Full Text:

**THE PURPOSE OF HYPOTHESES**

Hypotheses are the central tool of scientific observation. Because the core method of scientific investigation is the comparison of expectations against observations of the world, scientists need to make clear statements about their expectations. A hypothesis is a concise, falsifiable statement that is subjected to observational testing as part of a scientific investigation.

Scientific research generally starts with a question about the observable world. In the social sciences research questions focus on human behavior—especially behavior related to groups (e.g., communities, countries, or societies). The scientific method says nothing about the origins of these research questions (just as it says nothing about the content of the areas of research). The scientific method simply requires that a scientist state an answer to this question (the hypothesis) that can be tested with observations (hypothesis testing).

There is a bewildering array of potential research questions—and thus hypotheses—in the domain of social science. Hypotheses can focus on expectations about voting behavior, the tendency of nations to go to war, or the factors that contribute to juvenile delinquency or to decisions about where to live (among many, many other hypotheses).

The purpose of the hypothesis is to ease the task of testing an expectation with observations of the world. A good hypothesis, then, is one that is easily tested. The ease of testing contributes to a second key aspect of the scientific method: reproducibility of testing. A clearly worded hypothesis can be tested repeatedly by a scientist and, maybe more important, by other scientists (King, Keohane, and Verba 1994, pp. 28–29).

Consider the following example. A social scientist may hypothesize that smaller class sizes in secondary schools will lead to higher performance on standardized tests. Because it is easy to observe the number of students in a class and the standardized tests scores are also easily observable (though there may be questions of the validity of the test as a measure of “intelligence” or even “academic achievement”), this hypothesis is easy to test. The test itself is also easy to replicate by the original social scientist or by other investigators. The hypothesis is sufficiently clear that any observer would be able to tell whether people in the smaller classes actually performed better on standardized tests. The judgment, then, is not a product of the specific observer but is instead independent of the identity of the scientist (a subject of some controversy that is discussed in a later section).

Full Text:

**STRATEGIES OF HYPOTHESIS TESTING—QUANTITATIVE STRATEGIES**

One of the major strategies for hypothesis testing is quantitative research. The focus of this approach is on the quantification of social science concepts for purposes of comparison and hypothesis testing. For example, a social scientist might ask whether the U.S. president’s approval ratings have gone down over the past year. This could give some sense of the power the president might have in promoting his or her legislative agenda or the chances of the president’s party in an upcoming election.

The hypothesis would be the social scientist’s guess as to how the candidate will fare against the proposed opponent. A good hypothesis will be one that is well grounded Page 542  |  [Top of Article](http://go.galegroup.com/ps/i.do?action=interpret&id=GALE%7CCX3045301066&v=2.1&u=apollo&it=r&p=GVRL&sw=w&authCount=1#contentcontainer)in the available theory on elections and that is testable against observable data (in this case a survey). The hypothesis would predict whether the approval ratings of the president have gone down over the past year. It would provide a preliminary answer to the stated research question. More ambitious hypotheses that predict specific levels of support (that the president has lost 8 percent of support from the previous year) are possible, but these require highly developed theories. One can take as an example the basic hypothesis that the president’s approval rating has gone down in the past year.

Armed with a hypothesis, the scientist will conduct a survey of a sample of potential voters to test the hypothesis. The scientist cannot, or would not want to, survey all citizens of the United States. Instead, the scientist will select a small sample out of the U.S. population. The scientist might send a survey to 1,000 citizens and see whether the hypothesis is correct within this sample of voters. The results of the survey will give the scientist a sense of the president’s current approval rating for comparison to previous approval ratings (Babbie 1995, pp. 190–193).

Quantitative hypothesis testing—the comparison of numerically represented measurements for purposes of hypothesis testing—allows for some detailed comparisons. One can say, for the sake of argument, that the survey suggests 43 percent of respondents said they approved of the job the president was doing. The previous year’s survey had reported that 47 percent of respondents had said they approved. At first glance, the evidence provides support for the hypothesis.

A number of questions remain about this test of the hypothesis. To what extent is the observed dip in approval indicative of a general trend in the U.S. population? To what extent is the dip indicative of a persistent change in approval? Tools of probability and statistics provide some opportunity to address these questions. Sampling theory provides some sense of how reliable the results are that come from a sample of a larger population (Babbie 1995, pp. 195–203). Such theory helps scientists describe the range of possible values in the population given the size of the sample surveyed. One could describe the probability that the actual approval rating was 47 percent (the previous rating) while the sample happened to be skewed toward lower approval ratings. In general, the larger the sample size, the lower the probability of these sorts of discrepancies. Such theory also provides insight into whether the variation one sees is relatively permanent or just part of the inherent variability in measuring people’s approval of a public figure. It is the ability to assess these issues of sampling and fundamental uncertainty that have convinced many of the utility of quantitative hypothesis testing techniques.

Full Text:

**STRATEGIES OF HYPOTHESIS TESTING—QUALITATIVE STRATEGIES**

Many scholars pursue an alternative style of hypothesis testing. These scholars tend to be unsatisfied with the techniques of measurement for social concepts employed in many quantitative research projects. In lieu of quantitative measurements of large samples of observations, qualitative hypothesis testing involves the careful study of a smaller number of observations with detailed treatment of the context and meaning of the social concepts themselves.

A qualitative hypothesis testing strategy follows the basic procedure of hypothesis testing. The social scientist generates a hypothesis in response to a research question. The social scientist then compares his or her expectation against the observed world. The difference between the qualitative approach and the quantitative approach reviewed earlier is in the strategy for getting reliable observations of the world.

Qualitative hypothesis testing tends to focus on detailed histories and culturally sensitive accounts of the social systems that are being studied. The detail and contextual knowledge provide the qualitative hypothesis testing strategies leverage on the challenges of hypothesis testing in two ways. First, the detailed knowledge of the subjects under study allow for careful selection of cases for study. As opposed to the quantitative strategy of multiplying the number of observations to avoid the possibility of drawing the wrong lessons from a study, qualitative hypothesis testing involves carefully selecting a few observations to achieve the ideal contrast. Second, the detailed knowledge of the subjects also allows for greater attention to the measurement of variables. Proponents of qualitative research focus on the ability to really get to know the subjects as a means to understand the nuances of the proposed effects of policies (e.g., Brady and Collier 2004).

Qualitative research tends to investigate different types of hypotheses than quantitative research, though the barriers between the two have eroded somewhat since the late twentieth century. Whereas quantitative hypotheses tend to involve statements of correlation, qualitative hypotheses have tended to focus on issues of necessary and sufficient conditions. These hypotheses focus on the conditions whose presence guarantees that an effect will be present (a sufficient condition) or whose absence will guarantee that an effect will not be present (a necessary condition) (Goertz and Starr 2003).

Theda Skocpol’s (1979) research on social revolutions exemplifies this approach. In *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*, Skocpol studies the factors that are essential to the success of peasant revolutions in a selection of countries. To study Page 543  |  [Top of Article](http://go.galegroup.com/ps/i.do?action=interpret&id=GALE%7CCX3045301066&v=2.1&u=apollo&it=r&p=GVRL&sw=w&authCount=1#contentcontainer)complicated processes such as social revolution and its relation to the political structures of regimes, Skocpol focuses her attention on the contrast between France, Russia, and China. These detailed cases are contrasted with control cases such as England and Prussia. The control cases serve the comparative role of the prior approval ratings in the quantitative example given earlier. This approach allows Skocpol to study each of the cases in great detail and to have confidence in the measurement of such concepts as types of revolution and various aspects of regime structure. The result is a widely praised multidimensional account of the necessary conditions of success for peasant revolutions.

Full Text:

**FUNDAMENTAL CONTROVERSIES IN HYPOTHESIS TESTING**

While there is debate over the relative merits of qualitative and quantitative hypothesis testing, there are also more fundamental critiques of hypotheses testing. The quantitative and qualitative hypotheses are different strategies to accomplish the same goal. In both approaches observations are compared against hypotheses about the essential nature of the social world. In the qualitative example Skocpol is testing hypotheses about the underlying nature of social revolution. This assumes that there is an underlying nature of social revolutions. Some critics of the hypothesis testing contend that there is no singular underlying social nature. These authors, mostly associated with post structuralism, argue that there is no singular structure of society about which one can generalize or that one can discover through repeated observation (for a famous statement of this argument, see Derrida 1978).

Other authors focus their criticism not on the absence of a stable world to observe, but instead on the tools that social scientists have to observe the world (assuming that such a stable world exists). These critics allege that social measurement is inherently filled with biases. Observation, these critics allege, is inseparable from the observer. If this is the case, especially given the importance of social values to humans, there is no such thing as neutral observation of the social world. The result is that all hypothesis tests are suspect. Many of these critics recommend exploring the social world through admittedly biased accounts and narratives rather than the “at-a-distance” observation implied by the typical hypothesis testing framework (Shank 2002).

# Test Statistics

[BIBLIOGRAPHY](http://go.galegroup.com/ps/i.do?action=interpret&id=GALE%7CCX3045302732&v=2.1&u=apollo&it=r&p=GVRL&sw=w&authCount=1#A)

Hypothesis testing or significance testing is undoubtedly one of the most widely used quantitative methodologies in empirical research in the social sciences. It is one viable way to use statistics to examine a hypothesis in light of observations or sample information. The starting point of hypothesis testing is specifying the hypothesis to be tested, called the *null hypothesis*. Then a test statistic is chosen to summarize the sample information, and its value is taken as an indication of the strength of sample evidence against the null hypothesis.

Modern hypothesis testing dates to the 1920s and the work of Ronald Aylmer Fisher (1890–1962) on the one hand, and Jerzy Neyman (1894–1981) and Egon Pearson (1895–1980) on the other. Fisher (1925) refers to hypothesis testing as *significance testing* (this entry does not distinguish between the two terms). In the Fisherian approach, the observed test statistic is converted to the *P* -value, which is the probability of obtaining the observed or more extreme value of the test statistic under the null model; the smaller the *P* -value, the stronger the sample evidence against the null hypothesis. An early example of Fisher’s significance testing was conducted in 1735 by the father and son Swiss mathematicians Daniel Bernoulli (1700–1782) and John Bernoulli (1667–1748). They tested for the random/uniform distribution of the inclinations of the planetary orbits. A detailed discussion of their original results and subsequent modifications of their results can be found in Anders Hald (1998).

In the Neyman and Pearsonian (1928, 1933) approach, an alternative hypothesis is specified and the null hypothesis is tested against this alternative hypothesis. The specification of an alternative hypothesis allows the computation of the probabilities of two types of error: Type I error (the error of falsely rejecting a null hypothesis) and Type II error (the error of incorrectly accepting a null hypothesis). Page 332  |  [Top of Article](http://go.galegroup.com/ps/i.do?action=interpret&id=GALE%7CCX3045302732&v=2.1&u=apollo&it=r&p=GVRL&sw=w&authCount=1#contentcontainer)Type I error is also referred to as the *significance level* of the test, and one minus Type II error the *power* of the test. Given that the two types of error cannot be minimized simultaneously, the common practice is to specify the level of significance or Type I error and then use a test that maximizes its power subject to the given significance level. In the Fisherian approach, the *P* -value is reported without necessarily announcing the rejection or nonrejection of the null hypothesis, whereas in the Neyman and Pearsonian approach, the null hypothesis is either rejected in favor of the alternative hypothesis or not rejected at the given significance level. E. L. Lehmann (1993) provides a more detailed comparison of the two approaches.

In empirical research, a mixture of the two approaches is typically adopted. Consider the linear regression model:

  

where   is the set of observations on the dependent variable *Y* and the explanatory variables X 2, …, XK, and *ε*iis the unobserved error term. The parameters *β*2, …, βKmeasure the *ceteris paribus* effects of the explanatory variables on the dependent variable. The significance of these effects is routinely tested by the *t* -tests and *F* -test. The *t* -test was discovered by William Sealy Gosset (1876–1937) for the mean of a normal population and extended by Fisher in 1925 to other contexts, including regression coefficients. Gosset’s result was published in *Biometrika* under the pseudonym “Student” in 1908. The *F* -test was originally developed by Fisher in the context of testing the ratio of two variances. Fisher pointed out many other applications of the *F* -test, including the significance of the complete regression model.

For a given *j* = 2, …, *K*, the null hypothesis for the corresponding *t* -test is *H* 0j-βj = 0 and the *t* -statistic is

  

where bj denotes the ordinary least squares estimator of βj and *se* (bj) denotes the standard error of bj. Note that if the null *H* *0j*is true, the explanatory variable Xij would be absent from the regression model (1) and thus considered to be insignificant in explaining the dependent variable given the presence of the other explanatory variables. This is why *t* -tests are referred to as tests for the significance of individual variables as opposed to the *F* -test, which tests for the significance of the complete regression. The null hypothesis for the *F* -test is

*H*0:*β*2=*β*3=…=*βk*=0.

There are several equivalent formulas for computing the *F* -statistic, one of which is

  

where *R* 2 is the coefficient of determination. Since under *H* 0, all the explanatory variables can be dropped from (1), the *F* -test is a test for the significance of the complete regression.

Much packaged computer software routinely calculates the *t* -statistics and the *F* -statistic. For a given sample, the observed value of tj(*F*) summarizes the sample evidence on the significance of the explanatory variable Xj (the significance of the regression (1)). To either convert the observed value of tj (*F*) to the *P* -value or make a binary decision on the rejection or nonrejection of the null hypothesis *H* *0j*(*H* 0) at a given significance level, the distribution of tj(*F*) under the corresponding null hypothesis is required. On the basis of the null hypothesis being true and further assumptions on the nature of the sample and on the normality of the error in (1), the distribution of tj is known to be Student’s *t* with (*K*-1) degrees of freedom, denoted as *t* [*K*-1], and the distribution of *F* is the so-called *F* -distribution with {(*K*-1),(*n-K*)} degrees of freedom denoted as *F* [*K*-1, *n-K*] (see Goldberger [1991] for details). The known distribution of tj(*F*) under the null hypothesis allows the computation of the *P* -value or the computation of the appropriate critical value at a prespecified significance level with which the observed test statistic can be compared.

Like *t* -tests and the *F* -test, standard tests rely on further assumptions in addition to the truth of the null hypothesis, such as the assumption of a random sample and the normality of the error term. These further assumptions may not be met in typical applications in social sciences, and modifications are required of tests designed on the basis of these assumptions. For example, when normality of the error term is not met, the distributions of the *t* -statistic and *F* -statistic are no longer *t* [*K-1*]or *F* [*K-1*, *n-K*]. Fortunately, their asymptotic distributions are known under general conditions and may be used to perform these tests. Alternatively, resampling techniques, such as the bootstrap and subsampling, may be used to approximate the distributions of the test statistics under the null hypothesis (see Efron and Tibshirani [1993] and Politis et al. [1999] for an excellent introduction to these methods).

The issue that has generated the most debate in hypothesis testing from the beginning is the choice of significance level (Henkel 1976). Given any value of the test statistic, one can always force nonrejection by specifying a low enough significance level or force rejection by choosing Page 333  |  [Top of Article](http://go.galegroup.com/ps/i.do?action=interpret&id=GALE%7CCX3045302732&v=2.1&u=apollo&it=r&p=GVRL&sw=w&authCount=1#contentcontainer)a high enough significance level. Although reporting the *P* -value partly alleviates this arbitrariness in setting the significance level, it is desirable to report estimates of the parameters of interest and their standard errors or confidence intervals so that the likely values of the unknown parameters and the precision of their estimates can be assessed.