FIELD ESSAY

The Insignificance of Null Hypothesis Significance Testing

JEFF GILL, CALIFORNIA POLYTECHNIC STATE UNIVERSITY

The current method of hypothesis testing in the social sciences is under intense criticism, yet most political scientists are unaware of the important issues being raised. Criticisms focus on the construction and interpretation of a procedure that has dominated the reporting of empirical results for over fifty years. There is evidence that null hypothesis significance testing as practiced in political science is deeply flawed and widely misunderstood. This is important since most empirical work argues the value of findings through the use of the null hypothesis significance test. In this article I review the history of the null hypothesis significance testing paradigm in the social sciences and discuss major problems, some of which are logical inconsistencies while others are more interpretive in nature. I suggest alternative techniques to convey effectively the importance of data-analytic findings. These recommendations are illustrated with examples using empirical political science publications.

The primary means of conveying the strength of empirical findings in political science is the null hypothesis significance test, yet we have generally failed to notice that this paradigm is under intense criticism in other disciplines. Led in the social sciences by psychology, many are challenging the basic tenets of the way that nearly all social scientists are trained to develop and test empirical hypotheses. It has been described as a "strangle-hold" (Rozenboom 1960), "deeply flawed or else ill-used by researchers" (Serlin and Lapsley 1993), "a terrible mistake, basically unsound, poor scientific strategy, and one of the worst things that ever happened in the his-

Note: The author thanks Micah Altman, Gary King, Nick Theobald, Richard Tucker, and three anonymous referees for helpful and substantive comments. The original version of this essay was developed during Don Rubin's Graduate Seminar, Department of Statistics, Harvard University, 1997-1998.

Political Research Quarterly, Vol. 52, No. 3 (September 1999): pp. 647-674

tory of psychology" (Meehl 1978), "an instance of the kind of essential mindlessness in the conduct of research" (Bakan 1960), "badly misused for a long time" (Cohen 1994), and that it has "systematically retarded the growth of cumulative knowledge" (Schmidt 1996). Or even more bluntly: "The significance test as it is currently used in the social sciences just does not work" (Hunter 1997).

Statisticians have long been aware of the limitations of null hypothesis significance testing as currently practiced in political science research. Jeffreys (1961) observed that using p-values as decision criteria is backward in its reasoning: "a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." Another common criticism notes that this interpretation of hypothesis testing confuses inference and decision making since it "does not allow for the costs of possible wrong actions to be taken into account in any precise way" (Barnett 1973). The perspective of many statisticians toward null hypothesis significance testing is typified by the statement: "a P-value of 0.05 essentially does not provide any evidence against the null hypothesis (Berger, Boukai, and Wang 1997), and the observation that the null versus research hypothesis is really an "artificial dichotomy" (Gelman et al. 1995). Berger and Sellke (1987) show that evidence against the null given by correctly interpreting the posterior distribution or corresponding likelihood function "can differ by an order of magnitude."

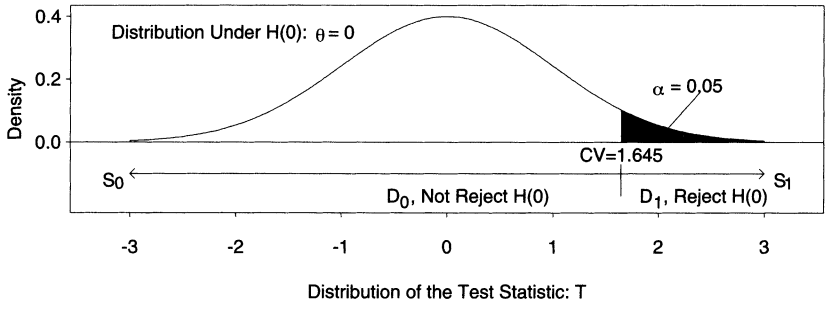
Political methodology has reemerged as an active and vibrant subfield developing important new methods and perspectives. However, the focus has not been on reviewing or reevaluating foundational issues, and there has been no discussion of the validity of the null hypothesis significance test as currently practiced in political science despite its pervasiveness. Noting this absence over quite some years, I discuss the history of the current hypothesis-testing paradigm, give evidence of some very serious misinterpretations of the approach, and present alternative procedures and interpretations.

Since political science is an overwhelmingly empirical discipline, the interpretation of these empirical results affect the interpretation of *substantive* conclusions. So a discussion about hypothesis-testing methodology is really a discussion about substantive political science issues. Thus the motivation for this review is not strictly methodological, nor is it confined to only a few data-analytic literatures.

The Current Paradigm: Null Hypothesis Significance Testing

The current, nearly omnipresent, approach to hypothesis testing in all of the social sciences is a synthesis of the Fisher *test of significance* and the Neyman-Pearson *hypothesis test*. In this "modern" procedure, two hypotheses are posited: a null or restricted hypothesis (H_0) which competes with an alternative or research hypothesis (H_1) describing two complementary notions about some phenomenon. The research hypothesis is the probability model which describes the author's belief about some underlying aspect of the data and operationalizes this belief through a

FIGURE 1
NULL HYPOTHESIS SIGNIFICANCE TESTING ILLUSTRATED



parameter: θ. In the simplest case,[1] described in every introductory text, a null hypothesis asserts that $\theta = 0$ and a complementary research hypothesis asserts that $\theta \neq 0$.

A test statistic (T), some function of θ and the data, is calculated and compared with its known distribution under the assumption that H_0 is true. Commonly used test statistics are sample means ($\bar{X}$), chi-square statistics (χ^2), and t-statistics in linear (OLS) regression analysis. The test procedure assigns one of two decisions (D_0, D_1) to all possible values in the sample space of T, which correspond to supporting either H_0 or H_1 respectively. The p-value ("associated probability") is equal to the area in the tail (or tails) of the assumed distribution under H_0 which starts at the point designated by the placement of T on the horizontal axis and continues to infinity. If a predetermined α level has been specified, then H_0 is rejected for p-values less than α, otherwise the p-value itself is reported.[2] Thus decision D_1 is made if the test statistic is sufficiently atypical given the distribution under H_0. This process is illustrated for a one-tail test at $\alpha = 0.05$ in Figure 1.

HISTORICAL DEVELOPMENT

The current null hypothesis significance test is a synthesis of two highly influential but incompatible schools of thought in modern statistics. Fisher developed

[1] More generally, the test evaluates a parameter vector: $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$, and the null hypothesis places restrictions on some subset ($\ell \leq m$) of the theta vector such as: $\theta_i = k_1\theta_j + k_2$ with constants k_1 and k_2.

[2] Formally, the sample space of T is segmented into two complementary regions (S_0, S_1) whereby the probability that T falls in S_1, causing decision D_1, is either one minus a predetermined null hypothesis cumulative distribution function (CDF) level: the probability of getting this or some more extreme value given a specified parametric form such as normal, F, t, etc. (α = size of the test, Neyman and Pearson), or the cumulative distribution function level corresponding to the value of the test statistic under H_0 is reported (p-value = $\int_{S_1} P_{H_0}(T = t)dt$, Fisher).

a procedure that produces significance levels from the data whereas Neyman and Pearson posit an intentionally rigid decision process which seeks to confirm or reject specified a priori hypotheses. The null hypothesis significance testing procedure is not influenced by the third major intellectual stream of the time: Bayesianism, except as a reaction against this approach.

Fisher Test of Significance

Fisher (1925a, 1934, 1955) posited a single hypothesis, H_0, with a known distribution of the test statistic T. As the test statistic moves away from its conditional expected value, $E(T \mid H_0)$, H_0 becomes progressively less plausible (less likely to occur by chance). The relationship between T and the level of significance produced by the test is established by the density outside the threshold established by T (one- or two-tailed), going away from the density region containing the expected value of T given H_0. The outside density is the p-value, also called the *achieved* significance level. Fisher hypothesis testing is summarized by the following steps:

1. Identify the null hypothesis.
2. Determine the appropriate test statistic and its distribution under the assumption that the null hypothesis is true.
3. Calculate the test statistic from the data.
4. Determine the achieved significance level that corresponds to the test statistic using the distribution under the assumption that the null is true.
5. Reject H_0 if the achieved significance level is sufficiently small. Otherwise reach no conclusion.

This construct naturally leads to the question of what p-value is sufficiently small as to warrant rejection of the null hypothesis. Although Fisher wrote in later years that this threshold should be established by the context of the problem[3] his influential work is full of phrases such as: "The value for Q is therefore significant on the higher standard (1 percent) and that for N_2 at the lower standard (5 percent)" (1971: 152-53). Furthermore, this determination of significance levels at 0.01 or 0.05 was made by Fisher in the context of agricultural and biological experiments.

Neyman and Pearson Hypothesis Testing

Neyman and Pearson (1928a, 1928b, 1933b, 1936a) reject Fisher's idea that only the null hypothesis needs to be tested. They argue that a more useful proce-

[3] This was partly defensive on Fisher's part as rigid significance threshold's are more conducive to Neyman-Pearson hypothesis testing: "In an acceptance procedure, on the other hand, acceptance is irreversible, whether the evidence for it was strong or weak. It is the result of applying mechanically rules laid down in advance; no *thought* is given to the particular case, and the tester's state of mind, or his capacity for *learning*, is inoperative" (Fisher 1955: 73-74).

dure is to propose two complementary hypotheses: Θ_A and Θ_B (or a class of Θ_{Bi}), which need not be labeled "null" or "alternative" but often are purely for convenience. Furthermore, Neyman and Pearson (1933b) point out that one can posit a hypothesis and consecutively test multiple admissible alternatives against this hypothesis. Since there are now two competing hypotheses in any one test, Neyman and Pearson can define an a priori selected α, the probability of falsely rejecting Θ_A under the assumption that H_0 is true, and β, the probability of failing to reject Θ_A when H_0 is false. By convention, the first mistake is called a Type I error, and the second mistake is called a Type II error. Note that α and β are probabilities *conditional* on two mutually exclusive events: α is conditional on the null hypothesis being true, and β is conditional on the null hypothesis being false. A more useful quantity than β is $1 - \beta$, which Neyman and Pearson (1933a, 1936a) call the *power* of the test: the long-run probability of accurately rejecting a false null hypothesis given a point alternative hypothesis.

In this construct it is desirable to develop the test which has the highest power for a given a priori α. To accomplish this goal, the researcher considers the fixed sample size, the desired significance level, and the research hypothesis, then employs the test with the greatest power. Neyman and Pearson's famous lemma (1936b) shows that under certain conditions[4] there exists a "uniformly most powerful" test which has the greatest possible probability of rejecting a false null hypothesis in favor of a point alternative hypothesis, compared to other tests.

To contrast the Neyman-Pearson approach with Fisher's test of significance, note how different the following steps are from Fisher's:

1. Identify a hypothesis of interest, Θ_B, and a complementary hypothesis, Θ_A.
2. Determine the appropriate test statistic and its distribution under the assumption that Θ_A is true.
3. Specify a significance level (α), and determine the corresponding critical value of the test statistic under the assumption that Θ_A is true.
4. Calculate the test statistic from the data.
5. Reject Θ_A and accept Θ_B if the test statistic is further than the critical value

[4] A sufficient condition is that the probability density tested has a monotone likelihood ratio. Suppose we have family of probability density functions $h(t|\theta)$ in which the random variable t is conditional on some unknown θ value to be tested. This family has a monotone likelihood ratio if for every $\theta_1 > \theta_2$, then:

$$\frac{h(t|\theta_1)}{h(t|\theta_1)}$$

is a non-decreasing function of the random variable t. Suppose further that we perform a test such as H_0: $\theta_1 \leq \theta_2$ versus H_1: $\theta_1 > \theta_2$ (θ_2 a known constant), where t is a sufficient statistic for θ_1, and $h(t|\theta_1)$ has a monotone likelihood ratio. The Karlin-Rubin Theorem (1956) states that if we set $\alpha = P(t > t_0)$ and reject H_0 for an observed $t > t_0$ (t_0 a known constant), then this test has the most power relative to any other possible test of H_0 with this α level (Casella and Berger 1900: 366-70; Lehmann 1986: 78).

from the expected value of the test statistic (calculated under the assumption that Θ_A is true). Otherwise accept Θ_A.

The Neyman-Pearson approach is important in the context of decision theory where the decision in the final step above is assigned a risk function computed as the expected loss from making an error.

Producing the Synthesis

The null hypothesis significance test attempts to blend the two approaches described above producing the "synthesis." With Fisher hypothesis testing, no explicit complementary hypothesis to H_0 is identified, and the p-value that results from the model and the data is evaluated as the strength of the evidence for the research hypothesis. Therefore there is no notion of the power of the test nor of accepting alternate hypotheses in the final interpretation. Conversely, Neyman-Pearson tests identify complementary hypotheses: Θ_A and Θ_B in which rejection of one implies acceptance[5] of the other and this rejection is based on a predetermined α level.

Neyman and Pearson's hypothesis test defines the significance level a priori as a function of the *test* (i.e., before even looking at the data), whereas Fisher's test of significance defines the significance level afterwards as a function of the *data*. The current paradigm in the social sciences straddles these two approaches by pretending to select a a priori, but actually using p-values (or asterisks next to test statistics indicating ranges of p-values) to evaluate the strength of the evidence. This allows inclusion of the alternate hypothesis but removes the search for a more powerful test.

The synthesized test also attempts to reconcile the two differing perspectives on how the hypotheses are defined. It adopts the Neyman-Pearson convention of two explicitly stated rival hypotheses, but one is always labeled as the null hypothesis as in the Fisher test. In some introductory texts the null hypothesis is presented only as a null relationship: $\theta = 0$ (no effect), whereas Fisher intended the null hypothesis simply as something to be "nullified." The synthesized test partially uses the Neyman-Pearson decision process except that failing to reject the null hypothesis is treated as a quasi-decision: "modest" support for the null hypothesis assertion. There is also confusion in the synthesized test about p-values and long-run probabilities. Since the p-value, or range of p-values indicated by stars, is not set a priori, it is not the long-run probability of making a Type I error but is typically treated as such. The synthesized test thus straddles the Fisher interpretation of p-values from the data and the Neyman-Pearson notion of error probabilities from the test.[6]

[5] Some people are surprised, but Neyman and Pearson actually do use the word "accept," see 1933b.

[6] It is very interesting to note that with the synthesized modern hypothesis test there is no claim of authorship. The acrimony, both intellectual and personal, between Fisher and Neyman & Pearson is legendary and continued until Fisher's death. So it is curious that no one was willing to claim

Many problems with the current paradigm result from the mixture of these two essentially incompatible approaches (Gigerenzer et al. 1989, Gigerenzer 1993, Gigerenzer and Murray 1987, MacDonald 1997). While both approaches seek to establish that some observed relationship is attributable to effects distinct from sampling error, there are important differences as noted above. *Neither Fisher nor Neyman and Pearson would have been satisfied with the synthesis.*Fisher objected to preselection of the significance level as well as the mandatory two-outcome decision process. Neyman and Pearson disagreed with interpreting p-values (or worse yet, ranges of p-values indicated by "stars") as the probability of Type I errors since they do not constitute a long-range probability of rejection. Neyman and Pearson also considered the interpretation of data-derived p-values to be subjective and futile (Neyman and Pearson 1933b: footnote 1).

Criticisms of the Null Hypothesis Significance Test

In this section I summarize some of the most important criticisms levied against null hypothesis significance testing. While this set of criticisms is not intended to be exhaustive, it includes most of those directly relevant to the current practice of hypothesis testing in political science. Some of these problems are simply common misinterpretations, while others are specific flaws in the logic of the synthesis.

Probabilistic Modus Tollens

The basis of the null hypothesis significance test rests on the logical argument of modus tollens (denying the consequent). The basic strategy is to make an assumption, observe some real-world event, and then check the consistency of the assumption given this observation. The syllogism works like this:

If A then B	If H_0 is true then the data will follow an expected pattern
Not B observed	The data do not follow the expected pattern
Therefore not A	Therefore H_0 is false.

The problem with the application of this logic to hypothesis testing is that the certainty statements above are replaced with probabilistic statements, causing the logic of modus tollens to fail. To see this, reword the logic above in the following way:

If A then B is highly likely	If H_0 is true then the data are highly likely to follow an expected pattern
Not B observed	The data do not follow the expected pattern
Therefore A is highly unlikely	Therefore H_0 is highly unlikely.

responsibility for a potentially bridging approach and it appeared in the textbooks anonymously (Gigerenzer 1987: 21). The timidity of these authors led them to try and accommodate both perspectives by denying that there were substantive differences.

Initially, this logic seems plausible. However, it is a fallacy to assert that obtaining data that is atypical under a given assumption implies that the assumption is likely false: almost a contradiction of the null hypothesis does not imply that the null hypothesis is almost false (Falk and Greenbaum 1955). For example (Cohen 1994; Pollard and Richardson 1987):

If A then B is highly likely	If a person is an American then it is highly unlikely she is a member of Congress
Not B observed	The person is a member of Congress
Therefore A is highly unlikely	Therefore it is highly unlikely she is an American.

From this simple little example and the resulting absurdity it is easy to see that if the P(**C**ongress|**A**merican) is low (the p-value), it does *not* imply that P(**A**merican|**C**ongress) is also low.

The Inverse Probability Problem

A common interpretive problem with null hypothesis significance testing is a misunderstanding of the order of the conditional probability. This problem was first discussed in political science by King (1989: 16-18) and this discussion extends his remarks by describing it more technically. Many people have a belief, which stems directly from Fisher's perspective, that the smaller the p-value, the greater the probability that the null hypothesis is false (Carver 1978, Cohen 1994, Meehl 1990). This incorrect interpretation is that the null hypothesis significance test produces $P(H_0|D)$: the probability of H_0 being true given the observed data D. But the null hypothesis significance test first posits H_0 as true then asks what is the probability of observing these or more extreme data. This is clearly $P(D|H_0)$. In fact, a more desirable test would be one that produces $P(H_0|D)$ because then we could simply search for the hypothesis with the greatest probability of being true given some observed data. Bayes law clarifies the difference between these two unequal probabilities:

$$P(H_0|D) = \frac{P(H_0)}{P(D)} P(D|H_0).$$

The two quantities, $P(D|H_0)$ and $P(H_0|D)$, are equal only if $P(H_0) = P(D)$, and there is absolutely no theoretical justification supporting such an equality.

The fact that we gain little direct insight about $P(H_0|D)$ from the null hypothesis significance test alone is illustrated with a simple example. Hypothetically assume that 2 percent ($P(M) = 0.02$) of the population of the United States are members of some right wing militia group (a fact some members attempt to hide, and will therefore not typically admit to an interviewer). A survey is 95 percent accurate on positive classification ($P(C|M) = 0.95$), and 97 percent accurate on negative classification ($P(C^c|M^c) = 0.97$), where C indicates positive classification and M indicates militia membership). What is the probability that, given the survey analysis indicates that the person is a member of a right wing militia, the person

really is a member? First we derive the unconditional probability of classifying a respondent as a militia member:

$$\begin{aligned} P(C) &= P(C \cap M) + P(C \cap M^c) \\ &= P(C|M)P(M) + [1 - P(C^c|M^c)]P(M^c) \\ &= (0.95)(0.02) + (0.03)(0.98) \cong 0.05 \end{aligned}$$

Now we can use this unconditional to apply Bayes law:

$$P(M|C) = \frac{P(M)}{P(C)} P(C|M) = \frac{0.02}{0.05} (0.95) = 0.38.$$

The highlighted difference is that the probability of correctly classifying an individual as a militia member given the person is a militia member is 0.95, yet the probability that an individual is a militia member given the person is positively classified is 0.38. Mistakenly inversing the conditional probability from the null hypothesis significance test from $P(D|H_0)$ to $P(H_0|D)$ can produce similarly dramatic differences.

Maximum likelihood estimation (Fisher 1925b[7]) finesses the inverse probability problem by substituting the unbounded notion of likelihood for the bounded definition of probability by treating the ratio:

$$f(D) = \frac{P(\theta)}{P(D)}$$

as an unknown function of the data independent of $P(D|\theta, H_0)$, and exploiting: $L(\theta|D, H_0) \propto P(D|\theta, H_0)$. The likelihood function, $L(\theta|D, H_0)$, is similar to the desired but unavailable inverse probability, $P(\theta|D, H_0)$, in that it facilitates testing alternate values of θ, to find a most probable value: $\hat{\theta}$. The likelihood function differs from the inverse probability in that it is necessarily a *relative* function since $f(D)$ is unknown.

If we are interested in obtaining the maximum likelihood estimate of the unconstrained parameter vector $\hat{\theta}$, given a parametric family of models, we need only maximize the function $L(\theta|D, H_0)$ with regard to the θ values. The result is the θ that is most likely to have generated the data given H_0 expressed through a specific parametric form.[8] This process thus produces the most likely value of θ *relative* to other possible values in the sample space of θ.

[7] Although Fisher undoubtedly developed the modern use of maximum likelihood estimation, its origins are generally credited to Gauss. See Stigler (1986: 141) or Brenner-Golomb (1993: 299) for interesting discussions.

[8] Note that this form of the null hypothesis is much more lenient than other forms discussed. The null hypothesis in the context of maximum likelihood estimation refers to the restricted model in the denominator of the likelihood ratio test.

Bayesian analysis addresses the inverse probability problem by making distributional assumptions about the unconditional distribution of the parameter vector, θ, *prior* to observing the data, D. This provides a means of integrating out θ to solve for the previously unknown $P(D, H_0)$:

$$P(D, H_0) = \int_{\theta \in \Theta} P(D|\theta, H_0)\, P(\theta, H_0) d\theta$$

With this construct, the conditional (posterior) distribution of θ is:

$$P(\theta|D, H_0) = \frac{P(\theta, H_0)}{P(D, H_0)} P(D|\theta, H_0).$$

This is also expressed in a more intuitive manner like the maximum likelihood formulation that essentially states that the posterior distribution is proportional only to the likelihood times the prior distribution: $P(\theta|D, H_0) \propto P(D|\theta, H_0)\, P(\theta, H_0)$. The primary advantage of the Bayesian approach (the maximum likelihood estimate is equal to the Bayesian posterior mode with the appropriate uniform prior, and they are asymptotically equal given *any* proper prior[9]), is that in a fairly tractable way we get a posterior *distribution* for θ, where the information in the data progressively subsumes a prior assumption as the sample size increases. A Bayesian posterior does not lead to an arbitrary decision about rejection. Instead the posterior distribution of θ explicitly supplies probabilities of interest such as $P(\theta_i > 0)$.

Model Selection

Typically when a political science model with a null hypothesis significance test is reported it is presented as if only two models were ever considered or ever deserved to be considered: the null hypothesis and the provided research hypothesis. The quality of a research finding is then solely judged on the ability to reject the complementary null hypothesis with a sufficiently low p-value. However, during the development of the reported model many differing alternate mixes of independent variables are tested. This is an "illusion of theory confirmation" (Greenwald 1975, Lindsay 1995) because the null hypothesis significance test is presented as evidence of the exclusivity of explanation of this single research hypothesis. Summary statistics are reported from the n^{th} equation as if the other $n - 1$ never existed and this last model is produced from a fully controlled experiment (Leamer 1978: 4). The null hypothesis significance test thus provides an infi-

[9] Despite this fact, there remains a division between Frequentists and Bayesians particularly in small sample problems where the asymptotic equivalence is not applicable. A common Frequentist criticism of the Bayesian approach is that subjective priors have great impact on the posterior distribution for problems with small sample sizes. There is a developing literature on Robust Bayes analysis that seeks to mitigate this problem by developing estimators that are relatively insensitive to a wide range of prior distributions (Berger 1984).

nitely strong bias in favor of a single research hypothesis against a huge amount of other reasonable hypotheses: all other distributional alternatives are assumed to have probability zero (Rozeboom 1960, Lehmann 1986: 68, Popper 1968: 113).

Two entirely reasonable and statistically significant, competing models can lead to substantively different conclusions using the exact same data (Raftery 1995). In many cases the decision criteria that led to the final model are based at least in part on intermediate significance levels and that the significance levels reported in the final published work have very different interpretations than the significance levels in intermediate models (Leamer 1978, Miller 1990, Raftery 1995). For example suppose that the p-value for some coefficient, X_1, in the final model is 0.01. However, in at least one model rejected by the researcher the significance of X_1 was 0.4 in the presence of another variable, X_2, omitted from the final model. Since X_2 has an effect on X_1, but is not itself included in the reported test of significance, the p-value for X has a substantively different interpretation in the two contexts.

Significance Through Sample Size

There are two misinterpretations about the role of sample size in null hypothesis significance testing. First is the belief that statistical significance in a large sample study implies real world importance. Many have observed that a null hypothesis significance test based on a large sample size almost always results in statistical significance (Leamer 1978, Macdonald 1997, Oakes 1986, Raftery 1995).[10] This is a concern in political science research since we do not want to infer that some subfields have greater legitimacy just because the corresponding data sets tend to produce smaller p-values: "a prejudice against the null" (Greenwald 1975: 1). Contrast survey research in American politics where studies routinely use data sets with thousands of cases such as the American National Election Study and the General Social Survey with comparative politics where small sample sizes are often the only resource available (Fearon 1991: 194, Lijphart 1971: 685). Researchers studying small sample events such as revolutions or major wars face substantial prejudice against their empirical findings when alpha levels appropriate to large sample survey research are applied.[11]

The correct interpretation is that as the sample size increases we are able to distinguish smaller population-effect sizes progressively. Effect size is the extent to which some measured phenomenon exists in the population. Finding population-

[10] It should also be noted that this statement says nothing at all about data *quality*. Large sample sizes based on convenience samples, those produced from non-random or arbitrary selection criteria, and those selected on the dependent variable provide particularly dubious results. Simply increasing sample size by these means to achieve statistical significance is an ill-advised strategy and almost certain to lead to incorrect results.

[11] Western and Jackman (1994: 413) describe this problem rather nicely: ". . . frequentist inference answers a question that comparative researchers are not typically asking."

effect sizes is actually the central purpose of political science research since any difference can be found to be statistically significant given enough data. Political scientists are more interested in the relative magnitude of effects (incumbency advantage, tendency for war, probability of voting, cultural influences, etc.), and making only binary decisions about the existence of an effect is not particularly informative. However, the null hypothesis significance test deflects us into an obsession with the strength of this binary decision measured through p-values.

The second misinterpretation is that for a given p-value in a study which rejects the null hypothesis, larger sample sizes imply more reliable results. This is false: two studies that reject the null with the same p-value are equally likely to make a Type I error even if they have dramatically different sample sizes. This mistake results from a poor understanding of *Type II* errors. Unlike the case above, two studies which *fail* to reject the null hypothesis and are identical in every way except sample size *are* different qualitatively. The test with the larger sample size is less likely to make a Type II error since the sample size is always in the denominator of the equation for statistical significance. To show why this is true Figure 2 displays two tests for: H_0: $\mu = 0$ versus H_1: $\mu > 0$, with sample sizes of 9 and 12 respectively and the true distribution normal (3, 3). The shaded region is the area under the true sampling distribution of $\bar{X}$ to the left of the area determined by setting $\alpha = 0.05$ under a false assumption that H_0 is true. Thus the shaded region is the probability we would fail to reject the false H_0 (i.e. $\beta = P$(Type II error)). It is clear from Figure 2 that the probability of a Type II error is much lower for the test with a larger sample size, *even though the probability of a Type I error is identical.* This distinction is not well known. The area under the true sampling distribution of $\bar{X}$ in either panel of Figure 2 not shaded is equal to $1 - \beta - P$(Type II error), which is the power of the test. It follows from above that the second test with larger sample size has greater power.

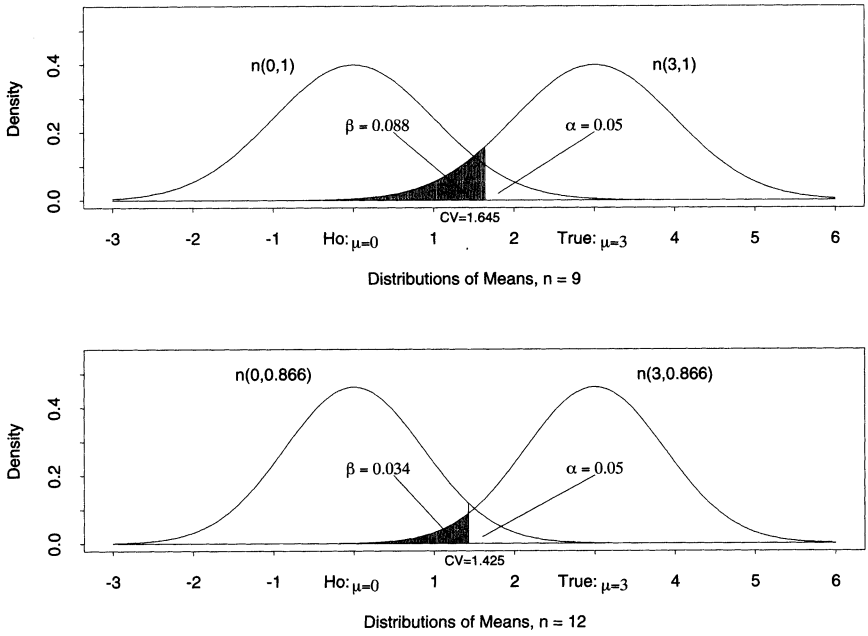
The Arbitrariness of Alpha

Fisher constructed the first significance tables and was therefore influential with regard to establishing rejection thresholds.[12] Since computers were rare and expensive during the adolescence of the null hypothesis significance test, the arbi-

[12] The firmly established significance thresholds of 0.01 and 0.05 come directly from the importance and influence of Fisher's work. The tenth edition of *Statistical Methods for Research Workers* (1946) includes a comprehensive listing of Fisher's published and unpublished work from 1912 to 1945. Of the 143 papers and books there are 74 works in biology and agronomy, 61 on statistical theory, 4 on climatology, 2 on playing cards, 1 on harmonics, and a presidential address. It is clear from the distribution of Fisher's interests that biological and agricultural experiments were the dominant applications of his ideas. Even if we assume that 0.01 and 0.05 are the ideal thresholds for high and moderate significance for experiments in biometry and biometrics (a very suspect assumption), there is absolutely no theoretical justification for wholesale importation into completely unrelated fields such as political science.

FIGURE 2

TYPE II ERRORS FOR $\bar{X}$, DIFFERING SAMPLE SIZES



trary significance levels in published tables were very convenient to researchers. These tables cannot give every possible value in the range of a test statistic, so particular discrete values had to be provided by convention. Therefore, a contributing factor in the enshrinement of 0.05, 0.01, and 0.001 is the ease with which tables for common distributions can be presented. However, this logic no longer applies as computers are pervasive in research and teaching environments.

On what basis do we decide that $p = 0.051$ is unacceptable but $p = 0.049$ is cause for rejoicing? Such a distinction relies on the assumption that there is virtually no measurement error, an assumption very few social scientists are willing to defend. There is no published work in political science that provides a theoretical basis for these thresholds. It is convenient to say "one time out of twenty" or "one time out of a hundred," but it is no less convenient to say "one time out of fifty" or "one time out of twenty-five" (i.e., $p = 0.02$ and $p = 0.04$). Fisher's justification rests on no scientific principle other than the belief that these levels represented some standard convention in human thought: "It is usual and convenient for experimenters to take 5 percent as a standard level of significance. . ." (1934: 15). Rosnow and Rosenthal (1989) note the arbitrariness with great flair: ". . . surely God loves .06 nearly as much as .05."

Replication Fallacy

One of the most heinous misinterpretations of null hypothesis significance testing is the belief that one minus the p-value is the probability of replication (producing significant results in repeated iterations of the same study). This is equivalent to the two misconceptions: $P(H_0) = \alpha$ and $P(H_1) = 1 - \alpha$. The term "replication fallacy" was coined by Falk and Greenbaum (1995) from their observation that many researchers believe that a low p-value such as $p = 0.05$ implies that 95 out of 100 replications will be statistically significant. The error is obvious when one recalls that $p = P(D|H_0)$ and is thus a function of a single data set producing a single test statistic. What really interests us in terms of replication is the *distribution* of this test statistic in repeated trials.

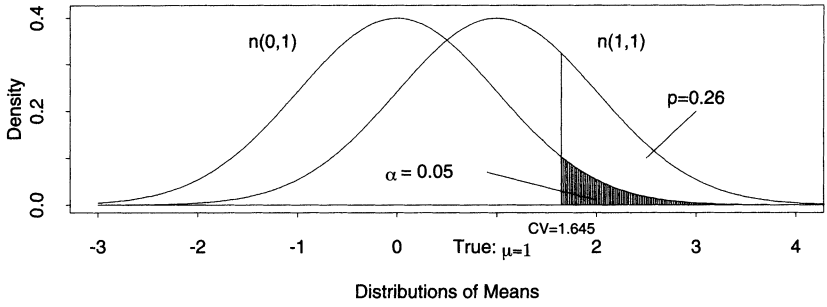
The probability of replication given a false null is actually the power of the test (Cohen 1992, 1977; Schmidt 1996; Sedlmeier and Gigerenzer 1989). This can be seen by revisiting Figure 2. The distribution on the right-hand side of the diagram (in either panel) is the true distribution of the test statistic over multiple replications. Recall that the shaded region is the probability that we would fail to reject the false distribution under the null hypothesis centered at zero. This region is the area under the true distribution but determined by the line that begins the alpha region under the false null hypothesis distribution. Values of the test statistic that fall along the x-axis to the right of the shaded region are those with which we would correctly reject the false null hypothesis. Therefore the unshaded region under the correct distribution is the probability of replicating statistically significant results, $1 - \beta$ (i.e., the power), given of course a false null hypothesis.

Asymmetry and Accepting the Null Hypothesis

We teach graduate students to be very careful when describing the occurrence of not rejecting the null hypothesis. This is because failing to reject the null hypothesis does not rule out an infinite number of other competing research hypotheses. Null hypothesis significance testing is asymmetric: if the test statistic is sufficiently atypical given the null hypothesis then the null hypothesis is rejected, but if the test statistic is insufficiently atypical given the null hypothesis then the null hypothesis is not accepted. This is a double standard: H_1 is held innocent until proven guilty and H_0 is held guilty until proven innocent (Rozeboom 1960).

Despite the fact that it is very dangerous implicitly or expressly to accept the conclusion from a non-rejected null hypothesis, instances are common. Eight of the twenty articles in the *American Political Science Review*, Volume 91 (1997), that used a null hypothesis significance test drew substantive conclusions from a fail to reject decision (40 percent). The *American Journal of Political Science*, Volume 41 (1997), contains 54 manuscripts, that either explicitly stated a hypothesis test, provided p-values, or furnished "stars," and 25 of these drew substantive conclusions from a failure to reject the null (46 percent). The same analysis for *Political Research Quarterly*, Volume 50 (1997), reveals 14 of 37 possible that made this error (38

FIGURE 3
POWER ANALYSIS AND EFFECT SIZE



percent). The *Journal of Politics*, Volume 59 (1997), has 18 out of 37 possible that made the error (51 percent).

There are two problems that develop as a result of asymmetry. The first is a misinterpretation of the asymmetry to assert that finding a non-statistically significant difference or effect is evidence that it is equal to zero or nearly zero. Regarding the impact of this acceptance error Schmidt (1996: 126) asserts that this: "belief held by many researchers is the most devastating of all to the research enterprise." This acceptance of the null hypothesis is damaging because it inhibits the exploration of competing research hypotheses. The second problem pertains to the correct interpretation of failing to reject the null hypotheses. Failing to reject the null hypothesis essentially provides almost no information about the state of the world. It simply means that given the evidence at hand one cannot make an assertion about some relationship: all you can conclude is that you *can't* conclude that the null was false (Cohen 1962).

Cross-Validation Studies

Replication of published work is an increasingly important part of scholarship in political science (King 1995, Meier 1997). Unless the replication study confirms the original findings (the type of replications that are unlikely to be published), the quality of a replication analysis can never be judged without understanding the power and effect size of the replication test.

Suppose we are interested in replicating a study in which the original author finds that a test statistic is sufficiently far from zero, given the observed variance, to conclude that this is evidence to infer the existence of some political effect. Suppose further that the effect size is really one: the *true* distribution of this test statistic is normal (1, 1), and therefore the author has reached the correct conclusion. We don't necessarily know that the author is correct or we would not bother doing the replication study. Our null hypothesis in the replication study is that the test statistic is normally distributed around zero so there is no corresponding political

effect in the population. These assumptions provide substantial simplification, but no real loss of generality.

Figure 3 shows the rejection region (shaded) for the test statistic determined by the $\mu = 0$ null hypothesis and $\alpha = 0.05$, but the vertical line is extended up to the true distribution with $\mu = 1$. This indicates that the power of this test is only 0.26. So 74 percent of the replication studies that attempt to show that the effect is zero, using $\alpha = 0.05$, will make the incorrect decision and assert that there is no evidence to show a statistically significant effect.

Recommendations

This section presents several alternatives to the null hypothesis significance test used by researchers in many scientific literatures. These techniques provide the means to report evidence of empirically observable effects without the associated problems described previously. Other effective interpretive and explanatory procedures not described here include: predicted values, expected values, first differences (King et al. 1998), graphical summaries (Cleveland 1993; Gill and Thurber 1999; Jacoby 1997), and simulation evidence (Gelman et al. 1995; Ripley 1987; Tanner 1996).

In general, misunderstanding the inverse probability problem is the primary cause for believing that a smaller p-value is greater evidence that the null hypothesis is false and subsequently adopting the flawed logic of the null hypothesis significance test. So the primary challenge is to interpret the implications of the observable $P(D|H_0)$ on the unobservable $P(H_0|D)$.

Confidence Intervals

Confidence intervals are an alternative to null hypothesis significance testing that provide the same information and more, without requiring a decision (Casella and Berger 1990: 406; Lehmann 1986: 89). Confidence sets (not necessarily contiguous), confidence intervals, and credible sets (Bayesian set estimates) are simply estimates of some parameter θ in which the uncertainty is expressed as a range of alternate values and a probability of coverage.[13] In one sense confidence intervals and null hypothesis significance tests present the same information: a linear regression coefficient with a $1 - \alpha$ confidence interval bounded away from zero is functionally identical to a null hypothesis significance test rejecting at $p \leq \alpha$ the hypothesis that the coefficient equals zero.

13 Bayesian credible sets measure the probability that the parameter is in the interval rather than the probability that the interval covers the true parameter. Thus credible sets are superior in terms of their intuitive appeal. Conversely, with confidence sets the set itself is the random quantity and the unknown parameter is fixed. So we cannot state that with any produced confidence set there is a known probability that the unknown parameter is contained, we have to say that we are $(1 - \alpha)$ percent confident that this set covers the true parameter value.

Confidence intervals have a great virtue: as the sample size increases the size of the interval decreases, correctly expressing our increased certainty about the parameter of interest. This is analogous to the correct interpretation of increasing power in a null hypothesis significance test as sample size increases. Most misunderstandings about sample size as a quality measure in null hypothesis significance testing stem from a poor understanding of Type II errors. Since there is no Type II error in confidence intervals, there is less potential for such confusion.

Unfortunately the confidence level of the interval is subject to the same arbitrary interpretations as α levels. Therefore confidence intervals require the same cautions with regard to sample size interpretations and unsupported conventions about α levels.

An even more basic and condensed form reporting of the confidence set is a table of parameter estimates with their associated standard errors omitting p-values and confidence intervals. This format provides readers with an immediate indication of location and uncertainty free from a format that leads to the problems reviewed previously. However, simply reporting parameter estimates and standard errors is a relatively unstructured presentation of results. One possible strategy is to divide each coefficient into one of three groups based on location and uncertainty in order to convey reliability with regard to distributional assumptions:

I. Highly reliable estimates with an interpretation that would likely overwhelm a wide range of reasonable distributional assumptions.
II. Highly suspect estimates which are either very unreliable based on the magnitude of their standard error, or with an interpretation that is likely to change with different distributional assumptions.
III. Estimates that fit neither of the previous two descriptions and for which we remain in doubt.

This typology places a loose structure on the results, but does not require the parametric assumptions and flawed decisions that create many of the problems described previously. The grouping is loosely derived from Neyman and Pearson's notion that the region of acceptance can be divided into subregions (1933b: Section I).

Example: Confidence Intervals and Leamer Bounds in a Discrete-Time Survival Analysis of Unionization

Since reported coefficients are sensitive to model specification, it is desirable to have some reporting mechanism for conveying the range that coefficients take in the unreported developmental models. Leamer (1983) proposes reporting extreme bounds for coefficient values during the specification search. A recent use of this approach is found in Western (1995). He argues that a decline in unionization during the 1980s is due to factors such as negative international economic conditions, passivity, fewer resources for the unions, and a diminished position in individual economies. To provide evidence Western develops a simple hazard-rate

TABLE 1
SURVIVAL ANALYSIS, PREDICTING YEAR UNION DECLINE:
18 OECD COUNTRIES, 1973-1989

Explanatory Variable	Quasi-Likelihood Coef.	Quasi-Likelihood Std. Error	99% Confidence Interval	Leamer Bounds	Estimated Power ($\alpha = 0.01$)
Highly Reliable Estimates					
Constant	–2.42	0.78	[–4.43:–0.41]	(–5.24:–0.98)	0.70
Year	1.41	0.22	[0.84: 1.98]	(0.24: 1.41)	0.90
Economic Openness	0.12	0.04	[0.02: 0.22]	(0.04: 0.12)	0.70
Unemployment	1.18	0.39	[0.34: 2.03]	(0.89: 1.18)	0.85
Union Density (lagged)	–0.23	0.04	[–0.33:–0.13]	(–0.23:–0.11)	0.99
Decentralization	2.77	0.64	[1.11: 4.43]	(0.33: 2.77)	0.97
Left Government	–4.41	0.86	[–6.64:–2.18]	(–4.41:–1.71)	0.99
Highly Suspect Estimates					
Strike Activity ($\times 10^4$)	0.11	2.75	[–6.97: 7.19]	(–1.50: 2.15)	0.01
Indeterminant Estimates					
(none)					

regression model of union decline using unionization data from 18 Organization for Economic Cooperation and Development (OECD) countries (1973 to 1989). A subset of Western's results are presented in Table 1 where the typologies suggested above are implemented. Table 1 also includes empirically estimated power using $\alpha = 0.01$, and the assumption that null hypothesis implies a similarly shaped marginal likelihood except centered at zero. Figure 2 displayed the calculation of power when there was perfect knowledge about the distribution of the test statistic under the null hypothesis and the research hypothesis. If the researcher is reasonably confident that the dispersion of the test statistic would not take on a substantially different form than that seen in the observed data, then this power estimate can be useful (Cohen 1977).

Although the coefficient for Left Government looks highly reliable, it is worth noting that it varied considerably during the model specification process. This should not be construed as making this variable less trustworthy in the final specification. Instead it should reasonably be considered a topic for further investigation. Conversely, the bounds for Economic Openness are very small providing confidence for the reliability of this coefficient across model specification. Also note that the only highly suspect coefficient estimate, Strike Activity, had Leamer extreme bounds which straddle zero, thus lending additional support for the classification.

Bayesian Analysis

Bayesianism is clearly in the midst of a renaissance period. Fueled by dramatic improvements in computing techniques, the field is flourishing as many previously

intractable problems are now realistically solvable. As a result, Bayesian applications have found their way into most social science fields.

In Bayesian analysis, inferences about unknown model parameters are not expressed as point estimates with the standard accompanying measure of variance. Instead statistical information about the unknown parameters is summarized in probability statements such as quantiles of the posterior distribution, the probability of occupying some region of the sample space, the posterior predictive distribution, and Bayesian forms of confidence intervals: the credible set and the highest posterior density region. Thus the dominant way of thinking is to consider that the parameter has some underlying distribution which should be described.

In the Bayesian setup the unnormalized posterior distribution for the unknown coefficients is calculated by: $P(\theta|D, H_0) \propto P(D|\theta, H_0)\, P(\theta|H_0)$. This means that the desired probability statement is a product of the likelihood function and some prior belief about the distribution of $\theta|H_0$. The proportionality is used instead of equality because the $P(D, H_0)$ term in Bayes law does not depend on θ and can therefore be dropped.

While the Bayesian assignment of a prior distribution for the unknown parameters strikes many as subjective, there are often strong arguments for particular forms of the prior: little or vague knowledge often justifies a diffuse or even uniform prior, certain probability models logically lead to particular forms of the prior, and the prior allows researchers to include additional information collected a priori (Gelman et al. 1995; Box and Tiao 1973; Press 1989; Berger 1985; Jeffreys 1961). A common defense of the "subjective priors" criticism is the observation that corresponding frequentist models are equally subjective: the choice of admissible hypotheses, the selected significance level, determination of an adequate sample size, and the adequacy of the test statistic (Howson and Urbach 1993).

Explicit hypothesis testing can also be performed in the Bayesian construct. Suppose Θ_A and Θ_B represent two competing hypotheses about the state of some unknown parameter, θ, which together form a partition of the sample space: $\Theta = \Theta_A \cup \Theta_B$. To begin, prior probabilities are assigned to each of the two outcomes: $\pi_A = p(\theta \in \Theta_A)$ and $\pi_B = p(\theta \in \Theta_B)$. This allows us to calculate the competing posterior distributions from the two priors and the likelihood function: $p_A = p(\theta \in \Theta_A|D, H_0)$ and $p_B = p(\theta \in \Theta_B|D, H_0)$. It is common to define the prior odds, π_A/π_B, and the posterior odds, p_A/p_B, as evidence for H_A versus H_B. A much more useful quantity, however, is $(\pi_A/\pi_B)/(p_A/p_B)$ which is called the *Bayes Factor*. The Bayes Factor is usually interpreted as odds favoring H_A versus H_B given the observed data. For this reason it leads naturally to the Bayesian analog of hypothesis testing.

Example: The 1982 Election Victory of the Spanish Socialist Party

Bernardo (1984) develops a precinct level Bayesian hierarchical model of vote choice for the Spanish election of 1982 in which the Socialist party obtained control of the government for the first time since the Civil War. The author defines n_{ij}

TABLE 2
HPD REGIONS: PREDICTED PERCENTAGES OF 1982 VOTES BY PARTY

	Party				
Valencia Province	Socialist	Conservative	Center	Center-Left	Communist
4 Weeks Before Election	[39.0:48.9]	[12.6:19.4]	[7.9:13.2]	[4.0: 7.8]	[5.1:9.2]
1 Week Before Election	[47.3:54.2]	[13.3:24.9]	[5.0:11.8]	[7.0:11.9]	[4.0:6.7]
First 100 Votes from 20 Polls	[49.0:57.7]	[23.5:31.2]	[2.3: 4.6]	[1.1: 2.9]	[4.0:6.2]
Total Vote from 20 Polls	[50.1:56.8]	[26.6:32.6]	[3.3: 4.6]	[1.8: 2.7]	[3.7:5.6]
Actual Province Results	53.5	29.4	4.4	2.3	5.3

as the number of voters in the i^{th} precinct voting for the j^{th} party. The data from the m precincts surveyed ($\{n_{1,j}, n_{2,j}, \ldots, n_{m,j}\}$, j = 1 to 5 major political parties) are assumed to be from an underlying multinomial distribution with unknown parameters θ_{ij} representing the probability of a vote for the j^{th} party in the i^{th} precinct with the constraints that these values are non-negative and sum to one. Bernardo specifies a uniform prior distribution on the θ_{ij} values and this leads naturally to a Dirichlet form (a multivariate generalization of the beta) of the posterior.

A substantively interesting aspect of the methodology is the scheduling of data collection and analysis. Data are collected in the province of Valencia at four points in time: by a survey four weeks before the election (n = 1000), by a survey one week before the election (n = 1000), by using the first 100 valid votes from 20 representative polling stations, and by using all valid votes from these same polling stations after the polls are closed. Data collection was performed with the full cooperation of the Spanish government and the results were immediately provided to the national media.

Bernardo presents the Bayesian estimates of predicted vote proportion by party as 0.90 highest posterior density regions. Highest posterior density (HPD) regions are essentially confidence intervals which cover the 100 $(1 - \alpha)$ percent of the posterior distribution with the highest probability regardless of whether or not the area is contiguous (although they are in this example). These results are summarized in Table 2.

One interesting result from this analysis is that the 0.90 HPD regions shrink as the election nears and better polling data are received. This reflects the growing certainty about the estimates as data quality improves. In addition, the estimates from actual polling data are remarkably accurate for the two parties receiving the largest vote share. Note that the uniform prior does not constrain the final, highly non-uniform, posterior distribution.

The simple example provided here demonstrates that Bayesian data analysis is essentially free from the previously described problems with the null hypothesis

significance test. Inferences are communicated to the reader without decisions, p-values, and confused conditional probability statements. The Bayesian approach also interprets sample size increases in a more desirable manner: larger sample sizes reduce the importance of prior information rather than guarantee a low but meaningless p-value.

Meta-Analysis

Meta-analysis offers the attractive proposition that the accumulation of knowledge on some research question can be compared and combined in a single procedure. These techniques look at how much of the variance in a set of studies is attributable to individual uncertainties such as sample variability and measurement error, and how much variability is truly the result of some real effect. The literature on this tool is notably young, starting with Cooper (1984); Glass (1976); Glass, McGaw, and Smith (1981); Hunter and Schmidt (1990); Schmidt and Hunter (1977); and Wolf (1986).

As with any other statistical procedure, meta-analysis has assumptions, and is therefore inappropriate under certain circumstances (Chow 1996: chap. 5). Meta-analysis assumes that the compared studies are all independent of each other and that variables are measured approximately the same way. No two studies will be identical, but the goal of meta-analysis is to calculate some commonality of effects from different studies. This is a strength because effect sizes that are large across a wide range of research designs have substantial supporting evidence. This is also a weakness because not finding a large meta-effect size can be caused by heterogeneity across individual models.

Meta-analysis can be performed for virtually any reported test statistic, but effect size has been a particularly popular choice because of its intuitive interpretation and relationship to the power of the individual studies.[14] The basic procedure develops an aggregated effect size and variance from separate studies accounting for sampling error and measurement error in the original estimates. In the simplest case, meta-analysis seeks to distinguish sampling error variance and the observed effect size variance to estimate the variance of the population effect size.

Example: Meta-Analysis of Federal Healthcare Reform Surveys

Consider an example in which five different surveys are conducted by five different polling organizations to gauge support for a federal program to guarantee affordable healthcare for all U.S. citizens. In each case respondents are asked whether they support this level of federal involvement in the healthcare industry.

[14] The test for effect sizes and correlation coefficients is the same since switching from one to the other is trivial:

$$r = \frac{d}{\sqrt{d^2 + 4}} \text{ and } d = \frac{2r}{\sqrt{1 - r^2}}.$$

TABLE 3
SAMPLE EFFECT SIZES FOR FIVE STUDIES

Study	Sample Size: n_i	Observed Effect Size: $d_{s,i}$
1	450	0.03*
2	500	0.09*
3	370	0.01
4	400	0.24
5	630	0.26*

$N = 2350$, $^*p < 0.05$

From an OLS regression model that includes a coefficient for this question, different effect sizes are determined, with varying levels of significance using the null hypothesis significance test.

Suppose for the sake of simplicity we ignore adjustments for measurement error and other heterogeneities. The goal is to obtain a meta-effect size, $\bar{d}$, and its corresponding variance which has two components: the average weighted observed variance from the studies, $VAR(d_s)$, and the sampling error variance, $VAR(s)$. From Table 3 the mean sample size is $\bar{n} = 470$ for the $k = 5$ studies. Thus the average weighted observed effect size and its variance are:

$$\bar{d}_s = \frac{1}{N} \sum_{i=1}^{k} n_i d_{s,i} = 0.137$$

$$VAR(d_s) = \frac{1}{N} \sum_{i=1}^{k} (n_i (d_{s,i} - \bar{d}_s)^2) = 0.0111$$

The sampling error variance, $VAR(s)$, is a difficult quantity to compute directly. Fortunately it can be approximated (Hunter and Schmidt 1990):

$$VAR(s) = \left(\frac{\bar{n} - 1}{\bar{n} - 3} \right) \left(\frac{4}{\bar{n}} \left(1 + \frac{\bar{d}_s^2}{8} \right) \right) = 0.0086$$

The estimated variance of the population effect size is the observed effect size variance minus the sampling error variance: $VAR(\bar{d}) = VAR(d_s) - VAR(s) = 0.0025$, and the standard error of the population effect is then $SE(\bar{d}) = 0.05$. This allows the construction of common confidence intervals for the effect size:

$$CI_{0.95,\ 2\text{-}tail}:\ 0.137 \pm 1.96(0.05) = [0.0391 : 0.2349]$$

$$CI_{0.99,\ 2\text{-}tail}:\ 0.137 \pm 2.575(0.05) = [0.0084 : 0.2657].$$

Note that each interval is bounded away from zero even though two of the five studies concluded that there exists no evidence to support a non-zero effect size (i.e. non-significance). This example shows how meta-analysis utilizes information from

a set of studies, but is not required to agree with the conclusions of any particular study. While many obvious simplifications were assumed in this example, the basic philosophy of meta-analysis in actual practice follows this logic. Hunter and Schmidt (1990) carefully review complications that can occur in practice such as heterogeneity of measurement error, sampling error, ranges on the variables, imperfect scale and variable construction, data problems across studies, and bias in correlations.

Conclusion

The basic problem with the null hypothesis significance test in political science is that it often does not tell political scientists what they think it is telling them. Most of the problems discussed here are interpretive in that they highlight misconceptions about the results of the procedure. From the current presentation of null hypothesis significance testing in published work it is very easy to confuse statistical significance with theoretical or substantive importance. It is also possible to have data which tells us something about an important political question, but which does not pass an arbitrary significance level threshold. In this circumstance, one would still want to know that there is empirical evidence of some phenomenon of interest. There is nothing mathematically wrong with p-values in these circumstances: they are simply not sufficient statistics for actual theoretical importance.

The interpretive problems described are not well known in political science. However, most of these have been discussed individually in statistics for some time. For instance as far back as 1959, Sterling (p. 30) considered the "publication bias" introduced by a fixation with p-values in the social sciences:

> There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs—an "error of the first kind"—and is published. Significant results published in these fields are seldom verified by independent replication. The possibility thus arises that the literature of such a field consists in substantial part of false conclusions resulting from errors of the first kind in statistical tests of significance.

Political science is highly empirical in its orientation, and therefore the device which declares particular results as substantial, or interesting, or worth publishing is a critical topic for discussion. Hopefully the null hypothesis significance test has not damaged accumulation of conclusions in the discipline to the degree in which Sterling predicts.

It is important to know that there exist effective alternatives which require only modest changes in empirical methodology: confidence intervals, Bayesian estimation, and meta-analysis. Confidence intervals are readily supplied by even the simplest of statistical computing packages, and require little effort to interpret. Bayesian estimation eliminates many of the pathologies described, albeit with a greater setup cost. Meta-analysis is sometimes a complex process, but it offers the

potential benefit of integrating and analyzing a wider scope of work on some political question.

So why is the null hypothesis significance test pervasive in political science? One reason is that it creates the illusion of objectivity by seemingly juxtaposing alternatives in an equivalent manner. Through naming conventions and procedural habits there is an implied fairness that does not really exist in making decisions about strengh of evidence. Furthermore, researchers fear publication rejection when there are an insufficient number of "stars" depicted in their tables. This problem is particularly damaging because it perpetuates the myth that more stars imply greater non-conditional strength of evidence, which is clearly wrong. There is also an impression among many political scientists that a testing process which is so pervasive and so long-lived cannot be seriously flawed. Destroying the misconception that the null hypothesis significance test is inherently sound because of its widespread and enduring use is the primary mission of this essay. We need at least to be aware that the undeniable longevity and popularity of the null hypothesis significance test are not a result of methodological integrity.

References

Bakan, David. 1960. "The Test of Significance in Psychological Research." *Psychological Bulletin* 66: 423-37.

Barnett, Vic. 1973. *Comparative Statistical Inference*. New York: Wiley.

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Berger, James O. 1984. "The Robust Bayesin Viewpoint (with discussion)." In Joseph B. Kadane, ed., *Robustness of Bayesian Analysis*. Amsterdam: North Holland.

Berger, James O., B. Boukai, and Y. Wang. 1997. "Unified Frequentist and Bayesian Testing of a Precise Hypothesis." *Statistical Science* 12: 133-60.

Berger, James O., and Thomas Sellke. 1987. "Testing a Point Null Hypothesis: The Irreconciliability of P Values and Evidence." *Journal of the American Statistical Society* 82: 112-22.

Bernardo, José M. 1984. "Monitoring the 1982 Spanish Socialist Victory: A Bayesian Analysis." *Journal of the American Statistical Society* 79: 510-15.

Brenner-Golumb, Nancy. 1993. "R. A. Fisher's Philosophical Approach to Inductive Inference." In G. Keren and C. Lewis, eds., *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Box, George E. P., and George C. Tiao. 1992. *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Carver, Ronald P. 1978. "The Case Against Statistical Significance Testing." *Harvard Education Review* 48: 378-99.

Casella, George, and Roger L. Berger. 1990. *Statistical Inference*. Belmont, CA: Wadsworth & Brooks/Cole.

Chow, Siu L. 1996. *Statistical Significance: Rationale, Validity and Utility*. London: Sage.

Cleveland, William. 1993. Visualizing Data. Summit, NJ: Hobart Press.

Cohen, Jacob. 1962. "The Statistical Power of Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65: 145-53.

____. 1977. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: Academic Press.

____. 1992. "A Power Primer." *Psychological Bulletin* 112: 115-59.

____. 1994. "The Earth is Round ($p < .05$)." *American Psychologist* December, 12: 997-1003.

Cooper, Harris. 1984. *The Integrative Research Review: A Systematic Approach*. Beverly Hills, Ca: Sage.

Falk, R., and C. W. Greenbaum. 1995. "Significance Tests Die Hard." *Theory and Psychology* 5: 396-400.

Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43, No. 2 (January): 169-95.

Fisher, Sir Ronald A. 1925a. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

____. 1925b. "Theory of Statistical Estimation." *Proceedings of the Cambridge Philosophical Society* 22: 700-25.

____. 1934. *The Design of Experiments*, 1st ed. Edinburgh: Oliver and Boyd.

____. 1955. "Statistical Methods and Scientific Induction." *Journal of the Royal Statistical Society* B, 17: 69-78.

____. 1971. *The Design of Experiments*, 9th ed. New York: Hafner Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. New York: Chapman & Hall

Gigerenzer, Gerd. 1987. "Probabilistic Thinking and the Fight Against Subjectivity." In Krüger, Lorenz, Gerd Gigerenzer, and Mary Morgan, eds., *The Probabilistic Revolution*, vol. 2. Cambridge, MA: MIT.

____. 1993. "The Superego, the Ego, and the Id in Statistical Reasoning." In G. Keren and C. Lewis, eds., *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, Gerd, and D. J. Murray. 1987. *Cognition as Intuitive Statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, Gerd, Zeno Swijtink, Theodore Porter, Lorraine Daston, John Beatty, Lorenz Krüger. 1989. *The Empire of Chance*. Cambridge: Cambridge University Press.

Gill, Jeff, and James Thurber. 1999. "Congressional Tightwads and Spendthrifts: Measuring Fiscal Behavior in the Changing House of Representatives." *Political Research Quarterly* 52: 387-402.

Glass, Gene V. 1976. "Primary, Secondary and Meta-Analysis of Research." *Educational Researcher* 5: 3-8.

Glass, G. V., B. McGaw, and M. L. Smith. 1981. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.

Greenwald, Anthony g. 1975. "Consequences of Prejudice Against the Null Hypothesis." *Psychological Bulletin* 82: 1-20.

Howson, Colin, and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach*, 2nd ed. Chicago: Open Court.

Hunter, John E. 1997. "Needed: A Ban on the Significance Test." *Psychological Science* January, Special Section 8: 3-7.

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Beverly Hills, CA: Sage.

Jacoby, William. 1997. "Statistical Graphics for Univariate and Bivariate Data." Beverly Hills, CA: Sage.

Jeffreys, Harold. 1961. *The Theory of Probability*. Oxford: Clarendon Press.

Karlin, S., and H. Rubin. 1956. "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio." *Annals of Mathematical Statistics* 27: 272-99.

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press.

____. 1995. "Replication, Replication." *PS: Political Science and Politics* 28, No. 3 (September): 443-99.

King, Gary, Michael Tomz, and Jason Wittenberg. 1998. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." Political Methodology Working Paper Archive: http://polmeth.calpoly.edu.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

____. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73, No. 1 (March): 31-43.

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*, 2nd ed. New York: Springer.

Lindsay, R. M. 1995. "Reconsidering the Status of Tests of Significance: An Alternative Criterion of Adequacy." *Accounting, Organizations and Society* 20: 35-53.

Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65, No. 3 (September): 682-93.

Macdonald, Ranald R. 1997. "On Statistical Testing in Psychology." *British Journal of Psychology* 88, No. 2 (May): 333-49.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Counseling and Clinical Psychology* 46: 806-34.

____. 1990. "Why Summaries of Research on Psychological Theories Are Often Uninterpretable." *Psychological Reports* 66: 195-244.

Meier, Kenneth. 1997. "The Value of Replicating Social-Science Research." *The Chronicle of Higher Education* 43 (February 7): 22.

Miller, Alan J. 1990. *Subset Selection in Regression*. New York: Chapman & Hall.

Neyman, Jerzy, and Egon S. Pearson. 1928a. "On the Use and Interpretation of Cer-

tain Test Criteria for Purposes of Statistical Inference. Part I." *Biometrika* 20A: 175-240.

____. 1928b. "On the Use and Interpretation of Certain Test Criteria for Purposes of of Statistical Inference. Part II." *Biometrika* 20A: 263-94.

____. 1933a. "On the Problem of the Most Efficient Test of Statistical Hypotheses." *Philosophical Transactions of the Royal Statistical Society* A 231: 289-337.

____. 1933b. "The Testing of Statistical Hypotheses in Relation to Probabilities *a priori*." *Proceedings of the Cambridge Philosophical Society* 24: 492-510.

____. 1936a. "Contributions to the Theory of Testing Statistical Hypotheses." Statistical Hypotheses." *Statistical Research Memorandum* 1: 1-37.

____. 1936b. "Sufficient Statistics and Uniformly Most Powerful Tests of Statistical Hypotheses." *Statistical Research Memorandum* 1: 113-37.

Oakes M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

Popper, Karl. 1968. *The Logic of Scientific Discovery*. New York: Harper & Row.

Pollard, P., and J. T. E. Richardson. 1987. "On the Probability of Making Type One Errors." *Psychological Bulletin* 102 (July): 159-63.

Press, S. James. 1989. *Bayesian Statistics: Principles, Models and Applications*. New York: Wiley.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." In Peter V. Marsden, ed., *Sociological Methodology*. Cambridge, MA: Blackwells.

Ripley, Brian D. 1987. *Stochastic Simulation*. New York: Wiley.

Rosnow, Ralph L., and Robert Rosenthal. 1989. "Statistical Procedures and the Justification of Knowledge in Psychological Science." *American Psychologist* 44: 1276-84.

Rozeboom, William W. 1960. "The Fallacy of the Null Hypothesis Significance Test." *Psychological Bulletin* 57: 416-28.

Schmidt, Frank L. 1996. "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for the Training of Researchers." *Psychological Methods* 1: 115-29.

Schmidt, Frank L., and John E. Hunter. 1977. "Development of a General Solution to the Problem of Validity Generalization." *Journal of Applied Psychology* 62: 529-40.

Sedlmeier, Peter, and Gerd Gigerenzer. 1989. "Do Studies of Statistical Power Have an Effect on the Power of Studies." *Psychological Bulletin* 105 (March): 309-16.

Serlin, Ronald C., and Daniel K. Lapsley. 1993. "Rational Appraisal of Psychological Research and the Good-enough Principle." In G. Keren and C. Lewis, eds., *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Stigler, Stephen M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Sterling, Theodore D. 1959. "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa." *Journal of the American Statistical Association* 54 (March): 30-34.

Tanner, Martin. 1996. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 3rd ed. New York: Springer-Verlag.

Western, Bruce. 1995. "A Comparative Study of Working-Class Disorganization: Union Decline in Eighteen Advanced Capitalist Countries." *American Sociological Review* 60 (April): 179-201.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88, No. 2 (June): 412-23.

Wolf, Frederic M. 1986. *Meta-Analysis: Quantitative Methods for Research Synthesis. Beverly Hills*, CA: Sage.

Received: May 27, 1998
Accepted: February 24, 1999
jgill@calpoly.edu