Journal of Parapsychology, Vol. 52, December 1988

SUCCESSFUL REPLICATION VERSUS STATISTICAL SIGNIFICANCE

BY JESSICA UTTS

ABSTRACT: The aim of this paper is to show that successful replication in parapsychology should not be equated with the achievement of statistical significance, whether at the .05 or at any other level. The *p* value from a hypothesis test is closely related to the size of the sample used for the test; so a definition of successful replication based on a specific *p* value favors studies done with large samples. Many "nonsignificant" studies may simply be ones for which the sample size was not large enough to detect the small magnitude effect that was operating. Conversely, "significant" studies may result from a small but conceptually insignificant bias, magnified by a very large sample.

The paper traces the history of the definition of statistical significance in parapsychology and then outlines the problems with using hypothesis-testing results to define successful replications, especially when applied in a cookbook fashion. Finally, suggestions are given for alternative approaches to looking at experimental data. These include calculating statistical power before doing an experiment, using estimation instead of, or in conjunction with, hypothesis testing, and implementing some of the ideas from Bayesian statistics.

Replication is a major issue in parapsychology. Arguments about whether a given research paradigm has been successful tend to focus on what the replication rate has been. For example, the recent review of parapsychology by the National Research Council includes statements such as "... of these 188 [RNG] experiments with some claim to scientific status, 58 reported statistically significant results (compared with the 9 or 10 experiments that would be expected by chance)" (Druckman & Swets, 1988, p. 185). In each section, the report critically evaluates "significant" experiments and ignores "nonsignificant" experiments. The extent to which nonsignificant experiments are ignored is exemplified by the following oversight, in which the total number of studies is equated with the number of "successful" studies: "Of the *thirteen* scientifically reported experiments [of remote viewing], *nine* are classified as successful in their outcomes by Hansen et al. ... As it turns out, all but one of the *nine* scientifically reported studies of remote viewing suffer from the flaw of sensory cueing" (p. 183, emphasis added). Apparently the authors decided that the four experiments that did not attain a *p* value of .05 or less did not even warrant acknowledgment.

The practice of defining a successful replication as an experiment that attains a p value of .05 or less is common in parapsychology, psychology, and some other disciplines that use statistics. However, like many other conventions in science, it is based on a series of historical events rather than on rational thought. In this paper, I will trace some of the history leading to this definition of a "successful" experiment, outline some problems with this approach, and suggest some methods that parapsychologists should consider in addition to the usual hypothesis-testing regimen. Rao (1984) and Honorton (1984) have discussed similar problems and solutions in the context of psi experiments.

History

It has not always been the case among parapsychologists that an experiment was deemed successful if it reached a significance level of $p = .05$. In 1917, John Edgar Coover, who was the Thomas Welton Stanford Psychical Research Fellow at Stanford University from 1912 to 1937, published a book with the results from several experiments he had conducted up to that time (Coover, 1917/1975). Although hypothesis testing as we know it today had not yet been formalized, he essentially conducted tests on many facets of this data and found no evidence for psi that was convincing to him. His conclusions regarding these results are typified by an example he gave in which the hit rate for 518 trials was 30.1%, when 25% was expected by chance (exact p value = .00476):

> We get 0.9938 [p-value = 1 − 0.9938 = 0.0062] for the probability that chance deviations will not exceed this limit [of 30.1 percent].... Since this value, then, lies within the field of chance deviation, although the probability of its occurrence by chance is fairly low, it cannot be accepted as a decisive indication of some cause beyond chance which operated in favor of success in guessing. (p. 82)

He then revealed what level of evidence would convince him that nonchance factors were operating: "... if we meet the requirement of a degree of accuracy usual in scientific work by making P = 0.9999779, when absolute certainty is P = 1, then [there is] satisfactory evidence for some cause in addition to chance" (p. 83). In other words, he was defining significance with a p value of 2.21×10^{-5}.

Coover was not alone in requiring that results conform to arbitrarily stringent significance levels. In 1940, when Rhine et al. pub-

lished *Extra-Sensory Perception After Sixty Years,* they included the following definitions in the glossary:

> *p*-value = probability of success in each trial
>
> SIGNIFICANCE: When the probability that chance factors alone produced a given deviation is sufficiently small to provide relative certainty that chance is not a reasonable expectation, the deviation is *significantly* above or below the chance level. Among ESP results, this is arbitrarily taken to mean a deviation in the expected direction such that the critical ratio is 2.5 times the standard deviation (or four times the probable error) or greater. (p. 423–424)

Thus, significance was defined by $z \geq 2.5$, or $p \leq .0062$.

Seventeen years later, in their book *Parapsychology: Frontier Science of the Mind,* Rhine and Pratt (1957) suggested that .01 was the appropriate threshold:

> In order for such judgments to have the necessary objectivity, a *criterion of significance* is established by practice and general agreement among the research workers in a particular field.... Most workers in parapsychology accept a probability of .01 as the criterion of significance. (p. 186)

Finally, the *Journal of Parapsychology* has included a definition of *significance* in its glossary for many years, but the appropriate *p* value has fluctuated back and forth between .01 and .02, finally settling at .02 in 1968. The following are excerpts from those glossaries:

> December 1949: "A numerical result is significant when it equals or surpasses some criterion of degree of chance improbability. Common criteria are: a probability value of .01 or less."
>
> March 1950 to June 1957: "The criterion commonly used in this Journal is a probability value of .02 or less."
>
> September 1957: "The criterion commonly used in this Journal is P = .01."
>
> December 1957 to December 1967: "The criterion commonly used in parapsychology today is a probability value of .01 or less."
>
> March 1968 to December 1986: "The criterion commonly used in parapsychology today is a probability value of .02 (odds of 50 to 1 against chance) or less.... Odds of 20 to 1 (probability of .05) are regarded as strongly suggestive."

March 1987: The term *significance* no longer appears in the glossary.

By the mid-1980's, despite the value of .02 given in the *Journal of Parapsychology,* significance seemed to have been determined to correspond to a p value of .05. For example, in their bibliography of remote-viewing research, Hansen, Schlitz, and Tart (1984) claim: "We have found that more than half (fifteen out of twenty-eight) of the published formal experiments have been successful, where only one in twenty would be expected by chance." As mentioned in my introduction, .05 was the value used by the National Research Council in their recent evaluation of parapsychology. Both Hyman (1985) and Honorton (1985) used .05 as the criterion for a successful ganzfeld study. In discussing the Schmidt REG experiments, Palmer (1985) implicitly used .05 as the cut-off for significance by observing: "Based on Z-tests ... 25 of the 33 (76%) were significant at the .05 level, two-tailed. In two of the seven non-significant studies...." (p. 102).

This definition of significance is obviously not unique to parapsychology. A popular introductory textbook in psychology states that:

> Psychologists used a statistical inference procedure that gives them an estimate of the probability that an observed difference could have occurred by chance. This computation is based on the size of the difference and the spread of the scores. By common agreement, they accept a difference as "real" when the probability that it might be due to chance is less than 5 in 100 (indicated by the notation $p < .05$). A **significant difference** is one that meets this criterion.... With a statistically significant difference, a researcher can draw a conclusion about the behavior that was under investigation. (Zimbardo, 1988, p. 54)

Given the weight that has been attached to .05 as the criterion for significance, one would think that it resulted from careful consideration of the issue by statisticians and psychologists. Unfortunately, such is not the case. Its roots apparently lie in the following passage published in 1926 by one of the founders of modern statistics, Sir Ronald A. Fisher:

> It is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials." ... If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a

> low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. (Fisher, 1926, p. 504; also quoted in Savage, 1976, p. 471)

Thus began the belief that an experiment is successful only if the null hypothesis can be rejected using $\alpha = 0.05$. As an immediate consequence of this belief, Fisher and his followers created tables of F statistics that included values only for tail areas of .05 and .01. Since researchers did not have access to computer algorithms to determine intermediate p values, success came to be measured in terms of these two values alone.

Problems with Hypothesis Testing

Misconceptions about p Values

Most modern research reports include p values instead of simply discussing whether an experimental result is significant at a prespecified level. Although this is somewhat better than the old method of "one star or two" (corresponding to a significant result at .05 or .01, respectively), it is still a misleading way to examine experimental results.

The problem is that many researchers interpret p values as being related to *the probability that the null hypothesis is true.* Even some sophisticated researchers tend to think that an extremely small p value must correspond to a very large effect in the population and that a large p value (say $> .10$) means that there is no effect. In other words, *the size of the p value is incorrectly interpreted as the size of the effect.* It should be interpreted as the probability of observing results as extreme or more so than those observed, *if* there is no effect.

To see how arbitrary it is to base a decision about the truth or falsity of a statement on a p value, consider a binomial study based on a sample of size n which results in $z = 0.30$, p value $= .38$, one-tailed. One would probably abandon the hypothesis under study and decide not to pursue the given line of research. Now suppose that the study had been run with a sample of size $100n$ instead and resulted in the exact same proportion of hits. Then we would find $z = 3.00$, p value $= .0013$. These results would be regarded as highly significant!

As another example, consider a chi square test for randomness based on a sequence of n numbers, each of which can take the values 1, 2, ... 10. Suppose that the test results in a chi-square value of 11.0, $df = 9$, p value = 0.28. Now suppose the sequence was three times as long but the proportions of each digit remained the same. Then each term in the numerator of the chi-square statistic would be multiplied by 3^2, whereas each term in the denominator would only be multiplied by 3. The degrees of freedom would not change, but the new result would be $\chi^2 = 33.0$, $df = 9$, p value = .00013. In the first case, the conclusion would be that the sequence was sufficiently random, yet a sequence three times as long with the same pattern would be seen to deviate considerably from randomness!

This problem was recognized more than 50 years ago by Berkson (1938):

> We may assume that it is practically certain that any series of real observations does not actually follow a normal curve *with absolute exactitude* in all respects, and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square P will be small if the sample has a sufficiently large number of observations in it.
>
> If this be so, then we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one, but since the result of the former test is known, it is no test at all. (pp. 526–527, emphasis in original)

Replication

Very often researchers simply do not understand the connection between the p value and the size of the sample. For example, Rosenthal and Gaito (1963) asked nine faculty members and ten graduate students in a university psychology department to rate their degree of belief or confidence in results of hypothetical studies with various p values and with sample sizes of 10 and 100. Given the same p value, one should have more confidence in a study with a *smaller* sample because it would take a larger underlying effect to obtain the small p value for a small sample. Unfortunately, these

respondents demonstrated that they were far more likely to believe results based on the large sample when the p values were the same. (For a discussion of this example and some other problems with hypothesis testing in psychology, see Bakan, 1967.)

One consequence of this misunderstanding is that researchers misinterpret what constitutes a "successful replication" of an experiment. Tversky and Kahneman (1982) asked 84 members of the American Psychological Association or the Mathematical Psychology Group the following question:

> Suppose you have run an experiment on 20 subjects, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group? (p. 23)

The median answer given was .85. Only 9 of the 84 respondents gave an answer between .40 and .60. Assuming that the value obtained in the first test was close to the true population value, the probability of achieving a p value $\leq .05$ on the second test is actually only about .47. This is because the sample size in the second study is so small. The effect would have to be quite large in order to be detected with such a small sample.

In the same survey, Tversky and Kahneman also asked:

> An investigator has reported a result that *you* consider implausible. He ran 15 subjects, and reported a significant value, $t = 2.46$. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data. You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate? (p. 28)

The majority of respondents considered $t = 1.70$ as a failure to replicate. But if the results from both studies are combined, then (assuming equal variances) the result is $t = 2.94$, $df = 29$, p value $= .003$. The paradox is that the new study *decreases* faith in the original result if viewed separately but *increases* it when combined with the original data!

This misunderstanding about replication is quite prevalent in the psi literature, as demonstrated by the emphasis on successful replication, where success is defined in terms of a specific p value, regardless of sample size. As an example of how unnecessarily discouraging this can be for researchers, I have shown elsewhere (Utts,

1986) that if the true hit rate in a binomial study (such as a ganzfeld experiment) is actually 33%, and 25% is expected by chance, then a study based on a sample of size 26 should be expected to be "successful" ($p \leq .05$) only about one fifth of the time. Even a study based on a sample of size 100 should be "successful" only about half of the time. It is no wonder that there are so many "unsuccessful" attempts at replication in psi.

As another example of the paradoxical nature of this definition of replication, consider the "unsuccessful" direct-hit ganzfeld studies covered by the meta-analyses of Hyman (1985) and Honorton (1985). Using those studies with p(hit) = .25, there were 13 out of 24 that were nonsignificant, $\alpha = 0.05$, one-tailed. (See Honorton, p. 84, Table A1.) But when these 13 "failures" are combined, the result is 106 hits out of 367 trials, $z = 1.66$, $p = .0485$!

Problems with Point Null Hypotheses

A point null hypothesis is one that specifies a particular value ("point") as the one being tested. Most hypothesis testing is done with point null hypotheses. The problem with this approach is that any given hypothesis is bound to be false, even if just by a minuscule amount. For example, in a coin-tossing experiment, the null hypothesis is that the coin is fair, that is to say, H_o: $P = .5000000$. This is never precisely true in nature. All coins and coin-tossers introduce a slight bias into the experiment. This slight bias can produce a very small p value if the sample size is large enough. If, for example, the true probability of heads is .5001, and the observed proportion of heads falls right at this value, then the null hypothesis will be rejected at .05 if the sample size is at least 6.7×10^7. As long as there is any bias *at all*, the p value can be made arbitrarily small by taking a large enough sample.

In practice, this problem was rarely serious before it became possible to collect large amounts of data rapidly using computers. Statisticians have often used ESP as an example of one of the few cases where it really is possible to specify an exact value for the null hypothesis. But even this view is changing, as shown by this comment from a recent issue of a popular statistics journal:

> It is rare, and perhaps impossible, to have a null hypothesis that can be exactly modeled as $\theta = \theta_o$. One might feel that hypotheses such as
>
> H_o: A subject has no ESP, or
>
> H_o: Talking to plants has no effect on their growth,
>
> are representable as exact (and believable) point nulls, but, even here,

> minor biases in the experiments will usually prevent exact representations as points. (Berger & Delampady, 1987, p. 320)

In summary, hypothesis testing as it is currently formulated tends to be a misleading approach to examining data. Small samples tend to lead to "nonsignificant" studies, whereas large samples can lead to extremely small p values, even if the null hypothesis is only slightly wrong. Many researchers do not understand the meaning of a p value and do not understand how closely replication issues are tied to sample size. Arguments about replication should not be based on p values alone.

SOLUTIONS

Power Calculations

If a hypothesis test is to be done at all, a researcher should at least determine in advance whether it is likely to be successful. The statistical power of a test is the probability that the null hypothesis will be rejected. It obviously depends on what the true underlying state of nature is. Because this information cannot be known (or there would be no point in doing the experiment), it is a good idea to look at power for a variety of possibilities *before* conducting the experiment. The results will tell you whether you are likely to be able to reject the null hypothesis, using the sample size you have planned, for specific values of the magnitude of the effect.

Statistical power is a function of the sample size, the true underlying magnitude of the effect, the level of significance for which the experiment would be considered a success, and the method of analysis used. It does not depend on the data.

As an example, suppose you are planning to conduct a test of the hypothesis H_o: $P = .25$ using a series of 10 independent trials. Power calculations would proceed as follows:

1. Find the cutoff point for the number of hits that would lead to rejection of H_o. In this case, the p value for 5 hits is .08, and for 6 hits it is .02, so 6 hits would probably be required to reject the null hypothesis.

2. Power for a specific alternative is the probability that the null hypothesis would be rejected if that alternative value is true. In this case, power = P(6 or more hits). This can be computed directly, using the binomial formula, for any specified hit rate. Here are some examples:

Hit rate	Power = P(6 or more hits)
0.30	.047
0.33	.073
0.40	.166
0.50	.377

Notice that even if the true hit rate is 50% instead of the chance level of 25%, the chances of a "successful" replication are poor, that is, only 37.7%. In most psi applications, 30% or 33% is probably a more realistic approximation to the true hit rate, so there would be a very small chance of having this experiment succeed with only 10 trials.

As a second example, suppose you are planning to run the same experiment with 100 trials and are planning to use the normal approximation instead of an exact test. Further, suppose you will reject the null hypothesis if $z \geq 1.645$, where z is the usual critical ratio, corrected for continuity: z = (number of hits − 0.5 − 25) / $\sqrt{(100 \times .25 \times .75)}$ = .23(number of hits − 25.5). Using simple algebra, note that $z \geq 1.645$ when the number of hits $\geq$ 32.65. Thus, the null hypothesis will be rejected if there are 33 or more hits, so power = P(33 or more hits). Computing this for the same hypothetical hit rates as in the previous example gives:

Hit rate	Power = P(33 or more hits)
0.30	.289
0.33	.538
0.40	.939
0.50	.9998

Now there is a more reasonable chance for a successful study, although it is still only 29% even if the true hit rate is 30%.

For studies in which the null hypothesis does not involve a single value, it can be more difficult to compute power because it is not so easy to specify a reasonable alternative. In these cases, it is still possible to look at the p value that can be expected if psychic functioning were to occur at specified levels for the sample size planned. For example, McClenon and Hyman (1987) conducted a remote-viewing study with eight trials, one for each of eight subjects, and used the preferential-ranking method of Solfvin, Kelly, and Burdick (1978) on the subject rankings. Each subject was asked to rank-order eight

choices of potential targets as compared to the response he or she had produced. By chance, the average rank should be 4.5. If psychic functioning had reduced the average rank to 4.0, the p value would have been .298, *not significant*. Even if the average rank had been reduced to 3.5, the study would still not have been significant, p value = .126. The average rank would have to be 3.0 before this study would achieve a significant result. A parapsychologist experienced in remote viewing should be able to determine in advance whether such a study would be likely to be successful with such a small sample.

The lesson here is that a "nonsignificant" study may be nothing more than a study with low power. Before investing time and money in a new study, it should be determined whether it is likely to succeed if psychic functioning is operating at a given level.

Estimation

An approach that avoids many of the problems with hypothesis testing is to construct a "confidence interval" or an "interval estimate" for the magnitude of an effect. This is done by computing an interval of values that almost certainly covers the true population value. The degree of certainty is called the *confidence coefficient* and is specified by the researcher. Common values are 95% and 99%.

As an example, consider a binomial study with 100 trials that results in 35 hits. Using the normal approximation, one would expect the proportion of hits in the sample to be within 1.96 standard deviations of the true hit rate 95% of the time. The appropriate standard deviation for the proportion P of hits is $\sqrt{P(1-P)/n}$. Thus, a 95% confidence interval for the true hit rate is found by adding and subtracting 1.96 of these standard deviations to the proportion of hits observed in the sample. The resulting interval in this case is 0.35 − 0.09 to 0.35 + 0.09, or 0.26 to 0.44. This tells us that with a fair amount of certainty (95%), the true hit rate is covered by the interval from 0.26 to 0.44. For the same proportion of hits in a study with 1,000 trials, the interval would be from 0.32 to 0.38. The larger the sample size, the shorter the width of the interval.

Consider two studies designed to test H_o: $P = .5$:

	Study 1	Study 2
z	3.60	2.40
p value	.0004	.0164
n	1,000	100

Which study provides more convincing evidence that there is a strong effect? In keeping with the results of Rosenthal and Gaito (1963) discussed earlier, most people would say that the first study shows a stronger effect, both because the p value is smaller and because it is based on a larger sample. In fact, the opposite is true. The number of hits for the two studies are 557 (55.7%) and 62 (62%), respectively; the smaller study had a higher hit rate. The 95% confidence intervals for the hit rates in the two studies are (0.53 to 0.59) and (0.53 to 0.72), respectively, so in both studies we are relatively sure that the hit rate is at least 53%, but in the second study it could be as high as 72% whereas in the first it is probably no higher than 59%.

In studies with huge sample sizes, confidence intervals make it evident that an infinitesimal p value does not correspond to an effect of large magnitude. For example, consider a study based on 100,000 trials and designed to test H_o: $P = .50$. Suppose there were 50,500 hits. Then $z = 3.16$, and the p value is 7.9×10^{-4}. But what does this mean in practical terms? A 95% confidence interval for the true hit rate is from 0.5019 to 0.5081. Thus, it appears that the true hit rate is indeed different from 0.50, but reporting the results in this way makes it clear that the magnitude of the difference is very small. The reader can decide whether an effect of this size has any meaning in the context of the experiment.

In summary, confidence intervals are preferable to hypothesis tests for the following reasons:

1. They show the *magnitude* of the effect.
2. They show that the accuracy of the conclusion is highly dependent on the sample size.
3. They remove the focus from decision making, which is arbitrary at best because of sample size problems.
4. They highlight the distinction between *statistical* significance and *practical* significance.
5. They allow the reader of a research report to come to his or her own conclusion.

Meta-Analyses

Meta-analytic techniques may be viewed by some parapsychologists as the solution to studying the issue of replication. Even though these techniques can address the replication issue in useful ways,

they also contain some dangerous pitfalls. For example, both Hyman (1985) and Honorton (1985) used "vote-counting" in their meta-analyses of the ganzfeld data base. In other words, they tallied the number of significant studies in the data base. This procedure inherits all of the problems associated with the original determination of whether a study was "significant" in the first place. A series of studies, each with low power, may all be determined to be nonsignificant, when the combined data may lead to an extremely significant result. Conversely, a series of studies based on large samples may all be significant, but the magnitude of the effect may be very small. A vote-count showing that most studies are significant could mislead researchers into believing that there was a large effect.

The concept of effect size was introduced to account for the fact that individual study results are highly dependent on sample size. Estimating the effect sizes for a series of studies and seeing whether they are similar is a useful way of studying replication. However, examining only the effect size for an individual study does not give any indication of the accuracy of the result. This should be done in conjunction with some estimate of the accuracy of the result, such as a confidence interval.

Bayesian Methods

Many statisticians believe that the conceptual framework of hypothesis testing and interval estimation is philosophically incorrect. Rather, they start by assigning prior probabilities, based on subjective belief, to various hypotheses, and then combine these "priors" with the data to compute final or "posterior" probabilities for the hypotheses. This is called the Bayesian approach to statistics. An introduction to the ideas of Bayesian analysis can be found in Berger and Berry (1988) or Edwards, Lindman, and Savage (1963). A more technical reference is Berger (1985).

Berger and Berry (1988), in a recent article in *American Scientist,* discussed the use of Bayesian methods instead of classical methods:

> The first step of this demonstration is to calculate the actual probability that the hypothesis is true in light of the data. This is the domain of Bayesian statistics, which processes data to produce "final probabilities" ... for hypotheses. Thus, the conclusion of a Bayesian analysis might be that the final probability of H is 0.30.
>
> The direct simplicity of such a statement compared with the convoluted reasoning necessary to interpret a P-value is in itself a potent ar-

> gument for Bayesian methods. Nothing is free, however, and the elegantly simple Bayesian conclusion requires additional input. To obtain the final probability of a hypothesis in light of the experimental data, it is necessary to specify the probability of the hypothesis before or apart from the experimental data.
>
> Where does this initial probability come from? The answer is simple. It must be subjectively chosen by the person interpreting the data. A person who doubts the hypothesis initially might choose a probability of 0.1; by contrast, someone who believes in it might choose 0.9. (p. 162)

They then provide an example of testing the hypothesis H: $P = .5$, where P is the proportion of hits expected in a binomial experiment. Suppose that in 17 trials there are 13 successes (76.5%). Then the p value is .049, two-tailed. Unless, of course, the experiment was designed to stop at the fourth failure instead of at the 17th trial. Then the p value, with the identical data, would only be .021. Such problems arise with classical methods, but not with Bayesian methods.

Using the Bayesian approach, suppose that one's prior belief that H is true is 50%. If H isn't true, the prior belief is that the true value of P is equally likely to be anywhere between $0.5 - c$ and $0.5 + c$ (where c is some constant), but could not possibly be farther than that from 0.5. The choice of c represents prior opinion about the strength of the effect, if there is one. Choosing $c = 0.1$ (the effect isn't likely to be very strong even if it exists) results in a final probability of 0.41 for H (given that there were 13 successes in 17 trials), whereas choosing $c = 0.4$ results in a probability of 0.21 for H. In other words, the final degree of belief in H is dependent on one's prior belief about the strength of the effect. It also depends on prior opinion about the veracity of H, and on the observed data.

One reason that Bayesian methods are not more widely used is that they are often difficult to apply. Another reason is that researchers are uncomfortable with having to specify subjective degrees of belief in their hypotheses. This approach makes particular sense for parapsychology, however, because most researchers have strong opinions about the probability that psi is real, and these opinions play a central role in how psi researchers and critics evaluate the evidence. Posterior probabilities in Bayesian analyses are a function of both the prior probabilities and the strength of the evidence; it may be informative to formalize these opinions and to see how much evidence would be needed to increase the posterior probability of a psi hypothesis to a non-negligible level when the prior probability was close to zero.

References

BAKAN, D. (1967). *On method: Toward a reconstruction of psychological investigation.* San Francisco: Jossey-Bass, Inc.

BERGER, J. O. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag.

BERGER, J. O., & BERRY, D. A. (1988, March–April). Statistical analysis and the illusion of objectivity. *American Scientist,* pp. 159–165.

BERGER, J. O., & DELAMPADY, M. (1987). Testing precise hypotheses. *Statistical Science,* **2**(3), 317–334.

BERKSON, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association,* **33,** 526–542.

COOVER, J. E. (1975). *Experiments in psychicial research.* New York: Arno Press. (Originally published 1917)

DRUCKMAN, D., & SWETS, J. A. (1988). *Enhancing human performance.* Washington, D. C.: National Academy Press.

EDWARDS, W., LINDMAN, H., & SAVAGE, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review,* **70,** 193–242.

FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain,* **33,** 503–513.

HANSEN, G. P., SCHLITZ, M. J., & TART, C. T. (1984). Bibliography, remote-viewing research, 1973–1981. In R. Targ & K. Harary, *The mind race* (pp. 265–269). New York: Villard Books.

HONORTON, C. (1984). How to evaluate and improve the replicability of parapsychological effects. In B. Shapin & L. Coly (Eds.), *The repeatability problem in parapsychology* (pp. 238–255). New York: Parapsychology Foundation, Inc.

HONORTON, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology,* **49,** 51–91.

HYMAN, R. (1985). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology,* **49,** 3–49.

MCCLENON, J., & HYMAN R. (1987). A remote viewing experiment conducted by a skeptic and a believer. *Zetetic Scholar,* Nos. 12/13, 21–33.

PALMER, J. (1985). An evaluative report on the current status of parapsychology. U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

RAO, K. R. (1984). Replication in conventional and controversial sciences. In B. SHAPIN & L. COLY (Eds.), *The repeatability problem in parapsychology,* (pp. 22–41). New York: Parapsychology Foundation, Inc.

RHINE, J. B., & PRATT, J. G. (1957). *Parapsychology: Frontier science of the mind.* Springfield IL: Charles C. Thomas.

RHINE, J. B., PRATT, J. G., STUART, C. E., SMITH, B. M., & GREENWOOD, J. A. (1940). *Extra-sensory perception after sixty years.* Boston: Bruce Humphries.

ROSENTHAL, R., & GAITO, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology,* **55,** 33–38.

SAVAGE, L. J. (1976). On rereading R. A. Fisher. *Annals of Statistics,* **4,** 441–500.

SOLFVIN, G. F., KELLY, E. F., & BURDICK, D. S. (1978). Some new methods for preferential-ranking data. *Journal of the American Society for Psychical Research,* **72,** 93–110.

TVERSKY, A., & KAHNEMAN, D. (1982). Belief in the law of small numbers. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

UTTS, J. M. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology,* **50,** 393–402.

ZIMBARDO, P. G. (1988). *Psychology and life.* Glenview, IL: Scott, Foresman and Co.

Division of Statistics
University of California
Davis, CA 95616