7

REPLICATION STATISTICS

PETER R. KILLEEN

We come finally, however, to the relation of the ideal theory to real world, or "real" probability. . . . To someone who wants [applications, a consistent mathematician] would say that the ideal system runs parallel to the usual theory: "If this is what you want, try it: it is not my business to justify application of the system; that can only be done by philosophizing; I am a mathematician." In practice he is apt to say: "try this; if it works that will justify it." But now he is not merely philosophizing; he is committing the characteristic fallacy. Inductive experience that the system works is not evidence.

Littlewood (1953, p. 73)

PROBABILITY AS A MODEL SYSTEM

For millennia, Euclidean geometry was a statement of fact about the world order. Only in the 19th century did it come to be recognized instead as a model system—an "ideal theory"—that worked exceedingly well when applied to many parts of the real world. It then stepped down from a truth about the world to its current place as first among equals as models of the world—the most useful of a cohort of geometries, each of differential service in particular cases, on spherical surfaces and relativistic universes and fractal percolates. In like manner, probability theory was born as an explanation of the contingent world—"'real' probability"—and, with the work of Kolmogorov among many others, it matured as a coherent model system, inheriting most features of the earlier versions of the probability calculus.

The abstraction of model systems from the world permits their development as coherent, clear, and concise logics. But the abstraction has another legacy: the eventual need for scientists to reconnect the model system to the empirical world. That such rapprochement is even possible is amazing; it stimulated Wigner's well-known allusion to "the unreasonable effectiveness of mathematics in describing the world." Realizing such "unreasonably effective" descriptions, however, can present reasonably formidable difficulties—difficulties that are sometimes overcome only by fiat, as noted by Littlewood, a mathematician of no mean ability,

Author's Note: The research was supported by National Science Foundation Grant IBN 0236821 and National Institute of Mental Health Grant 1R01MH066860.

and M. Kline (1980), a scholar of comparable acuity. The toolbox that helps us apply the "ideal theory" of probability to scientific questions is called inferential statistics. These tools are being continually sharpened, with new designs replacing old.

Intellectual ontogeny recapitulates its cultural phylogeny. Just as we must outgrow naive physics, we must outgrow naive statistics. The former is an easier transition than the latter. Not only must we as students of contingency deal with the gamblers' fallacies and exchange paradoxes; we must also cope with the academics' fallacies and statistical paradoxes that are visited upon us as idols of our theater, the university classroom. The first step, one already taken by most readers of this volume, is to recognize that we deal with model systems, some more useful than others, not with truths about real things. The second step is to understand the character of the most relevant tools for their application, their strengths and weaknesses, and attempt to determine in which cases their marriage to data is one of mere convenience and in which it is blessed with a deeper, Wignerian resonance. That step requires us to remain appreciative but critical craftsmen. It requires us to look through the halo of mathematics that surrounds all statistical inference to assess the goodness of fit between tool and task, to ask of each statistical technique whether it gives us leverage or just adds decoration.

This chapter briefly reviews—briefly, because there are so many good alternative sources (e.g., Harlow, Mulaik, & Steiger, 1997; R. B. Kline, 2004)—the most basic statistical technique we use, null hypothesis statistical testing (NHST) and its limits. It then describes an alternative statistic, p_{rep}, that predicts replicability. We remain mindful of Littlewood's (1953) observation that "inductive experience that the system works is not evidence [that it is true]." But then Littlewood was a mathematician, not a scientist. The search for truth about parameters has often befuddled the progress of science, which recognizes simpler goals as well: to understand and predict. If we "try [a tool, and] it works," that can be very good news and may constitute a significant advance over what has been. So, try this new tool, and see if it works for your inferential problems.

Connecting Probability to Data

You are faced with two columns of numbers, data collected from two groups of subjects. What do you want to know? Not, of course, "whether there's a significant difference between them." If they are identical, you would have looked for the clerical error. If they are different, they are different. You can review them 100 times, and they will continue to be different, hopefully 100 times; $p = 1.0$. "*Significantly* different," you might emphasize, irritated. But what does that mean? "That the probability that they would be so different by chance is less than 5%," you recite. OK. Progress. Now we just need clarification of *probability, so,* and *chance.*

Probability. Probability theory is a deductive calculus. One starts with probability generators, such as coins or cards or dice, and makes deductions about their behavior. The premises are precise: coins with a probability of heads of .50, perfectly balanced dice, perfectly shuffled cards. Then elegant theorems solve problems such as "Given an unbiased coin, what is the probability of flipping 6 heads in a row?" But scientists are never given such ideal objects. Their modal inferences are inductions, not deductions: An informant gives them a series of outcomes from flipping a coin that landed heads six times in a row, and they must determine what probability of heads should be assigned to the coin. They can solve this mystery either as Dr. Watson or as Mr. Holmes in the cherished tale, "The Case of the Hypothesis That Had No Teeth." As you well remember, Dr. Watson studiously purged his mind of all prior biases and opined that the probability of the coin being fair was manifestly $(½)^6, < .025$ and, further, that the best estimate of the probability of a heads was $1 - 2^{-6}$. Mr. Holmes stuffed his pipe; examined the coin; spun it; asked about its origin, how the coin was released and caught, and how many sequences were required to get that run of six; and then inquired about the bank account of the informant and his recent associates. Dr. Watson objected that that was going beyond the information given; in any case, how could one ever combine all those diverse clues into a probability statement that was not intrinsically subjective? "Elementary," Mr. Holmes observed, "probability theory this is not, my dear Watson; nor is it deduction. When I infer a state

of nature from evidence, the more evidence the better I infer. My colleague shall explain how to concatenate evidence in a later chapter of this sage book I saw you nodding over."

How do we infer probability from a situation in which there is no uncertainty—the six heads in a row, last week's soccer cup, your two columns of experimental data? There are two root metaphors for probability: For frequentists, probability is the long-run relative frequency of an outcome; it does not apply to novel events (no long run) or to accomplished events (*faits accomplis* support no probability other than unity). For Bayesians, probability is the relative odds that an individual gives to an outcome.

You do not have resources or interest to reconduct your experiment thousands of times, to estimate the relative frequency of two means being so different. And even if you did, you'd just be left with a much larger sample of accomplished data. This seems to eliminate the frequentist solution. On the other hand, the odds that you give to your outcome will be different from the odds your reviewers or editor or your significant other gives to your outcome. This eliminates any unique Bayesian probability. What next? Just imagine.

Instead of conducting the experiment thousands of times, just *imagine* that it has been conducted thousands of times. This is Fisher's brilliant solution to the problem of connecting data to the "ideal theory" of probability. Well then, just how big an effect should we imagine that your experiment yielded? Here we must temper imagination with discipline: We must imagine that the experiment never worked—that there was never a real effect in all those imaginary trials but that the outcomes were distributed by that rogue called Chance. Next, graph the proportion of outcomes at various effect sizes to give a sampling distribution. If your measured outcome happens only rarely among this cohort of no real effects, you may conclude that it really is not of their type—it does not belong on the group null bench. It is so deviant, you infer, because more than Chance was at work—the experimental manipulation was at work! This last inference is, as we shall see, as common, and commonly sanctioned, as it is unjustified.

If the thousands-of-hypothetical-trials scenario taxes the computational resources of your imagination, then imagine instead that the data were drawn from a large, normally distributed population of data similar to those of the control group. Increase the power of the test by estimating the population variance from the variances of both the experimental and control groups. Then a theoretical sampling distribution, such as the *t* distribution, can be directly used to infer how often so deviant an outcome as what you measured would have happened by chance under repeated sampling. This set of tactics is the paradigmatic modus operandi for statistical inference. In modern applications, the theoretical sampling distribution may be replaced with an empirical one, obtained by Monte Carlo elaboration of the original empirical distribution function (e.g., Davison & Hinkley, 1997; Mooney & Duval, 1993).

So in this standard scenario, *probability* means the long-run relative frequency that you *stipulate* in your test of the behavior of an ideal object. It is against this that you will test your data. Unlike a Bayesian probability, the parameters tested (e.g., the null hypothesis that the means of the two populations are equal) are not inferred from data. Indeed, authorities such as Kyburg (1987) argue that use of Bayesian inverse inference, leading up from statistics to parameters, undermines all direct inference thereafter, including NHST.

So. Let us assume your experiment recorded a standardized difference between the responses of 30 control subjects and 30 experimental subjects of $d = 0.50$, from which you calculated a p value of $< .05$. What does that mean? Does it mean that the probability of getting a d of 0.50 under the null is less than 5%? No, because we know a priori that the probability of getting exactly that value is always very close to zero; in fact, the more decimal places in your measurements, the closer to zero. That's true for any real number you might have recorded, even $d = 0.00$, which Chance favors.

Again an impasse, but again one that can be solved through imagination (providing inductive evidence supporting Einstein's chestnut "Imagination is more important than knowledge"). To get a probability requires an interval on the *x*-axis. We could take one around the observed value: say, $d \pm 0.05$. This would work; as we let the interval shrink toward zero, comparison with a similar extent around 0

would give us a likelihood analysis of the null versus the observed (Royall, 1997). Fisher was a pioneer of likelihood analyses but could not get likelihoods to behave like probabilities, and so he suggested a different interval on the evidence axis: He gave the investigator credit for everything more extreme than the observed statistic (a benefice that amazed and pleased many generations of young researchers, who might have been told instead, "The null will give effects *up to* that size 95% of the time"). So, what *so* means here is "more extreme than" what you found, including *d* scores of 1, 2, . . . 100. . . . Don't ask why those extents of the *x*-axis never visited by data should play a role in the decision; take the $p < .05$ and run.

Chance. "The probability [relative frequency under random sampling] that your data would be so [at least that] extreme by chance." Here *chance* means your manipulation did not work and only the null did. How does the null work? Typically, thanks to the central limit theorem, in ways that result in a Gaussian distribution of effects (although for other test statistics or inferences, related distributions such as the *t* or *F* are the correct asymptotic distributions). Two parameters completely determine the Gaussian: its mean and variance. Under the null, the mean is zero. Its variance? Since we have no other way to determine it, we use the data you brought with you, the variance of your control group. Well . . . the experimental group could also provide information. Hoping that the experimental operations have not perturbed that too much, we will pool the information from both sources. Then *chance* means that "with no help from my experimental manipulation, the impotent null could have given rise to the observed difference by the luck of the draw [of a sample from our hypothetical population]." If it would happen in less than 5% of the samples we hypothetically take, we call the data *significantly* different from that expected under the null.

You knew all that from Stat 101, but it is worth the review. What do you conclude with a gratifying $p < .05$? Also as learned in Stat 101, you conclude that those are improbably extreme data. But what many of us *thought* we heard was that we could then reject the null. That, of course, is simplistic at best, false at worst.

Foundational Problems With Statistical Inference

The Inverse Inference Problem. Assume the probability of the statistic S given the null (N) is less than a prespecified critical number, $p(S|N) < \alpha$, where the null may be a hypothesis such as $\mu_E - \mu_C = 0$. It does not then follow that the probability of the null given the statistic, $p(N|S)$, is less than α. We can get from one to the other, however, via Bayes theorem:

$$p(N|S) = p(S|N)p(N)/p(S).$$

By the (prior) probability of the null, $p(N)$, we mean the probability of the mathematical implication of the null (for instance, that two population means are equal, $\mu_E - \mu_C = 0$). That prior must be assigned a value before looking at the new data (S). By the probability of the statistic, $p(S)$, we mean its probability under relevant states of the world—in this case, the probability of the data given that the null is true, plus the probability of the data given that the null is false: $p(S) = p(S|N)p(N) + p(S|{\sim}N)(1 - p(N))$. By normalizing the right-hand side of Bayes's equation, $p(S)$ makes the posterior probabilities sum to 1. Only in the unlikely case that the prior probability of the null equals the probability of the statistic, $p(N) = p(S)$, does $p(N|S) = p(S|N)$. Otherwise, to have any sense of the probability of the null (and, by implication, of the alternative we favor, that our manipulation was effective: $\mu_E - \mu_C > 0$), we must be able to estimate $p(N)/p(S)$. This is not easy. Even if we could agree on assigning a prior probability to the null, Bayes would not give us the probability of our favored hypothesis (unless we defined it broadly as "anything but the null"). In light of these difficulties, Fisher (1959) made it crystal clear that we generally cannot get from our *p* value to *any* statement about the probability of the hypotheses. But it was to make exactly such statements about our hypotheses, with the blessings of statistical rigor, that we attended all those statistics courses. We were misled, but it was not, we suspect, the first or last time that happened, which goes some distance to explaining the conflation of *statistics* with *lies* in the public's mind.

"It is important to remember that [relative frequency] is but one interpretation that can be given to the formal notion of probability" (Hays,

1963, p. 63); given our inferential imbroglio, we may well wonder if that interpreter ever spoke the language of science. Neyman and Pearson solved the problem of inverse inference by emphasizing comparison with an alternate hypothesis, setting criterial regions into which our statistic would either fall, or not, and noting that, whereas we have no license to change our beliefs about the null even if our statistic falls into such a critical ($p < \alpha$) region, it would nonetheless be prudent to change our behavior, absent a change in belief. Like Augustine's *credo quia impossibile,* this tergiversation does solve the problem—but only for those whose faith is stronger than their reason.

What did we ever do that got us into this mess? We wanted to know if our manipulation (or someone else's, perhaps Nature's, if this was an observational study) really worked. We wanted to know how much of the difference between groups was caused by the factor that we used to sort the numbers into two or more columns. We wanted to know if our results will replicate or if we are likely to be embarrassed by that most odious of situations, publication of unreplicable results. Null hypothesis statistical tests cannot get us from here to there. Let us go back to basics and see if we cannot find a viable alternate route to some of our valid goals.

Defining Replication

Curious, isn't it, that we have license to use the measured variance (s^2) in estimating a parameter of the population under the null hypothesis (σ^2), but we do not use the measured mean to estimate a parameter? Why evaluate data with one hand tied behind our back? Assay the following hypothesis instead: H_A: $\delta = d$, where δ is the value of the population effect size, and d is the effect size you measured in your experiment. But to ask the probability that this alternative to the null is true seems to be creating a new logical fallacy: *post hoc, ergo hoc.* (Gigerenzer [2004] called a similar solecism "The Feynman Fallacy" because Feynman was so exasperated when a young colleague asked him to calculate the probability of an accomplished event.) But we can ask a different, noncircular question, one that takes advantage of the first two moments of the observed data. What is the probability that, using the same experimental operations and population of subjects, another investigator can replicate those results? This can be computed once we agree on the meaning of *replicate.* Consider this definition:

To *replicate* means to repeat the empirical operations and to record data that support the original claim.

- If the claim is as modest as "This operation works [generates a positive effect]," then any replication attempt that finds a positive effect could be deemed a successful replication. Although this may seem a too-modest threshold for replication, put it in the context of what traditional significance tests test: In a one-tailed test, $1 - p$ gives us the probability that our statistic d has the same sign as the population parameter (Jones & Tukey, 2000).
- If the claim is "This operation generates an effect size of at least d_L," then only replication attempts that return $d \geq d_L$ count as successful replications.
- If the claim is "This operation generates a significant effect," then only replication attempts that return a $p < .05$ are successful replications.
- If the claim is "This operation is essentially worthless, generating effect sizes less than d_U," then any replication attempt that returned a $d < d_U$ could be deemed a successful replication. One would have to decide beforehand if a d less than, say, –0.5 was as consistent with the claim or constituted evidence for a stronger alternative claim, such as "This operation could backfire."

In the following, we shall show how to compute such probabilities.

Predicting Replicability Rather Than Inferring Parameters

How. How to estimate these probabilities? The sampling distribution of effect sizes in replication, $p(d_2|d_1)$, is required.

Effect size is calculated as

$$d = \frac{M_E - M_C}{s_p}, \qquad (1)$$

with M_E the mean of the experimental group, M_C the mean of the control group, and s_p the estimate of the population standard deviation based on the pooled standard deviations of both groups (see the appendix for more information). Consider an effect size $d = d_1$ based on a total of n observations randomly sampled from a population with unknown mean δ and variance σ^2. The effect size d_2 for the next m observations constitutes the datum that we wish to predict based on the original observations: $f(d_2|d_1)$. Because d_1 provides information about d_2, these statistics are not independent. They are only independent when considered as samples from a large population whose parameter is δ. Solution proceeds by considering the joint density of d_2 and δ conditional on the primary observations. This is developed in the appendix, leading to the distribution shown as the flatter curve in Figure 7.1. In evaluating a claim concerning experimental results, calculate replicability by integrating that density between the appropriate limits:

$$p_{\text{rep}} = \int_{d_L}^{d_U} n(d_1, s^2_{d_R}) = \Phi \frac{d_U - d_1}{s_{d_R}} - \Phi \frac{d_L - d_1}{s_{d_R}}, \quad (2)$$

In cases where a positive effect has been claimed, and we stipulate that the hypothetical replication would have the same power as the accomplished experiment—everything is done exactly as in the original experiment—then $d_L = 0$, $d_U = \infty$, and $S^2_{d_R} = 2S^2_d$. This is shown as the gray area of the predictive distribution on the right in Figure 7.1. If the claim is that the effect size is greater than d^*, the probability of getting supportive evidence is predicted by p_{rep} with $d_L = d^*$ and $d_U = \infty$. The probability of finding a significant effect size in replication is given by p_{rep} with $d_L = S^2_{d_R} z_\alpha$. Other claims take other limits.

For convenience in calculation of the standard case, Equation 2 may be rewritten as

$$p_{\text{rep}} = \int_{-\infty}^{d_1/\sigma_{dR}} n(0, 1) = \Phi\, d_1/s_{d_R}, \quad (3)$$

with σ_{d_R} estimated by S_{d_R}.

The variance of d is

$$s^2_d \approx \frac{n^2}{n_E n_C (n-4)},$$

with $n = n_E + n_C$. When experimental and control groups are of equal size, $n_E = n_C$, then

$$s^2_d \approx \frac{4}{n-4},$$

and $s^2_{d_R} = 2s^2_1$. Predicting the effects in a different-size prospective sample requires adjusting the variance. These considerations are reviewed in Killeen (2005a), whose derivation was corrected along the lines of the appendix by Doros and Geier (2005).

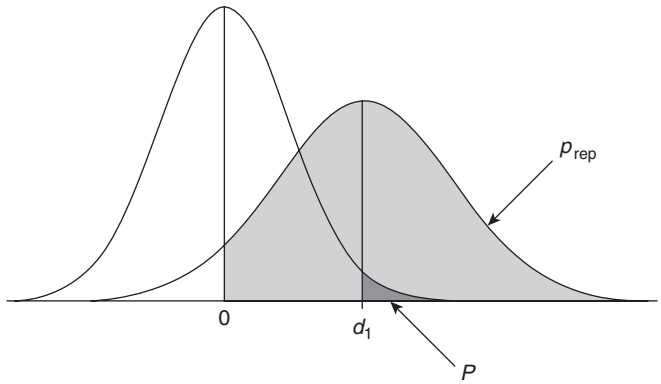

Figure 7.1 The dark area to the right of the observed effect size d_1 gives the probability of finding data more extreme than d_1, given the null. The gray area to the right of 0 gives the probability of finding a positive effect in replication (Equation 2). The figure is reproduced from Killeen (2005a), with permission.

Relation to p. How does the probability of replication compare to traditional indices of replicability such as *p?* The *p* value is the probability of rejecting the null hypothesis given that the datum, d_1, is sampled from a world in which the null is true. It is shown as the area to the right of d_1 in Figure 7.1, under a normal density centered at 0 and having a variance of σ_d^2, estimated from S_d^2. The value of p_{rep} for the same data—the probability of finding a positive effect in replication—is the shaded area to the right of 0 in the normal curve centered on d_1 and having a variance of σ_{dR}^2, estimated from $S_{d_R}^2 = S_d^2$. As d_1 moves to the right, or as the variance of d_1, S_{d1}^2, decreases, p_{rep} will increase and *p* decrease in complement. For use in spreadsheets such as Excel, Cumming (2005) suggested the notation $p_{rep} = \text{NORMSDIST}([\text{NORMSINV}(1 - p)]/\sqrt{2})$, where NORMSDIST is the standardized normal distribution function, and NORMSINV is the normal *p*-to-*z* transformation. Drawn in probability coordinates, the relation between these two probabilities is transparent: The *z* scores for p_{rep} decrease linearly with the *z* scores for *p*, $z_{Prep} = -kz_p$, with $k = 1/\sqrt{2}$.

Advantages of p_{rep} *over* p. Given the affinity between *p* and p_{rep}, why switch? Indeed, a first reading may give the impression that *p* is preferable: Sentences that contain it also contain *hypotheses,* and those are what scientists are interested in. Conversely, p_{rep} does not give the probability that your hypothesis is true, or that the null is true, or that one or the other or both are false. It gives the long-run probability that an exact replication will support a claim. Both the original and the replicate may work for the wrong reasons—widdershins sampling errors in both cases. Rational scientists would prefer to know whether their hypotheses are true or false, not just whether their data are likely to replicate.

But that knowledge is not granted by statistical inference. Null hypothesis statistical tests never let us assign probabilities to hypotheses (Fisher, 1959). This is a manifestation of the unsolved problem of inverse statistical inference (Cohen, 1994; Killeen, 2005c). Students are admonished, "Never use the unfortunate expression 'accept the null hypothesis'" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599); without priors on the null, they should also be admonished, "Never use the equally unfortunate expression 'reject the null hypothesis.'" One can make no claims more interesting than "My data are absurdly improbable under the null," leaving the implication unsaid. Stipulating the null keeps its truth value off the table as a conclusion.

A researcher *can* say, "My hypothesis led me to predict a positive effect from this manipulation. I found one, and if you repeat my manipulation, you have a probability of approximately p_{rep} of also finding one." As ever, the investigator can assert that the manipulation caused the effect, and the hypothesis that birthed the manipulation was correct, only to the extent all confounds were eliminated, as other causes may have been more operative than the ones manipulated; those alternative causes may have been systematic, or possibly discernable only as "error [sampling] variance." In the long run (as *n* increases), nonetheless, exact replications should succeed with probability p_{rep}. Whereas *failure to reject the null* is not easily interpreted, a $p_{rep} = .85$ is just as interpretable as a $p_{rep} = .95$. Other advantages are discussed in Killeen (2005a), who did not adequately emphasize that p_{rep}, based on a single experiment, is only an estimate of replication probability (Cumming, 2005; Iverson, Myung, & Karabatsos, 2006). Like confidence intervals and *p* values, its accuracy depends on just how representative of the population the original sample happened to be. Any one estimate of replicability may be off, but in the long run, p_{rep} provides a reliable estimate of replicability. Evidence of this is seen in p_{rep}'s ability to predict the proportion of replications in meta-analyses (Killeen, 2005a).

The Distribution of *p*, p_{rep}, and Log-Likelihood Ratios

In order to compare p_{rep} with its cousins, *p* and the log-likelihood ratio (LLR), a simulation was conducted to generate an empirical sampling distribution of effect sizes d_i. Twenty thousand samples of size $n = 60$ were taken from a normal distribution with mean 0.5 and variance $\sigma_d^2 = 4/(60 - 4)$. This corresponds to the sampling distribution of differences between the means of experimental and control groups of 30 observations each, when the true difference between them is half a standard deviation ($\delta = 0.5$). This should yield a typical $p_{rep} = .907$ and $p = .029$. Values of p_{rep}, *p,* and LLR were calculated, along with the *z* score transformation of

p_{rep}, NORMSINV(p_{rep}). As expected, the distribution of p_{rep} is negatively skewed (coefficient of skewness [CS] = –1.60; mean = .859; median = .906; see Figure 7.2 and Cumming, 2005). This skew has been cited as a fault of p_{rep} (Iverson et al., 2006). However, the distribution of p is even *more* strongly skewed (CS = 2.41; mean = .093; median = .031). Even though the means are biased, however, both cases the median values of the test statistics are very close to their predicted values. The sampling distribution for the z scores of p_{rep} closely approximated a normal density, having a CS of 0.02. To aggregate values of p_{rep} over studies, it is therefore the z transform of p_{rep} that should be averaged (inversely weighted by the number of observations in each sample, if those differ).

For comparison, consider the likelihood of the data (d_i) given that the parameter equals zero, divided by the likelihood given that the parameter equals that observed, here 0.5. In the present scenario, its expected value is the ratio of the ordinate of the normal density at $d = 0.5$ under the null (with mean at 0) to that of the density under the alternate (with mean at 0.5). The natural logarithm of this ratio is the LLR, which is simply computed as $-z^2/2$. For the present exercise, this is $-(.5)^2/(4/(60-4))/2 = -1.75$. The distribution of LLRs is shown in the right panel of Figure 7.2. The likelihood ratio is positively skewed (CS = 0.93) and, as visible in Figure 7.2, the LLR is negatively skewed, to about the same degree as p_{rep} (CS = –1.46; mean = –2.25, median = –1.75, just as predicted). Readers may experiment with these and related distributions at the excellent site maintained by Cumming (2006).

The Bayes factor is an analogous statistic favored by Bayesians such as Iverson et al. (2006) and M. D. Lee and Wagenmakers (2005). If the hypotheses are simple and their prior plausibilities equal, than the Bayes factor is the likelihood ratio (P. M. Lee, 2004). If these conditions are not met, then a prior distribution for the parameters must be chosen and then integrated out (see, e.g., Wagenmakers & Grünwald, 2006). The distribution of the Bayes factor will depend on the nature of that prior distribution.

Refutation and Vindication

Predicting a positive effect of any size in replication may seem too weak a prediction to merit

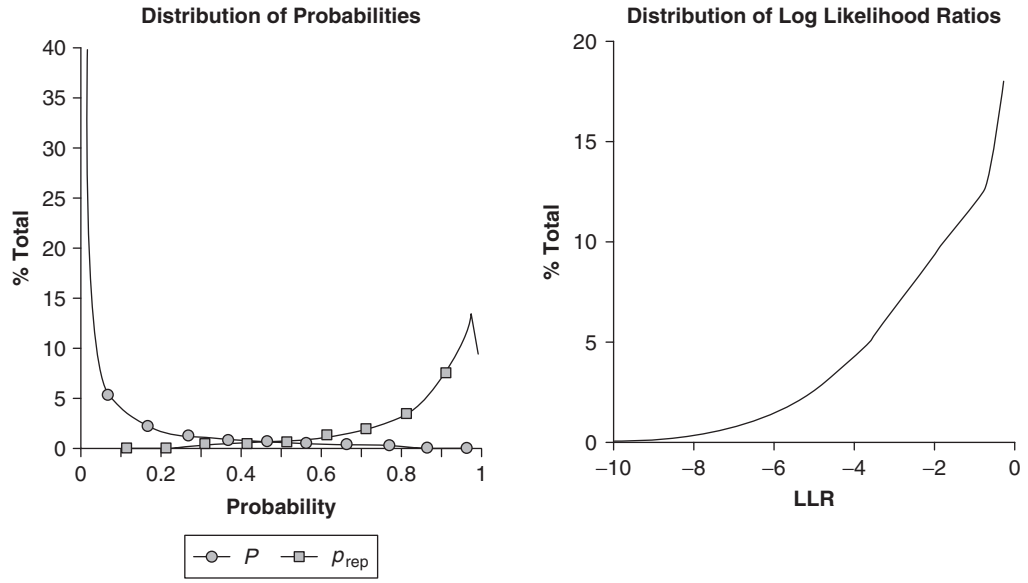

Figure 7.2 Twenty thousand samples of a normally distributed random variable, with mean $\delta = 0.5$ and variance $\sigma^2 = 4/(60-4)$, yielded these distributions of three test statistics.

attention. It is therefore useful to identify two kindred indices of replicability, the probability of strong support, p_{sup}, and the probability of strong contradictory results, p_{con}. Let us measure the degree of support that a real replication gives as its own value of p_{rep}. Set some threshold for calling a result *strong*, say, $p^* = .8$. Then a replication that returns a $p_{rep} > p^*$ is called "strong support," and one that returns a p_{rep} greater than p^* *in the wrong direction* (i.e., the replication's $p_{rep} < 1 - p^*$) is called "strong contradiction." The middling outcomes between these are called weak support or contradiction depending on their sign. The criteria for just what constitutes strong support and strong refutation are arbitrary; we could use p^*s of .75 or .8 or .9 or any other percentile. It is easy to calculate the resulting probabilities of support and refutation:

$$p_{sup} = 1 - \text{normsdist}[\text{normsinv}(p^*) - \text{normsinv}(p_{rep})];$$

$$p_{con} = 1 - \text{normsdist}[\text{normsinv}(p^*) + \text{normsinv}(p_{rep})].$$

Selected values of these expectations are found in Table 7.1. The first criterion in that table is $p^* = .5$. This returns the probability that a replication will have an effect of the same sign. This is obviously just p_{rep}. The probability of a contradiction—an effect of an opposite sign—is the complement of p_{rep}. Consider next the row indexed by $p_{rep} = .95$. This corresponds to the threshold LLR that Bayesians consider strong evidence, as well as to a one-tailed critical region for p of .01. The probability of strong support at a $p^* = .8$ is .789. The probability of a replication retuning an effect that is significant at $p < .05$ is given by $p^* = .88$, close to $p^* = .9$. For $p_{rep} = .95$, this happens about 2/3 of the time.

The probability of bad news—strong contradiction—is also found in Table 7.1. For $p_{rep} = .95$, $p^* = .8$, it is .006: Fewer than one attempted replication out of a hundred will go that far in the opposite direction from the original data. Experiments that yield a p_{rep} in excess of .9 are unlikely to be refuted (on statistical grounds, at least!). Variations in the execution of replication attempts (e.g., those deriving from changes in the measurement instruments and experimental or observational context) will inevitably add realization variance and, to the extent that they do so, will make these estimates optimistic.

Readers familiar with signal detection theory will immediately see that effect size d is nothing other than their familiar index of discriminability d'. The criterion p^* is analogous to bias: Changes in p^* do not move the criterion from left to right but move two criteria in and out. The data in Table 7.1 can be represented as points along receiver operating characteristics (ROCs), with p_{sup} corresponding to hits and p_{con} to false alarms. The resulting isosensitivity functions for the first few columns of Table 7.1 are shown in Figure 7.3, with the criterion p^* increasing from .5 to .98 in smaller steps than shown in that table to draw the curves. These

Table 7.1 The Probability That a Result Will Be Replicated or Refuted at Different Levels of Confidence Given the Strength of the Original Results

Criteria for strong support or contradiction		0.5		0.75		0.8		0.85		0.9	
p	p_{rep}	p_{sup}	p_{con}	p_{sup}	p_{con}	p_{sup}	p_{con}	p_{sup}	p_{con}	p_{sup}	p_{con}
0.1000	0.818	0.818	0.182	0.592	0.057	0.526	0.040	0.448	0.026	0.354	0.014
0.0500	0.878	0.878	0.122	0.687	0.033	0.626	0.022	0.550	0.014	0.453	0.007
0.0250	0.917	0.917	0.083	0.762	0.020	0.707	0.013	0.637	0.008	0.542	0.004
0.0100	0.950	0.950	0.050	0.834	0.010	0.789	0.006	0.729	0.004	0.642	0.002
0.0050	0.966	0.966	0.034	0.874	0.006	0.836	0.004	0.784	0.002	0.705	0.001
0.0025	0.976	0.976	0.024	0.905	0.004	0.874	0.002	0.829	0.001	0.759	0.001
0.0010	0.986	0.986	0.014	0.935	0.002	0.910	0.001	0.875	0.001	0.817	0.000

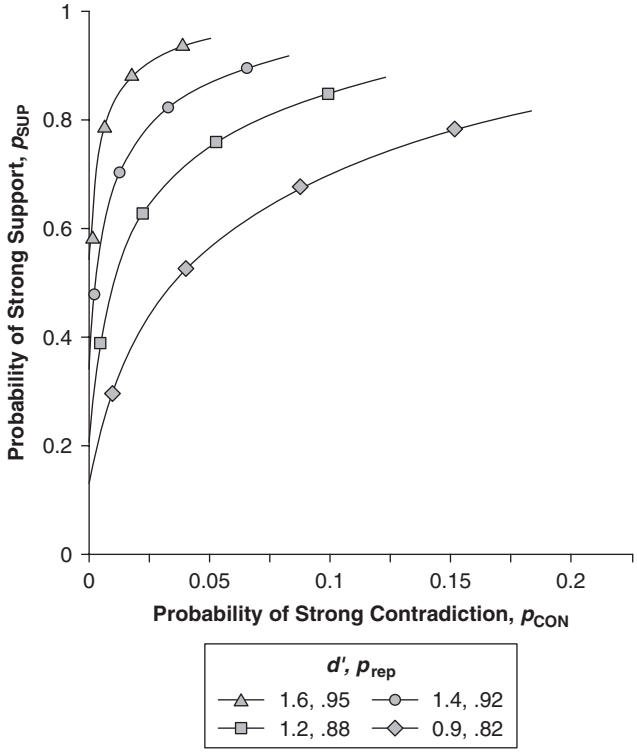

Figure 7.3 The probability of strong support (ordinates) and probability of strong contradiction (abcissae) both increase as the criterion for *strong* ($p^{\star}$) is reduced from an austere .98 (lower left origin of curves) to a magnanimous .50. The parameter is p_{rep} and its corresponding *z* score, *d′*.

are not the traditional *yes/no* ROCs, but *yes/no/uncertain,* with the last category corresponding to replications that will fall between p_{sup} and p_{con} in strength.

One may also calculate the probability of a nil effect in replication. This is analogous to the probability of the null. If one takes a nil effect to be one that falls within, say, 10% of chance ($.4 < p_{rep} < .6$), the probability of a nil effect in replication when $p_{rep} = .95$ is 5%. This probability may easily be calculated as $p_{nil} = P^{L}_{sup} - P^{U}_{sup}$, where P^{L}_{sup} corresponds to the probability of support returned by the lower limit ($p^{\star} = .4$ in the example), and P^{U}_{sup} corresponds to the probability of support returned by the upper limit ($p^{\star} = .6$ in the example). For further discussion, see Sanabria and Killeen (2007).

Credible Confidence Intervals. A small but increasing number of editors are encouraging researchers to report measures of effect size (Cumming & Finch, 2005). A convenient way to specify the margin of error in effect size is by bounding it with confidence intervals (CI; see Chapter 17). Unfortunately, most investigators do not really understand what confidence intervals mean (Fidler, Thomason, Cumming, Finch, & Leeman, 2004). This confusion can be remedied by study of Cumming and Finch (2001), Loftus and Masson (1994), and Thompson (1999), along with the handy chapters of this volume. An alternate approach provides a more intuitive measure of the margin of error in data. One such measure is the range over which a replication will fall with some stipulated probability. A convenient interval to use is the standard error of the statistic. Approximately half the replication attempts will fall within ± 1 standard error, centered on the measured statistic (Cumming, Williams, & Fidler, 2004). I call this kind of CI a *replication interval* (RI). Its interpretation is direct, its calculation routine, and its presentation as error bars hedging the datum unchallenged (Estes, 1997).

Evidence, Belief, and Action

What prior information should be incorporated in the evaluation of a research claim? If there is a substantial literature in relevant areas, then the replicability of a new claim can be predicted more accurately by incorporating that information. For some uses, this is an optimal tactic and constitutes a running meta-analysis of the relevant literature. The evaluation of the evidence at hand would then, however, be confounded with the particular prior information that was engaged to optimize predictions. Other consumers, with other background information, would then have to deconvolute the experimental results from those priors. For most audiences, it seems best to bypass this step, letting the data speak for themselves and letting the consumers of the results add their own qualification based on their own sense of the prior results in the area (Killeen, 2005b). This decision is equivalent to assuming that the prior distribution of the population effect size is flat and is consistent with some analysts' advice to use only likelihood ratios (e.g., Glover & Dixon, 2004; Royall, 1997), not Bayesian posteriors, thereby eschewing the difficulties in the choice of priors. Royall (2004) is perhaps the most cogent, noting three questions that our statistical toolbox can help us address:

1. How should I evaluate this evidence?
2. What should I believe?
3. What should I do?

1. The first question confronts us when data are first assembled. Here, considerations of the past (priors) and the future (different prospective populations) confound analysis of the data on the table. Such externals "can obfuscate formal tests by including information not specifically contained within the experiment itself" (Maurer, 2004, p. 17). Royall (2004) and Maurer (2004) argue for an evidential approach using the likelihood ratios. They eschew the use of priors to convert these into the probability of the null because this ties the analysis to a reference set that may not be shared by other interested consumers of the evidence.

2. *Belief,* however, should take into account prior information, even information that may be particular to the individual. Each of us carries a unique reference set with which we update our beliefs in light of evidence. This is why serious crimes are evaluated by large juries of peers: large, to accommodate a range of reference sets; peers, to relate those priors to ones most relevant to the defendant. The transformation of evidence into belief (concerning a hypothesis or proposition) transforms a public datum into a personal probability. Shared evaluation of strong evidence will bring those personal probabilities toward convergence, but they will be identical only for those with identical reference sets. Savage (1972) took such personal probabilities to be "the only probability concept essential to science" (p. 56). Royall (2004) and Maurer (2004) disagree. Belief is indeed best constructed with Bayesian updating and motivates the acceptance or rejection of scientific theories, but evidence evaluation must be kept insulated from priors. Once that evaluation is executed, belief adjustment is natural. But beliefs should concern claims and hypotheses; they should not contaminate evidence.

3. What we should *do* depends both on what we believe and what we value. Decision theory tells us how to combine these factors to determine the course of action yielding the greatest expected benefit. In particular, the posterior predictive distribution shown in Figures 7.1 and 7.3 estimate the probability of different effect sizes in replication. If we can assign utility to outcomes as a function of their expected effect size, we can determine what values of d_1 fall above a threshold of action. The sigmoid function in Figure 7.4 represents such a utility function. Multiplying the probability of each effect size in replication by the utility function and summing gives the expected utility of replicating the original results.

Given a posterior predictive distribution, it is straightforward to construct a decision theory to guide our action. Winkler (2003) teaches the basics, and Killeen (2006) applies them to recover traditional practices and extends them in nontraditional ways. The execution may employ flat priors and identical prospective populations; it then will guide disposition of the evidence: whether it be admitted to a corpus or rejected, whether the paper be published or not. Or the execution may employ informative priors and may take into consideration the realization variance involved in generalizing to new

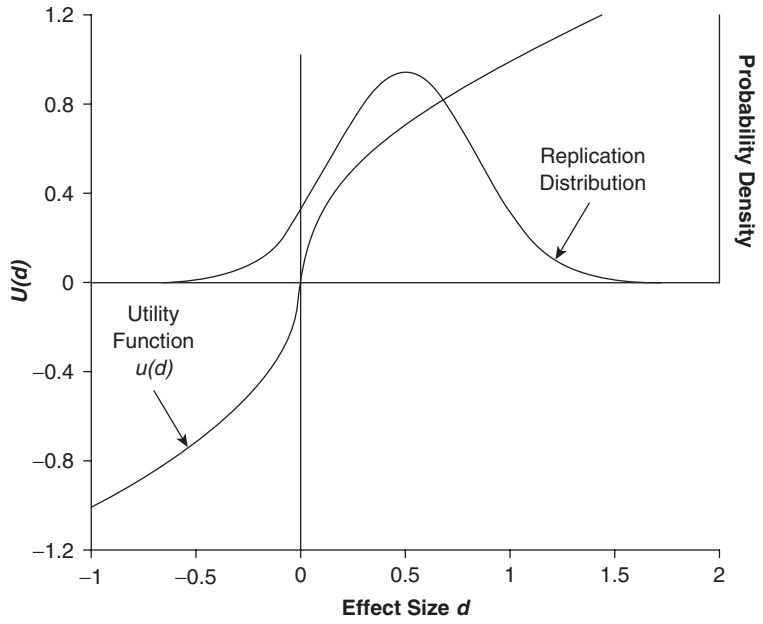

Figure 7.4 The Gaussian density is the posterior predictive distribution for a result with an effect size of $d_1 = 0.5$. The sigmoid is an example of a utility function on effect size that expresses decreasing marginal utility as a function of effect size and treats positive and negative effects symmetrically.

prospective populations. It then becomes a pragmatic guide for action, an optimal tool to guide medical, industrial, and civic programs.

Realization Variance

Whenever an attempt is made to replicate, it is inevitable that details of the procedure and subject population will vary. Successful replication with different instruments, instructions, and kinds of subjects lends generality to the results—but it also increases the risk of different results. This risk may be represented as *realization variance*, σ_{δ_j}, the uncertainty added by deviations from exact replication (Raudenbush, 1994; Rubin, 1981; van den Noortgate & Onghena, 2003). The subscript j indicates that this variance is indexed to a particular field of research. The estimated variance of effect size in replication becomes

$$s^2_{d_R} = (s^2_{d_1} + s^2_{\delta_j}) + (s^2_{d_2} + s^2_{\delta_j}). \quad (4)$$

If the replication involves the same number of subjects, then $s^2_{d_2} = s^2_{d_1}$, and the estimated standard error of replication is

$$s_{d_r} = \sqrt{2(s^2_{d_1} + s^2_{\delta_j})}. \quad (5)$$

In a meta-analysis of studies involving 8 million participants, Richard, Bond, and Stokes-Zoota (2003) reported a mean *within-literature* variance of $\sigma^2_\delta = 0.092$ (median = 0.08; $\sigma^2_\delta = 0.30$), after correction for sampling variance (Hedges & Vevea, 1998). This substantial realization variance puts an upper limit on the probability of replicating results, even with n approaching infinity. The limit is given by Equation 3 with $d_L = -\infty$ and

$$d_U = d_1/\sqrt{2}s_{\delta_j}.$$

This limit on replicability is felt most severely when d_1 is small: For the typical realization variance within a field (0.08), the asymptotic p_{rep} for $d_1 = 0.1$, 0.2, and 0.3 is .60, .69, and .77. It requires an effect size of $d_1 = 0.5$ to raise replicability into the respectable range ($p_{\text{rep}} = .9$). Alas, that is substantially larger than the effect size typical of the social psychological literature, found by Richard and associates to be $d_1 \approx 0.3$. It appears that we can expect a typical ($d_1 = 0.3$)

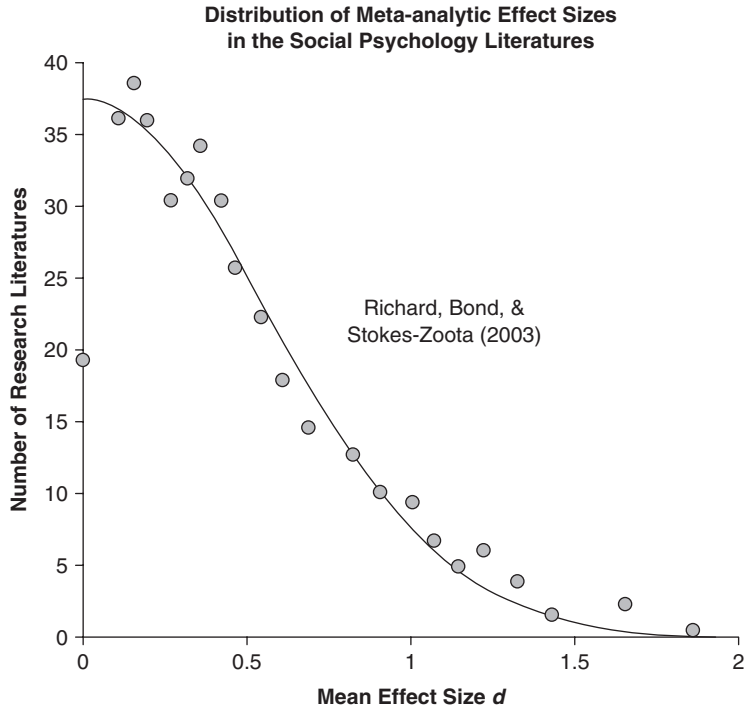

Figure 7.5 The distribution of effect sizes measured in 474 research literatures in social psychology. The data were extracted from Richard et al. (2003, Figure 1). The curve is Gaussian with mean 0 and overall variance 0.3.

research finding to receive positive support at any level in replication only about 75% of the time.

Validation. Simulations were conducted to validate the logic of p_{rep} and the accuracy of normal approximation for the noncentral t sampling distribution of d for small n. The analysis of Richard et al. (2003) provided a representative set of parameters. These investigators displayed the distribution of effect sizes for social psychological research involving 457 literatures, comprising 25,000 studies. Their measure of effect size, r, was converted into d by the relation $d = 2r(1 - r^2)^{-1/2}$ (Rosenthal, 1994). The authors reported absolute values of effect sizes because the direction of effect often reflects an arbitrary coding of dependent variables; however, it also may inflate the impression of replicability. The smooth curve through the data in Figure 7.5 provides another perspective on their report. (The nonpublication of small or conflicting effects—the "file drawer effect"—might be responsible for the outlier near $d = 0$.) The data in Figure 7.5 are used to create a population from which the simulation will sample effect sizes. The within-literature standard deviation σ_δ was set to 0.30, consistent with Richard and associates' estimate. Given these fingerposts, the simulation is described in Figure 7.6.

The relative frequency of successful replication was gauged for values of $n = n_C + n_E$ ranging from 6 to 200, for nine ranges of $|d|$ starting at 0 with upper limits of 0.08, 0.16, 0.24, 0.33, 0.43, 0.53, 0.65, 0.81, and 1.10. These frequencies, expressed as probabilities of replication (p_{rep}), are the ordinates of Figure 7.7. The abcissae are from Equation 2, with d_1 taken as the midpoints of the nine ranges. s_d^2 was calculated from the simulata (see the appendix). The only parametric information used in the predictions was the realization variance σ_δ^2, set to 0.09. The predictions held good down through very small ns, with average absolute deviations of 1.2 percentage points, on the order of binomial variability around exact predictions.

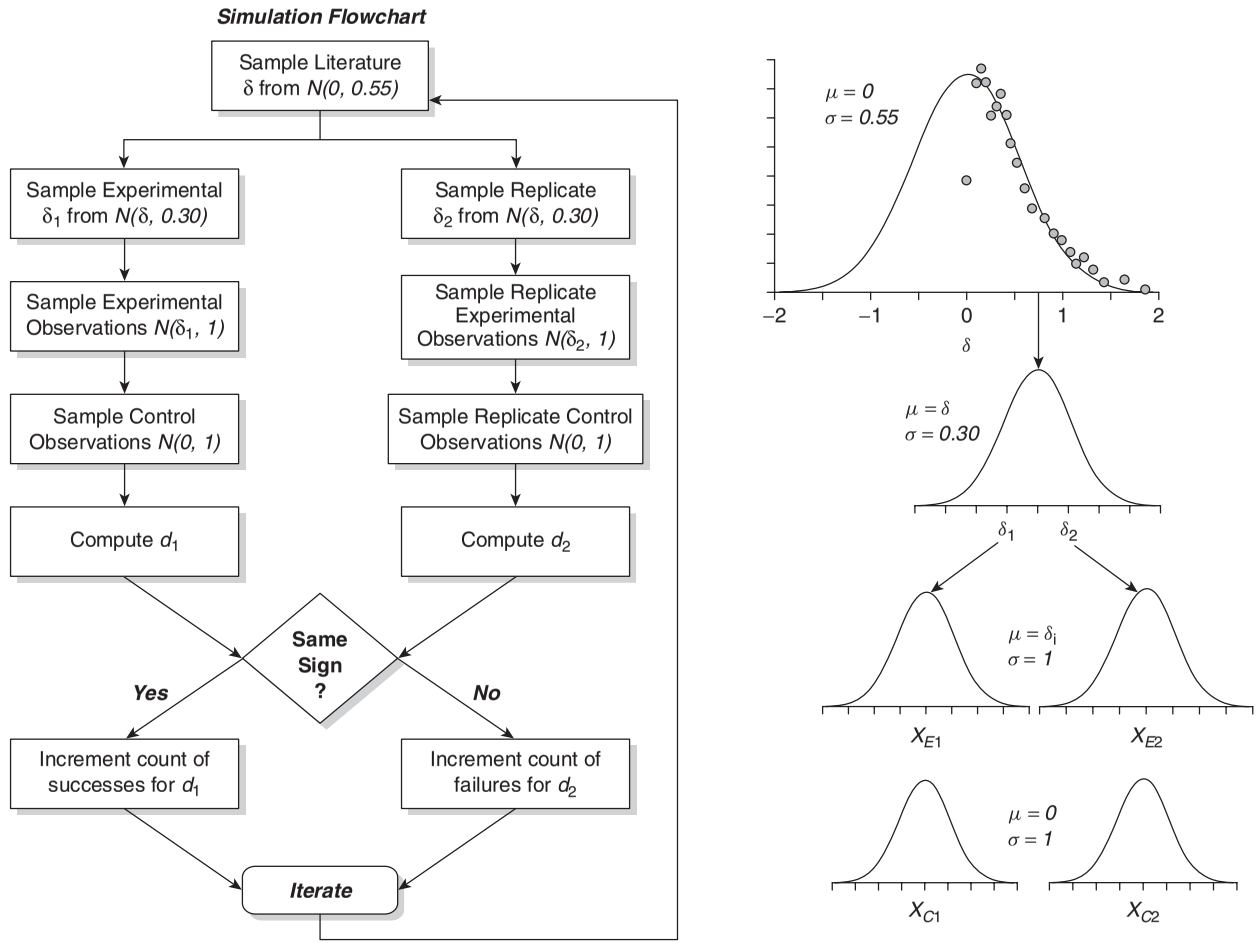

Figure 7.6 The simulation: A representative effect size of δ was first sampled from the distribution shown in Figure 7.5. Then two instances of δ_j from the literature it represents were selected, one for the original study and one for the replicate. The data were normally distributed random variables for the control and experimental groups, with the latter sampled from a population shifted by δ_1 (original study) or δ_2 (replicate). The difference between these parameters arises from the realization variance of 0.09 ($\sigma_\delta = 0.3$). The proportion of times an effect size from the original study predicted the sign of the replicate was tallied; information about the magnitude of the replication effect was not retained. The process was iterated 20,000 times for each *n* studied.

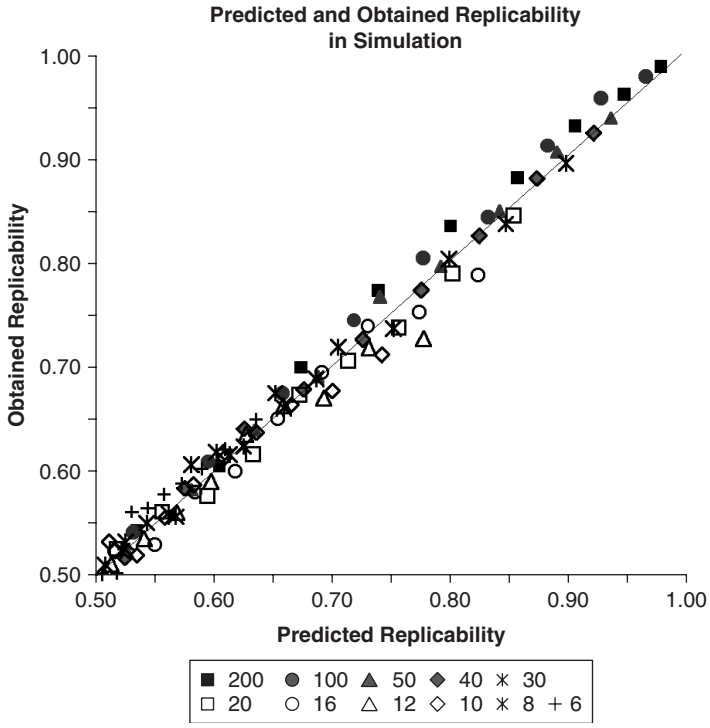

Figure 7.7 Results of the simulation. The obtained proportions of successful replications are plotted against the area over the positive axis of $N(d_1, s^2_{d_R})$. Symbols indicate the total number of observations in the experimental and control groups combined.

Meta-Analysis

Despite the width of the distributions of p_{rep} shown in Figure 7.2, p_{rep} proves generally accurate in predicting replicability, both in simulations such as those above and in meta-analytic compendia. For any one study, conclusions should always be tempered by the image of Figure 7.2 and the recognition that our test statistics were sampled from one like it. This is why real replications are crucial for science. These replications are best aggregated using meta-analyses. Consider, for example, the recent meta-analysis of body satisfaction among women of different ethnic groups (Grabe & Hyde, 2006). Ninety-three studies contributed data, whose median effect size was 0.30, indicating that Black women were more satisfied with their body image than were White women. The authors of the meta-analysis reported a random effects variance of 0.09, which is the average value Richard et al. (2003) reported for social-psychological literatures in general. Using that realization variance, the median p_{rep} for these studies was 0.78, and the proportion predicted by the 20% trimmed mean (Wilcox, 1998) of the $z_{p_{rep}}$ was $p_{rep} = .80$. The percentage of published studies showing a positive effect was close to these predictions, .85. These are unweighted predictions because the point is to demonstrate predictive validity of single estimates of p_{rep}, not to combine studies in an optimal fashion. Note that with realization variance set to zero, p_{rep} would seriously overestimate replicability ($p_{rep} = .93$), as expected.

Based on comparison of published and unpublished studies, as well as of studies that focused on ethnicity and those not so focused, the authors concluded that there was little evidence of publication bias—the inflation of effect sizes due to nonpublication of nonsignificant effects. Another way of estimating the influence of studies left in the "file drawer" because they did not yield a significant p value is to use the posterior predictive distribution to estimate the number of studies that should have reported nonsignificant effects. For the median number of observations in each

group of the meta-analysis, the effect size would have to exceed 0.36 for significance ($\alpha = .05$, assuming a fixed effect model with zero realization variance, as is typical in conducting tests of significance, and a one-tailed critical region). Integration of the posterior predictive distribution (Equation 3) from $d_L = -\infty$ up to $d_U = 0.36$ gives an expected number of studies with nonsignificant effects equal to 56; 53 were found in the literature. This corroborates the authors' inference of little publication bias. Assuming instead that authors of the original studies used two-tailed tests requires us to compare the expected and observed number of studies with effect sizes falling between ±0.43. Integration between these limits predicts 66 such results, 10 greater than the 56 observed. If 10 additional studies are inferred to reside only in file drawers because their effect sizes averaged around 0, when these are added to the 93 analyzed, the median effect size decreases from 0.30 to 0.27, leaving all conclusions intact. The same would be the case if this process were iterated, assuming the same proportion were filed for the larger hypostatized population. This further corroborates the authors' inference of little effect from publication bias.

A serious attempt to generalize—to predict replicability—must incorporate realization variance, just as traditional random and mixed effects models and hierarchical Bayesian models internalize it. But few original studies attempt to incorporate such estimates, resulting in discomfiture when results do not replicate (Ioannidis, 2005a, 2005b). This is compounded by the prevalent notion that "to replicate" means getting a significant effect in replication, rather than increasing our confidence in the original claim or narrowing our confidence intervals around an estimate. The posterior predictive distribution of Figure 7.1 and its integration over relevant domains by Equation 2 provide useful tools in the meta-analysis of results.

Deployment

This development of p_{rep} concerns only the simplest scenario, corresponding to a *t* test. Most inferential questions are more complicated. How does one calculate and interpret p_{rep} values in more interesting scenarios, such as analyses of variance (ANOVAs) with multiple levels, multivariate ANOVAs (MANOVAs), and multiple regression analysis? The provisional answer, given users' familiarity with traditional analyses, is to calculate a p value and transform it to p_{rep} using the standard conversion, $z(p_{rep}) = -z(p)/\sqrt{2}$. The p values returned from ANOVA stat-packs are analogous to t^2 and thus allow deviation among scores in either direction (even though they employ just the right tail of the F distribution). They should therefore be halved before converting to p_{rep}. In the case of multiple independent comparisons, the probability of replicating each of the observed effects equals the product of the constituent p_{rep}s. This may be contrasted with the probability of finding positive effects if the null was true in all k cases, 2^{-k}. Cortina and Nouri (2000) show how to adjust d for correlated measures, and Bakeman (2005) provides detailed recommendations for the use of generalized eta squared as the preferred effect size statistic for repeated-measure designs. Manly (1997) and Westfall and Young (1993) describe resampling techniques for multiple testing. It is trivial to adjust such resampling to calculate p_{rep} directly: (a) Resample from within the control data and independently from within the experimental data, (b) calculate the resampled statistics (e.g., mean difference, trimmed mean difference, etc.) over half the resampled numbers to double the variance for the predictive distribution, (c) count the number of statistics of the same sign as the measured test, and (d) divide by the total number of resamples to estimate p_{rep}.

Realization variance (σ_δ^2) inflates the sampling variance of the effect size. This is awkward to introduce into the resampling process, but an adjustment can be easily made after the fact. The variance of d in replication is approximately $8/(n-4) + 2\sigma_\delta^2$. The resampling operation is essentially a compound Bernoulli process that can, for these purposes, be approximated by the normal distribution. Compute the z score corresponding to the p_{rep} resampled as above with no allowance for realization variance, and divide it by

$$\sqrt{1 + \sigma_\delta^2 (n-4)/4}\,.$$

The normal transform of this adjusted z score gives the p_{rep} that can be expected in realistic attempts to replicate.[1]

Conversion of the voluminous statistical literature into p_{rep}-native applications remains a task for the future—one that will be most expeditiously and accurately accomplished with randomization techniques, discussed below.

Problems With P_{REP}

Frequentists will object to the introduction of distributions of parameters needed for the present derivation of p_{rep}. Parameters are by their definition fixed, if unknown, quantities. Frequentists will also be concerned that the choice of any particular informative prior can introduce an element of subjectivity into the calculation of probability. The introduction of uninformative priors, on the other hand, will expose the arbitrariness of choosing the particular version of them: Should an ignorance prior for the mass of a box be uniform on the length of its side or uniform on the cube of that length (Seidenfeld, 1979)? This debate has a long and nuanced history.

Bayesians (e.g., Wagenmakers & Grünwald, 2006) object that p_{rep} distracts us from an opportunity to compare alternative hypotheses. If credible alternate hypotheses are available, both the Neyman-Pearson framework and Bayesian analyses are to be preferred to the present one. But if those are not available, postulating them reintroduces the very sources of subjectivity and dependence on context-sensitive perspectives that p_{rep} permits us to sidestep.

An inelegance in the above analyses is that they invoked the unknown population parameter δ, only to marginalize it. Why not go directly from d_1 to d_2? As noted by O'Hagan and Forster (2004), "If we take the view that all inference is directly or indirectly motivated by decision problems, then it can be argued that all inference should be predictive, since inference about unobservable parameters has no direct relevance to decisions. . . . An extreme version of the predictivist approach is to regard parameters as neither meaningful nor necessary" (p. 90). Alas, as these authors, as well as Cumming (2005) and Doros and Geier (2005), note, unless d_1 and d_2 are treated as random samples from a population, they are not independent of one another, and the requisite evaluation of joint distributions of nonindependent variables is difficult. Fisher spent many of his latter years attempting to accomplish such direct inference with his fiducial probability theory, but his work was judged unsuccessful (Macdonald, 2005; Seidenfeld, 1979). We must resort to the introduction and marginalization of parameters described in the appendix.

More important than the above objections is the invalidity of a fundamental assumption in all of the above analyses. Traditional statistical tests, as well as p_{rep} as developed to this point, assume that the data are randomly sampled from the population to which generalization is desired—from the population whose parameter under the null (e.g., δ_0) is typically assumed to be zero. But this is a feat that is rarely attempted in science. Ludbrook and Dudley (1998; cited in Lunneborg, 2000, p. 551) surveyed 252 studies from biomedical journals and found that experimental groups were constructed by random sampling in only 4% of the cases. Of the 96% that were randomly assigned (rather than randomly sampled), 84% of the analyses employed inappropriate *t* or *F* tests—inappropriate because those tests assume random samples, not convenience samples. The situation is unlikely to be different in the fields known to the readers of this book. One can speculate why analyses are so misaligned with data. The potential causes are multivariate and include bad education, limitations to otherwise convenient stat-packs, the desire for statistics that permit generalization to a population (even though the data collection technique a fortiori prohibits such generalization), and the ubiquity of reviewers who recognize that, even though statistical tests are merely arbitrary conventional filters that cannot legitimately be used to reject null hypotheses, they retain pragmatic value for rejecting dull hypotheses (Nickerson, 2000).

Permutation Statistics

There is another way. It was introduced by Fisher, developed for general cases by Pitman (Pitman, 1937a, 1937b), and realized in a practical manner by modern randomization techniques. Randomization, or rerandomization, or permutation tests are ways of comparing distributions of scores. They do not, like their computerized siblings the bootstrap tests, attempt to estimate or conditionalize on population parameters. Instead, they ask how frequently a random reassignment of the observed scores would generate differences at least as large as those observed. In executing such tests, the observed scores are randomly reassigned (without replacement) to ad hoc groups, the relevant statistics (mean, trimmed mean, median, variance, *t* score, etc.) computed, another random reassignment executed, and so on, thousands of times to create

a distribution of the sampling error expected under chance. Calculate the percentile of the observed statistic in that distribution. One minus that percentile gives an analog to the *p* value. If the observed statistic is in the 95th percentile, only 5% of the time would random assignments give deviations that large or larger.

What the Permutation Distribution Means. In a deterministic world, no effects are uncaused, although the causes may be varied and complex and different for each observation (Hacking, 1990). *Chance* is the name for these otherwise unnamed, and generally unnamable, causes; it appears as the error variance that is added to our regression equations to permit them to balance. In resampling, we give the error variance free rein. The resulting randomization distributions may be viewed as random samples from thousands of these unnamed "hypotheses"—each corresponding to a different pattern in the data—that might account for the observations with more or less accuracy. The empirical distribution function generated by the random shuffles of data among groups gives the final rankings of hypotheses. All hypotheses except the investigator's are vague ("chance") and post hoc, so the investigator's are typically preferred, unless there are too many alternate reshuffles that sort the data into more extreme configurations—that is, unless they constitute more than, say, 5% of the distribution.

Another way of thinking about the partitioning is as the result of the random motion of particles (data) in space. This is a problem in thermodynamics, for analysis of which Boltzman created the measure of randomness called *entropy.* Using similar logic and mathematics, we may calculate the amount by which entropy is reduced by learning to which group—experimental or control—each observation belongs. If there is no real effect, the information transmitted by the group designation will be approximately 0. The information transmitted by knowledge of the group grows with effect size and with the logarithm of the number of observations. Information-theoretic measures such as Kullback-Leibler (K-L) distance, as well as its unbiased realization in the Akaike information criterion, are modern extensions of Boltzman's approach. The K-L distance is the average information that each observation adds toward discriminating the experimental and control groups. There is a natural affinity between permutation techniques and information-theoretic analyses: As the information gained from distinguishing groups increases, there will be a corresponding decrease in the number of alternate hypotheses that will provide more information than the investigator's. Replicability may be measured in terms of the probability of finding effects in replication that continue to make the distinction between groups worthwhile.

Although these considerations can give deeper meaning to the analysis, most consumers of statistics will be content to understand the results of permutation analyses as an analog of the *p* value, the proportion of randomizations that provide a more informative sorting of the observations than the experimenter's labels "experimental" and "control." There are many good programs available to carry out this analysis; see Cai (2006) and the references in Good (2000), Higgins (2004), and Manly (1997).

Permutation tests ask, "How often would this happen by chance?" not "How likely is this to happen again by design?" The short answer to "How can I generalize this result?" is the same as that given to users of traditional statistical design who have not sampled randomly from the populations to which they would generalize: "At your own hazard." To predict replicability in an attempt with an *n* of the same size, follow the same steps as given for bootstrap techniques above, including within-group shuffling, half-sizing, and correction for realization variance. For permutation techniques, however, the randomization is without replacement (whereas in bootstrapping, it is with replacement). The half-sizing makes this analysis a hybrid of permutation and bootstrap techniques—the bootstrap is constructed not out of the unlimited population of the bootstrap but out of populations twice the size of the original investigation. It permits extrapolation to a replication that samples from a small population derived from individuals identical to those in the original experiment, one from which the original was also ostensibly sampled. Because permutation techniques are generally more powerful than bootstrap techniques (see Mielke & Berry, 2001, and Chapter 19, this volume), predictions of replicability will be higher than for bootstrap techniques, making inclusion of nonzero realization variance even more important for realistic

projections. The same post hoc correction described above may be used: Divide the *z* score of replicability by

$$\sqrt{1 + \sigma_{\delta}^2 (n-4)/4}\,.$$

As noted by Higgins (2004), Lunneborg (2000), and others, computer-implemented permutation tests are the gold standard to which modern techniques such as ANOVA are an approximation; they are worth learning to use.

The Three Paths of Statistical Inference

Traditional Fisher/Neyman-Pearson statistics have been the primary mode of inference in the field for half a decade, despite the fact that "frequentist theory is logically inadequate for the task of uncertain estimation (it provides right answers to wrong questions), . . . the bridge from statistical technologies to actual working science remains sketchy" (Dempster, 1987, p. 2). Twenty years have not greatly changed that assessment. Littlewood, if you remember from the epigraph, did not see it as his "business to justify application of the system"—an Olympian view shared by some modern statisticians. Because NHST in particular cannot provide mathematical estimates of the probabilities of hypotheses, it is constitutionally unfit for deciding between null and alternate hypotheses. Introductory statistics texts should carry warning labels: "The Statistician General warns that use of the algorithms contained herein justifies no inferences about hypotheses and no generalization to populations unsampled. Their assumptions seldom match their applications. Their critical regions can displace critical judgments. Their peremptory authority can damage unborn hypotheses. Their punctilio shifts authority from scientists to stat-packs. Addictive."

Bayesian analysis (Chapter 33, this volume) is a step forward. It provides the machinery for deriving the posterior predictive distribution on which p_{rep} is based and on which a decision theory for science may be erected. It has a vigorous literature (e.g., Howson & Urbach, 1996; Jaynes & Bretthorst, 2003). As currently deployed, it is sometimes hobbled by continuing to maintain the frequentists' focus on parameters. It has been blamed for making probabilities subjective, but as long as reference sets are unique, probabilities must always be conditional on those priors.

Permutation techniques are another step forward, in that they more closely model the scientific process. Modern computer packages make them easy to implement and easier to teach than traditional statistical pedagogy based on Fisher/Neyman-Pearson inference.

What this chapter hopes to convey is that by setting our sights a bit lower—down from the heavens of Platonic parameters to the earthier Aristotelian enterprise of predicting replicability—our inferences will be simpler and more useful. Much work needs yet to be accomplished: generalizing replicability statistics to cases with multiple degrees of freedom in the numerator, securing their implementation with permutation tests, and utilizing those tests for predicting replicability in contexts involving substantial realization variance. But accomplishing these tasks should be straightforward, and their execution will bring us closer to the fundamental task of scientists: to validate observations through prediction and replication.

Appendix

The Statistics of Effect Size

Pooled variance is

$$s_p^2 = \frac{s_C^2 (n_C - 1) + s_E^2 (n_E - 1)}{n - 2}. \qquad \text{(A1)}$$

The Route to Posterior Predictive Distributions

Bayesian statistics provides a standard way to calculate $f(d_2|d_1)$(see, e.g., Bolstad, 2004; Winkler, 2003): Predicate the unknown parameter, such as the population mean effect size δ; update that predication with the observed data; and then calculate the posterior predictions over all possible values of the parameter, weighted by the probability of the parameter given the observed data. The predication is a nuisance, and eliminating the nuisance parameter by integrating it out in the last step is called marginalization.

$$f(d_2|d_1) = \int f(d_2, \delta|d_1)\, d\delta$$
$$= \int f(d_2|\delta,d_1)\, f(\delta|d_1)\, d\delta$$
$$= \int f(d_2|\delta,d_1)\, f(d_1|\delta)\, f(\delta) d\delta$$

where $f(\delta|d_1)$ is the posterior distribution of the parameter in light of the observations, and $f(\delta)$ is the prior distribution of the parameter. Integration of the last line over all population parameters δ delivers the posterior predictive distribution. If $f(\delta)$ is assumed to have a very large variance (*flat* or *ignorance* priors), then the observed data dominate the result. If a credible prior distribution is available (it is generally sought by "empirical Bayesians"), then the final predictions of replicability will be more accurate.

Consider the case in which the prior on the population mean δ is normally distributed. Then its posterior $f(\delta|d_1)$ is $n(d_1', s_1'^2)$, where the primed variables are weighted averages of the priors and the observed statistics, with the weights proportional to the precisions (reciprocal variances) of the means (s_{prior}^{-2} and s_1^{-2}). If we are relatively ignorant a priori of the value of the parameter, its distribution is flat relative to that of the observed statistic ($s_{\text{prior}}^2 >> s_1^2$), and then $d_1' \approx d_1$ and $s_1' \approx s_1$. This is the case developed here. When the sampling distribution of the statistic is also approximately normal—a reasonable assumption for measures of effect size even when n is relatively small (Hedges & Olkin, 1985)—then factoring, completing the square, and removing constants leads eventually (Bolstad, 2004; Winkler, 2003) to a normal density with mean d_1 and variance $s_{dR}^2 = s_1^2 + s_2^2 = 2s_1^2$:

$$f(d_2|d_1) \propto e^{-\frac{1}{2\left(s_2^2+s_1^2\right)}(d_2-d_1)^2}.$$

Integration of this between appropriate limits, as in Equations 2 and 3, leads to the central results of this chapter.

The Simulations

In the simulations for Figures 7.6 and 7.7, d' was calculated using Equations 1 and A1. Variance was calculated using Equation 5 and

$$s_d^2 \approx \frac{4}{n-4}.$$

All simulations in this chapter used Resampling Stats© software (Bruce, 2003), which creates random numbers with Park and Miller's (1988) "Real Version 1" multiplicative linear congruential algorithm.

NOTE

1. The half-sizing recommended in Step b can be bypassed by replacing the 1 in this radical with 2, which allows for the increased variance inherent in the posterior predictive distribution.

REFERENCES

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37,* 379–384.

Bolstad, W. M. (2004). *Introduction to Bayesian statistics.* Hoboken, NJ: John Wiley.

Bruce, P. (2003). Resampling Stats [Excel Add-in]. Arlington, VA: Resampling Stats, Inc.

Cai, L. (2006). Multi-response permutation procedure as an alternative to the analysis of variance: An SPSS implementation. *Behavior Research Methods, 38,* 51–59.

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

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs.* Thousand Oaks, CA: Sage.

Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science, 16,* 1002–1004.

Cumming, G. (2006). *ESCI.* Retrieved from http://www.latrobe.edu.au/psy/esci/

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–575.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60,* 170–180.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3,* 299–311.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application.* New York: Cambridge University Press.

Dempster, A. P. (1987). Probability and the future of statistics. In I. B. MacNeill & G. J. Umphrey (Eds.), *Foundations of statistical inference* (Vol. 2, pp. 1–7). Dordrecht, Holland: D. Reidel.

Doros, G., & Geier, A. B. (2005). Comment on "An alternative to null-hypothesis significance tests." *Psychological Science, 16,* 1005–1006.

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

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*, 119–126.

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

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-economics, 33*(5), 587–606.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11*, 791–806.

Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (2nd ed.). New York: Springer-Verlag.

Grabe, S., & Hyde, J. S. (2006). Ethnicity and body satisfaction among women in the United States: A meta-analysis. *Psychological Bulletin, 132,* 622–640.

Hacking, I. (1990). *The taming of chance.* New York: Cambridge University Press.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.

Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart and Winston.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486–504.

Higgins, J. J. (2004). *An introduction to modern nonparameteric statistics.* Pacific Grove, CA: Brooks/Cole.

Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (2nd ed.). Chicago: Open Court.

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

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Iverson, G., Myung, I. J., & Karabatsos, G. (2006, August). *P-rep, p-values and Bayesian inference.* Paper presented at the Society for Mathematical Psychology, Vancouver, BC.

Jaynes, E. T., & Bretthorst, G. L. (2003). *Probability theory: The logic of science.* Cambridge, UK: Cambridge University Press.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5,* 411–414.

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, 16,* 1009–1012.

Killeen, P. R. (2005c). Tea-tests. *The General Psychologist, 40*(2), 16–19.

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review, 13,* 549–562.

Kline, M. (1980). *Mathematics, the loss of certainty.* New York: Oxford University Press.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kyburg, H. E., Jr. (1987). The basic Bayesian blunder. In I. B. MacNeill & G. J. Umphrey (Eds.), *Foundations of statistical inference: Vol. 2. Biostatistics* (pp. 219–232). Boston: D. Reidel.

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

Lee, P. M. (2004). *Bayesian statistics: An introduction* (3rd ed.). New York: Hodder/Oxford University Press.

Littlewood, J. E. (1953). *A mathematician's miscellany.* London: Methuen.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in medical research. *American Statistician, 52,* 127–132.

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

Macdonald, R. R. (2005). Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science, 16*(12), 1007–1008.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology.* New York: Chapman & Hall/CRC.

Maurer, B. A. (2004). Models of scientific inquiry and statistical practice: Implications for the structure of scientific knowledge. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 17–50). Chicago: University of Chicago Press.

Mielke, P. W., & Berry, K. J. (2001). *Permutation methods: A distance function approach.* New York: Springer.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference* (Vol. 95). Newbury Park, CA: Sage.

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

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics: Vol. 2B. Bayesian inference* (2nd ed.). New York: Oxford University Press.

Park, S. K., & Miller, K. W. (1988). Random number generators: Good ones are hard to find. *Communications of the ACM, 31*(10), 1192–1201.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society, 4*(1), 119–130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations: II. The correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society, 4*(2), 225–232.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331–363.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.

Royall, R. (2004). The likelihood paradigm for statistical evidence. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 119–152). Chicago: University of Chicago Press.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics, 6,* 377–400.

Sanabria, F., & Killeen, P. R. (2007). Better statistics for better decisions: Rejecting null hypotheses statistical tests in favor of replication statistics. *Psychology in Schools, 44,* 471–481.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.

Seidenfeld, T. (1979). *Philosophical problems of statistical inference: Learning from R. A. Fisher.* London: D. Reidel.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory Psychology, 9*(2), 165–181.

van den Noortgate, W., & Onghena, P. (2003). Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods. *Behavior Research Methods, Instruments, & Computers, 35,* 504–511.

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing: A comment on Killeen (2005). *Psychological Science, 17,* 641–642.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for* p*-value adjustments.* New York: John Wiley.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

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

Winkler, R. L. (2003). *An introduction to Bayesian inference and decision* (2nd ed.). Gainesville, FL: Probabilistic Publishing.