that is, p_{rep} = *normsdist(normsinv(1–p)/sqrt(2))*, where *normsdist* is the cumulative distribution function and *normsinv* is its inverse. These may be found in the back of any statistics text, or issued as commands in a spreadsheet such as Excel®.

Circular Files To rule out reporting results because they don't achieve significance rules out the possibility of efficient and unbiased cumulation of the results by a later reviewer. This is known as the "file drawer problem", although to some it is the "circular file problem". The new vehicle for inference, p_{rep}, doesn't force the misperception that failure to achieve significance is tantamount to failure to replicate. It doesn't force us to trash data that, aggregated with others, can have real value for the community.

A Significant Difference Why bother with all this if p and p_{rep} are kissing cousins? Because, *viva la difference*, kissing cousins are not identical twins. One can never make a positive claim with NHST ("Never use the unfortunate expression 'accept the null hypothesis'"; Wilkinson & Task Force on Statistical Inference, 1999, p. 599); and, without priors, one can never make the negative claim of rejecting the Null. But p_{rep} permits positive claims, such as: "My data will replicate approximately 70 [or 80, or 90] percent of the time." Values of p_{rep} greater than 0.9 correspond to significant p-values. But even if your p_{rep} is (only) 0.8, that still permits a positive and informative statement concerning the replicability of your data; you are left with something better to hold onto than the foul bag of Failure to Reject the Null.

Fear of the Unknown One of the tedious aspects of statistics is remembering the details. Unless you teach statistics, your ability to distinguish between Type I and Type II Errors and give a quick definition of the latter will be less than perfect. Feel guilt no longer. Because p_{rep} is not predicated on the truth or falsity of the Null, it does not incur either type of error. A large value of p_{rep} does suggest that the Null is false; but the utility of p_{rep} is not *predicated* on the Null being true, as is the case for NHST. No need to stay awake at night wondering whether to use Neyman and Pearson's critical regions or Fisher's p values (Christensen, 2005). Use p_{rep}.

Does p_{rep} really predict replicability? It provides an estimate whose accuracy depends on the similarity of procedure and subjects. It also depends, like any probabilistic event, on the luck of the draw (Cumming, 2005). If variables are measured or behavior motivated differently, then "realization variance" must be added to the sampling variance to predict the results. This is a realistic random effects model of prediction (Killeen, 2005a). But the burden of adding that realization variance belongs to the replicator, who chooses how deviant the conditions will be, not to the originator.

How likely is it that the original results were a fluke, and will not replicate despite a large p_{rep}? Call a value of p_{rep} equal to ps "strong" evidence. The probability that a replication will provide strong support is 1 - NORMSDIST(NORMSINV(*ps*) - NORMSINV(p_{rep})). The probability that it will strongly contradict the original is 1 - NORMSDIST(NORMSINV(*ps*) + NORMSINV($\mathrm{p}_{\mathrm{rep}}$)). If we set $ps = .8$, and the original had a p_{rep} of .9, then the probability that a replication will provide strong support is 0.67; the probability of strong contradiction is 0.02.

Feel Real Confidence Replicability analysis, like traditional statistical analysis, is only half the story. Effect sizes are equally important, and should always be reported. An optimal inferential procedure would integrate effect sizes with the probability of replication, to achieve a true scientific decision theory. Presenting effect sizes in terms of a confidence interval is less than optimal, because confidence intervals are the alter-ego of NHST, and inherit the same difficulties of interpretation.

Whereas NHST takes a null effect as a default and hedges it with critical "significance" regions, CIs take the measured statistic as default and hedges it with limits. But a confidence interval is the difference between the population parameter and sample statistic, not territory on the *x*-axis. If the Null is true, the CI should be centered on 0; but if the statistic happens to equal the population parameter, then CI should be centered on that statistic. But if you knew which was the case, why do statistics? And if you don't know which is the case, you shouldn't put it anywhere (Estes, 1997).

"What to construct CIs around—and how to display them—remain issues for debate" (Fidler, Thomason, Cumming, Finch and Leeman, 2005, p. 495). They remain issues because their proper explanation is convoluted: "If the experiment were repeated 100 times and 100 confidence intervals like yours computed, approximately 95 of them would contain the population mean". Just what this means for your particular data is so difficult to understand that standard reference manuals either get it wrong (e.g., Zwillinger, 1996, p. 608) or make a strategic decision to misrepresent it.

Life really can be much simpler. The familiar standard error bars are, *mirabile dictu*, replication intervals. Drawn flanking the measured statistic, they can interpreted as the limits within which replications will fall approximately half the time (Cumming, 2005).

The First Chapter There's a prequel to my story, one told by Fisher about a test conducted with a hypothetical lady who averred she could taste the difference when tea was poured into milk, versus milk into tea. He used the story to introduce permutation tests (as Salsburg, 2001, used it to name his charming history of statistics). Permutation tests are much better than traditional statistics for analyzing most psychologists' research (Lunneborg, 2000), and can be used in concert with p_{rep} (Killeen, 2005b).

The Last Chapter It will require some experimentation to become comfortable with p_{rep}. The new statistic deserves its own treatment, but in the interim you can simply translate a *p*-value from any traditional test into a p_{rep}, and interpret it as above. Once you are comfortable with it, try using p_{rep} in your classes. You'll find fewer students like mine, brought to tears by *t*. How long will it take journals to come around? I only have two data, both positive. Given the small database, I conjured some subjective priors by visiting an establishment of divination where bones were thrown, palms read, and tea leaves swirled. There I met, of all people, my old student! She bore no malice, but carried instead a Tarot, *t*-tables, and a certificate in tasseography. Reading my leaves, she predicted: "Eventually all editors will cease fumbling with the knot, dispatch the null *bête noire,* and evaluate manuscripts by their significance, effect size, and replicability. But on that happy day, *significance* will mean what it means to their mothers, not what it means to their statisticians." Ahh … the pride we take in successful students!

Author Note

The research was supported by National Science Foundation Grant IBN 0236821, National Institute of Mental Health Grant 1R01MH066860, and by the Centre for Advanced Study at the Norwegian Academy of Sciences and Letters. I thank them and Bob Johnson for his patient midwifery. A spreadsheet for calculation of p_{rep} is available at: www.asu.edu/clas/psych/research/sqab/ .

References

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59,* 121-126.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997-1003.

Cumming, G. (2005). Understanding the Average Probability of Replication: Comment on Killeen (2005). *Psychological Science.* (in press)

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review, 4,* 330-341.

Fisher, R. A. (1936). Uncertain inference. *Proceedings of the American Academy of Arts and Science, 71,* 245-258.

Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). New York: Hafner Publishing Company.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294*(2), 218-228.

Killeen, P. R. (2005a). An alternative to null hypothesis significance tests. *Psychological Science, 16,* 345-353.

Killeen, P. R. (2005b). *Replicability, confidence and priors. Psychological Science.* (in press)

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112,* 662-668.

Lunneborg, C. E. (2000). *Data analysis by resampling: concepts and applications.* Pacific Grove, CA: Brooks/Cole/Duxbury.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth century.* New York: W. H. Freeman and Company.

Wilkinson, L. and the Task Force on Statistical Inference. (1999). Statistical methods in psychology: guidelines and explanations. *American Psychologist, 54,* 594-604.

Zwillinger, D. (Ed.). (1996). *CRC Standard mathematical tables and formulae* (30th ed.). Boca Raton, FL: CRC Press.