where Var W_n is the variance of W_n and Z is a standard normal random variable. We now have the basis for an approximate test, for example, we would reject $H_0: \theta \leqslant \theta_0$ at level 0.05 if $(W_n - \theta_0)/\sqrt{\mathrm{Var}\, W_n} > 1.645$. Note that Var W_n could depend on θ_0 and we can still use it in the test statistic. This type of test, where we use the actual variance of W_n, is called a *score test*.

If Var W_n also depends on unknown parameters we could look for an estimate S_n^2 of Var W_n with the property that $(\mathrm{Var}\, W_n)/S_n^2$ converges in probability to one. Then, using Slutsky's Theorem (see Casella and Berger 2001, Sect. 5.5), we can deduce that $(W_n - \theta)/S_n$ also converges in distribution to a standard normal distribution. The large-sample test based on this statistic is called a *Wald test*.

4. Conclusions

Hypothesis testing is one of the most widely used, and some may say abused, methodologies in statistics. Formally, the hypotheses are specified, an α-level is chosen, a test statistic is calculated, and it is reported whether H_0 or H_1 is accepted. In practice, it may happen that hypotheses are suggested by the data, the choice of α-level may be ignored, more than one test statistic is calculated, and many modifications to the formal procedure may be made. Most of these modifications cause bias and can invalidate the method. For example, a hypothesis suggested by the data is likely to be one that has 'stood out' for some reason, and hence H_1 is likely to be accepted unless the bias is corrected for (using something like Scheffe's method—see Hsu 1996).

Perhaps the most serious criticism of hypothesis testing is the fact that, formally, it can only be reported that either H_0 or H_1 is accepted at the prechosen α-level. Thus, the same conclusion is reached if the test statistic only barely rejects H_0 and if it rejects H_0 resoundingly. Many feel that this is important information that should be reported, and thus it is almost required to also report the p-value of the hypothesis test.

For further details on hypothesis testing see the classic book by Lehmann (1986). Introductions are also provided by Casella and Berger (2001) or Schervish (1995), and a good introduction to multiple comparisons is Hsu (1996); see also *Hypothesis Tests, Multiplicity of.*

See also: Explanation: Conceptions in the Social Sciences; Hypothesis Testing: Methodology and Limitations

Bibliography

Berger R L 1982 Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**: 295–300

Casella G, Berger R L 2001 *Statistical Inference*, 2nd edn. Wordsworth/Brooks Cole, Pacific Grove, CA

Hsu J C 1996 *Multiple Comparisons, Theory and Methods.* Chapman & Hall, London

Hwang J T, Casella G, Robert C, Wells M T, Farrell R H 1992 Estimation of accuracy in testing. *Annals of Statistics* **20**: 490–509

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

Schervish M 1995 *Theory of Statistics.* Springer, New York

G. Casella and R. L. Berger

Hypothesis Testing: Methodology and Limitations

Hypothesis tests are part of the basic methodological toolkit of social and behavioral scientists. The philosophical and practical debates underlying their application are, however, often neglected. The fruitful application of hypothesis testing can benefit from a clear insight into, the underlying concepts and their limitations.

1. The Idea of Hypothesis Testing

A test is a statistical procedure to obtain a statement on the truth of falsity of a proposition, on the basis of empirical evidence. This is done within the context of a model, in which the fallibility or variability of this empirical evidence is represented by probability. In this model, the evidence is summarized in observed data, which is assumed to be the outcome of a stochastic, i.e., probabilistic, process; the tested proposition is represented as a property of the probability distribution of the observed data.

1.1 Some History

The first published statistical test was by John Arbuthnot in 1710, who wondered about the fact that in human births, the fraction of boys born year after year appears to be slightly larger than the fraction of girls (cf. Hacking 1965). He calculated that this empirical fact would be exceedingly unlikely (he obtained a probability of 1/483 600 000 000 000 000 000 000) if the probability of a male birth were exactly 0.5, and argued that this was a proof of divine providence, since boys—some of whom will be soldiers—have a higher risk of an early death, so that a higher ratio of male births is needed to obtain an equal ratio of males among young adults. We see here the basic elements of a test: the proposition that the male birth ratio is 0.5, related to data by regarding these as outcomes of stochastic variables, the calculation that the data would be unlikely if the proposition were true, and a further conclusion interpreting the falsity of the proposition.

One of the first statistical procedures that comes close to a test in the modern sense was proposed by Karl Pearson in 1900. This was the famous chi-squared test for comparing an observed frequency distribution to a theoretically assumed distribution. Pearson derived the now well-known statistic to test the proposition, or *hypothesis*, that the probabilities of the various possible (finitely many) outcomes of some random variable are given by certain preassigned numbers. Pearson proved that this statistic has (in a large-sample approximation) the chi-squared distribution; this distribution can therefore be used to calculate the probability that, if the hypothesis holds, the test statistic will assume a value equal to or larger than the value actually observed.

The idea of testing was further codified and elaborated in the first decades of the twentieth century, mainly by R. A. Fisher (e.g., 1925). In his significance tests the data are regarded as the outcome of a random variable X (usually a vector or matrix), which has a probability distribution which is a member of some family of distributions; the tested hypothesis, also called the null hypothesis, is an assertion which defines a subset of this family; a test statistic $T = t(X)$, which is a function of X, is used to indicate the degree to which the data deviate from the null hypothesis; and the significance of the given outcome of the test statistic is calculated as the probability, if the null hypothesis is true, to obtain a value of T which is at least as high as the given outcome. If the probability distribution of T is not uniquely determined by the null hypothesis, then the significance is the maximal value of this probability, for all distributions of T compatible with the null hypothesis. The significance probability is now often called the p-value (the letter p referring to probability). With Fisher originates the convention to consider a statistical testing result as 'significant' if the significance probability is 0.05 or less—but Fisher was the first to recognize the arbitrary nature of this threshold.

A competing approach was proposed in 1928 by J. Neyman and Egon Pearson (the son of Karl). They criticized the arbitrariness in Fisher's choice of the test statistic and asserted that for a rational choice of test statistic one needs not only a null hypothesis but also an alternative hypothesis, which embodies a proposition that competes with the proposition represented by the null hypothesis. They formalized the testing problem as a two decision problem. Denoting the null hypothesis by H_0 and the alternative H_1, the two decisions were represented as 'reject H_0' and 'do not reject H_0.' (The second decision was also called 'accept H_0,' but it will be discussed below that this is an unfortunate term.) Two errors are possible: rejecting a true H_0, and failing to reject a false H_0. Neyman and Pearson conceived of the null hypothesis as a standard situation, the burden of proof residing with the researcher to demonstrate (if possible) the untenability of this proposition.

Correspondingly, they called the error of rejecting a true H_0 an error of the first kind and the error of failing to reject a false H_0 an error of the second kind. Errors of the first kind are considered more serious than errors of the second kind. The probability of—correctly—rejecting H_0 if H_1 is true, which is 1 minus the probability of an error of the second kind, given that the alternative hypothesis is true, they called the power of the test. Neyman and Pearson proposed the requirement that the probability of an error of the first kind, given that the null hypothesis is indeed true, do not exceed some threshold value called the significance level usually denoted by α. Further they proposed to determine the test so that, under this essential condition, the power will be maximal.

In the Neyman–Pearson formulation, we obtain richer results (namely, specific precepts for constructing good tests) at the cost of a more demanding model. In addition to Fisher's null hypothesis, we also need to specify an alternative hypothesis; and we must conceive the testing problem as a two-decision situation. This led to vehement debate between Fisher on the one hand, and Neyman and E. Pearson on the other. This debate and the different philosophical positions are summarized by Hacking (1965) and Gigerenzer et al. (1989), who also give a further historical account. The latter study also discusses how this controversy was resolved in the teaching and practice of statistics in the social sciences by a hybrid theory which combines elements of both approaches to testing, and which has been treated often as an objective precept for the scientific method, glossing over the philosophical controversies.

Examples of this hybrid character are that, in accordance with the Neyman–Pearson approach, the theory is explained by making references to both the null and the alternative hypotheses, and to errors of the first and second kind (although power tends to be treated in a limited and often merely theoretical way), whereas—in the spirit of Fisher—statistical tests are regarded as procedures to give evidence about the particular hypothesis tested and not merely as rules of behavior that will in the long run have certain (perhaps optimal) error rates when applied to large numbers of hypotheses and data sets. Lehmann (1993) argues that indeed a unified formulation is possible, combining the best features of both approaches.

Instead of implementing the hypothesis test as a 'reject/don't reject' decision with a predetermined significance level, another approach often is followed: to report the p-value or significance probability, defined as the smallest value of the significance level at which the observed outcome would lead to rejection of the null hypothesis. Equivalently, this can be defined as the probability, calculated under the null hypothesis, of observing a result deviating from the null hypothesis at least as much as the actually observed result. This deviation is measured by the test statistic, and the *p*-value is just the tail probability of the test statistic. For a given significance level α, the null hypothesis is rejected if and only if $p \leqslant \alpha$.

2. Hypothesis Tests in Empirical Research

2.1 The Role of the Null Hypothesis

In the social and behavioral sciences, the standard use of hypothesis testing is directed at single research questions, practically always fragments of larger investigations, about the existence of some or other effect. This effect could be a difference between two group means, the effect of some explanatory variable on some dependent variable as expressed by a regression coefficient in multiple linear regression analysis, etc. Expressed schematically, the researcher would like to demonstrate a research hypothesis which states that the effect in question, indeed, exists. The negation of this hypothesis, then, is understood as the proposition that the effect is absent, and this proposition is put up as the null hypothesis. The research hypothesis, stating that the effect exists, is the alternative hypothesis. Rejecting the null hypothesis is interpreted as support for the existence of the hypothesized effect. In this way, the burden of proof rests with the researcher in the sense that an error of the first kind is to state that there is evidence supporting the existence of the effect if, in reality, the effect is absent. The usual welcoming phrase for the rejection of the null hypothesis is the statement that a significant result has been obtained.

2.2 Example: The t-test

Consider, as a major example, the proposition that two groups, arbitrarily labeled *A* and *B*, are different with respect to some numerical characteristic *X*. Available data are measurements X_{Ai} of this characteristic for some individuals *i* in group *A* and measurements X_{Bi} of the characteristics for other individuals *i* in group *B*. The first step in modeling this is to consider the observed values *X* as outcomes of random variables, usually with the assumption that they are stochastically independent and have a probability distribution not depending on the individual *i*. The next step is, usually, to focus on the expected values, i.e., population means in the two groups, denoted here by μ_A and μ_B. The tested proposition is formalized as the statement that μ_A and μ_B are different (two-sided alternative hypothesis) or as the statement that one (e.g., μ_A) is bigger than the other (e.g., μ_B) (one-sided alternative hypothesis). The null hypothesis H_0 is defined as the statement that μ_A and μ_B are equal.

The most commonly used test for this testing problem is *Student's t-test*, called after the pseudonym of W. Gosset, who laid the mathematical basis for this test in 1908. The test statistic is

$$T = \frac{M_A - M_B}{\sqrt{\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right) S^2}}$$

where M_A and M_B are the two sample means, S^2 is the pooled within-group variance, and n_A and n_B are the two sample sizes. (Definitions of these quantities can be found in any statistics textbook.) This formula illustrates the property of many test statistics that an observed effect (here the difference between the two sample means) is compared to a measure of variability (based here on the within-group variance). Student/Gosset showed that, if the variable *X* has a normal distribution with the same variance in the two groups, then *T* has, if the null hypothesis is true, the so-called *t* distribution on $df = n_A + n_B - 2$ degrees of freedom. For the two-sided alternative hypothesis, H_0 is rejected for large values of the absolute value of *T*, for the one-sided hypothesis H_0 is rejected for large values of *T* itself. The threshold beyond which H_0 is rejected is called critical value, and is determined from the *t*-distribution in such a way that the significance level has the pre-assigned value. The one-sided threshold at significance level α is denoted $t_{df;\,\alpha}$, so that the 'rejection region' for the one-sided *t*-test is given by $\{T > t_{df;\,\alpha}\}$. The power of the one-sided test is larger than that of the two-sided test for $\mu_A > \mu_B$ (corresponding to the one-sided alternative hypothesis) and smaller for $\mu_A < \mu_B$. This is, in the Neyman–Pearson formulation, the reason for using the one-sided test if the alternative hypothesis is one-sided (in which case parameter values $\mu_A < \mu_B$ are not taken into consideration).

The conventional value of the significance level in the social and behavioral sciences is 0.05. For large values of the combined sample size $n_A + n_B$, the critical value of the *t*-test approximates the critical value that can be computed from the standard normal distribution. This is because the fact that the variance is estimated and not known beforehand becomes less and less important as the combined sample size gets larger; if the population variance was known beforehand and substituted for S^2, the test statistic would have the standard normal distribution, and the test would be the so-called *z*-test. For this test, the critical value is 1.645 for the one-sided test and 1.960 for the two-sided test. The power depends on many quantities: the sample sizes, the population means and variances, and the significance level. As examples, the power of the one-sided *z*-test for $\alpha = 0.05$ is equal to 0.50 for

$$\frac{\mu_A - \mu_B}{\sigma} = 1.645\sqrt{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

where σ is the within-group standard deviation. The power is equal to 0.95 if $(\mu_A - \mu_B)/\sigma$ is equal to twice this value.

2.3 The Role of Assumptions

The probability statements that are required for statistical tests do not come for free, but are based on certain assumptions about the observations used for the test. In the two-sample *t*-test, the assumptions are that the observations of different individuals are outcomes of statistically independent, normally distributed, random variables, with the same expected value for all individuals within the same group, and the same variance for all individuals in both groups. Such assumptions are not automatically satisfied, and for some assumptions it may be doubted whether they are ever satisfied exactly. The null hypothesis H_0 and alternative hypothesis H_1 are statements which, strictly speaking, imply these assumptions, and which therefore are not each other's complement. There is a third possibility: the assumptions are invalid, and neither H_0 nor H_1 is true. The sensitivity of the probabilistic properties of a test to these assumptions is referred to as the lack of robustness of the test. The focus of robustness studies has been on the assumptions of the null hypothesis and the sensitivity of the probability of an error of the first kind to these assumptions, but studies on robustness for deviations from assumptions of the alternative hypothesis have also been done, cf. Wilcox (1998).

One general conclusion from robustness studies is that tests are extremely sensitive to the independence assumptions made. Fortunately, those assumptions are often under control of the researcher through the choice of the experimental or observational design. Traditional departures from independent observations are multivariate observations and within-subject repeated measures designs, and the statistical literature abounds with methods for such kinds of dependent observations. More recently, methods have been developed for clustered observations (e.g., individual respondents clustered within groups) under the names of multilevel analysis and hierarchical linear modeling.

Another general conclusion is that properties of tests derived under the assumption of normal distributions, such as the *t*-test, can be quite sensitive to outliers, i.e., single, or a few, observations that deviate strongly from the bulk of the observations. Since the occurrence of outliers has a very low probability under normal distributions, they are 'assumed away' by the normality assumption. The lack of robustness and sensitivity to outliers have led to three main developments.

First, there are non-parametric tests, which do not assume parametric families of distributions (such as the normal). Most of these are based on the ranks of the observations rather than their numerical values. They are standard fare in statistical textbooks. Second, robust tests have been developed, based on numerical values of the observations, which are less sensitive to outliers or heavy-tailed distributions, e.g., by some kind of automatic downweighting of outlying observations (e.g., Wilcox, 1998). The purpose of such tests is to retain a high power as much as possible while decreasing the sensitivity to deviations from the normal distribution or to departures from other assumptions.

Third, diagnostic means have been developed to spot single observations, or groups of observations, with an undue influence on the result of the statistical procedure (e.g., Cook and Weisberg 1999, Fox 1991). The idea behind most of these diagnostics is that most of the observations come from a distribution which corresponds well enough to the assumptions of the statistical model, but that the observations could be contaminated by a small number of poorly fitting observations. Ideally, these observations should also be recognizable by close inspection of the data or data collection process. After deleting such poorly fitting observations one can proceed with the more traditional statistical procedures, assuming that the remaining data do conform to the model assumptions.

3. Confidence Intervals

Confidence intervals are closely related to hypothesis tests. They focus on a parameter in the statistical model. Examples of such parameters are, in the two-sample situation described above, the difference of the two population means, $\mu_A - \mu_B$, or the within-group standard deviation, σ. To frame this in general terms, consider a one-dimensional statistical parameter denoted by θ. A null hypothesis test is about the question whether the data is compatible with the hypothesis that the parameter θ (in this case $\mu_A - \mu_B$) has the particular value 0. Instead, one could put forward the question: what are the values of θ with which the data is compatible? This question can be related to hypothesis testing by considering the auxiliary problem of testing the null hypothesis $H_1\colon \theta = \theta_0$ against the alternative hypothesis $H_1\colon \theta \neq \theta_0$, for an arbitrary but fixed value θ_0. The data may be said to be compatible with any value of θ_0 for which this null hypothesis would not be rejected at the given level of significance. Thus, the confidence interval is the interval of non-rejected null hypotheses.

Another way of viewing the confidence interval is as follows. A confidence coefficient $1 - \alpha$ is determined by the researcher (a conventional value is 0.95). The confidence interval is an interval with lower boundary L and upper boundary U, calculated from the data and therefore being random variables, with the property that the probability that $L \leqslant \theta \leqslant U$, i.e., the interval contains the true parameter value, is at least $1-\alpha$. It can be proved mathematically that the interval of non-rejected null hypotheses has precisely this property.

The confidence interval can be inverted to yield a hypothesis test for any null hypothesis of the form H_0: $\theta=\theta_0$ in the following way. For the most usual null hypothesis that $\theta=0$, the hypothesis is rejected if the value 0 is not included in the confidence interval. More generally, the null hypothesis $H_0{:}\,\theta=\theta_0$ is rejected whenever θ_0 is outside the confidence interval. This shows that a confidence interval is strictly more informative than a single hypothesis test.

4. *Problems in the Interpretation and Use of Hypothesis Tests*

While hypothesis tests have been applied routinely by hosts of social and behavioral scientists, there have been continuing debates, sometimes vehement and polemic, about their use, among those inclined to philosophical or foundational discussion. Many contentious issues can be found in Harlow et al. (1997), Nickerson (2000), and Wilkinson and TFSI (1999). Within the context of the present contribution I can only give a personally colored discussion of some main points which might contribute to a better understanding and a more judicious application of hypothesis tests.

The reason that there has been so much criticism of the nature and use of hypothesis tests, is in my view the difficulty of reasoning with uncertain evidence and the natural human tendency to take recourse to behavioral rules instead of cogitate afresh, again and again, about how to proceed. The empirical researcher would like to conclude unequivocally whether a theory is true or false, whether an effect is present or absent. However, experimental and observational data on human behavior are bound to be so variable that the evidence produced by these data is uncertain. Our human tendency is to reason—if only provisionally—as if our conclusions are certain rather than tentative; to be concise rather than accurate in reporting our arguments and findings; and to use simple rules for steering how we make inference from data to theories. We are tempted to reason as if '$p>\alpha$' implies that the investigated effect is nil, and the presumed theory is false, while '$p\leqslant\alpha$' proves that the effect indeed exists and the theory is true.

However, such cognitive shortcuts amount to serious misuse of hypothesis testing. Much of the criticism of the use of hypothesis testing therefore is well-founded. In my view this implies that better interpretations of hypothesis testing should be promoted, and that tests should be used less mechanically and combined with other argumentations and other statistical procedures. In any case, the user of tests (or other statistical procedures) should realize that these procedures are based on the assumption that there is random variability in the data which cannot be completely filtered out of the results. Whether a test result is significant or not depends partly on chance, and each researcher obtaining a significant result should be aware that this could be an error of the first kind, just as a non-significant result could be an error of the second kind.

4.1 *Some Misinterpretations*

The following list tries to correct misinterpretations that have haunted applications of statistical tests. A more extensive discussion can be found, e.g., in Nickerson (2000).

(a) A common misinterpretation is that non-rejection implies support for the null hypothesis. Nonrejection should be interpreted, however, as an undecided outcome: there is not enough evidence against the null hypothesis, but this does not imply that there is evidence for the null hypothesis. Maybe the sample size is small, or error variability is large, so that the data does not contain much information anyway. Usually the alternative hypothesis contains probability distributions that approximate the null distribution; e.g. when testing $\mu_A-\mu_B=0$ against $\mu_A-\mu_B\neq 0$, the difference between the two population means can be tiny, while the alternative hypothesis is still true. In other words, the power of practically all tests is quite low in a sizeable part of the alternative hypothesis, so if one of the distributions in this part would prevail, chances would be low of rejecting the null hypothesis. Therefore, nonrejection provides support for those parts of the alternative practically as strongly as for the null hypothesis itself, and non-rejection may not be interpreted as support for the null hypothesis and against the alternative hypothesis.

(b) If one wishes to get more information about whether a nonsignificant result provides support for the null hypothesis, a power study is not the answer. Statistical power is the probability to reject the null hypothesis, if a given effect size obtains. Its interpretation cannot be inverted as being the degree of support of the null hypothesis in the case of nonsignificance. Power studies are important in the stage of planning a study. Once the study has been conducted, and one wishes to see in more detail to what extent the null and alternative hypotheses are supported, a confidence interval is the appropriate procedure (also see Wilkinson and TFSI, 1999, p. 596).

(c) Test results tell us nothing about the probabilities of null or alternative hypotheses. The probabilities known as the significance level or the power of the test are probabilities of certain sets of outcomes given the condition that the null or, respectively, the alternative hypothesis is true.

(d) A significant result is not necessarily important. Very low p-values do not in themselves imply large effect sizes. A small estimated effect size still can yield a low p-value, e.g., if the corresponding standard error is very small, which can be caused by a large sample or by good experimental control. Investigation of effect sizes is important, and is a different thing than hypothesis testing. The importance of reporting effect sizes is stressed in Schmidt (1996) and Wilkinson and TFSI (1999).

(e) The alternative hypothesis is not the same as a scientific theory. If a researcher is investigating some theory empirically, this can be based on tests of which the alternative hypotheses are deduced from the theory, but always under the assumption that the study was well designed and usually under additional technical assumptions such as, e.g., normality of the distributions. Since the alternative hypothesis is only a consequence and not a sufficient condition for the theory, it is possible that the alternative hypothesis is true but the theory is not. To find support for theories, one has to find corroborations of many consequences of the theory under various circumstances, not just of one of its implications. On the other hand, it is possible that the theory is true but the alternative hypothesis deduced from it is not, e.g., because the experiment or observation was ill-designed or because the auxiliary technical assumptions are not satisfied.

(f) The null hypothesis is just what it says: a hypothesis. The test is not invalidated by the mere fact that there are other reasons for thinking, or knowing, that the null hypothesis is wrong. Even if there may be other reasons for which the null hypothesis is wrong, still it can be sensible to check whether the data at hand are, or are not, compatible with it.

4.2 Limitations of the Neyman–Pearson Approach to Hypothesis Testing

The Neyman–Pearson (1928) formulation changed the concept of hypothesis testing because it provided a rational basis for the choice of a test statistic. This was obtained at the cost of a quite narrow frame for the notion of a hypothesis test: a two-decision problem in which one error ('type I') is considered much more serious than the other error ('type II'). This model for decision-making is often inappropriate and, when it is useful, it is more commonly a useful approximation than a faithful representation of the actual decision situation. Some of its limitations are discussed briefly in this section.

(a) Even when a dichotomous decision must be taken as a result of a scientific investigation, it is possible that both types of error should be considered equally serious. This is the case, e.g., when two therapies are compared and the most promising one must be chosen for further study or for practical application. If the costs of both kinds of error (choose therapy B, while A is better; or choose therapy A, while B is better) are about the same, then there is no asymmetry between the two competing hypotheses, and a significance test is not in order.

(b) Often there are more than two outcomes of the decision situation. For example, in the two-sample situation discussed above, where the population means are μ_A and μ_B, and where the researcher entertains a theory implying that $\mu_A > \mu_B$, it is natural to define three decision outcomes defined as '$\mu_A > \mu_B$ (support for the theory),' 'undecided,' and '$\mu_A < \mu_B$ (evidence against the theory).' Usually, however, such a situation is represented by a Neyman–Pearson testing problem with null hypothesis H_0: '$\mu_A = \mu_B$' and alternative H_1: '$\mu_A > \mu_B$.' If the t-statistic for $\mu_A - \mu_B$ yields a strongly negative outcome ('significance in the wrong direction') then the researcher only concludes that the null hypothesis is not rejected, whereas the data conveys the message that the alternative hypothesis should be rejected. The data then have to be considered as evidence against the theory embodied in H_1. Such a three-decision formulation is closer to the actual purpose of hypothesis tests in many situations where a one-dimensional parameter is being tested, cf. Leventhal (1999).

(c) The Neyman–Pearson approach is formulated as if the data at hand is the only evidence available about the null and alternative hypotheses: it is an *in vacuo* formulation. This is almost never realistic. Often the propositions under study are theoretically plausible or implausible, there is earlier empirical evidence, there may be converging evidence (as in the case of triangulation studies) in the same investigation, and other evidence may become available in the future. In the light of this surrounding evidence, it is unrealistic—perhaps arrogant—to presume that the researcher has to take either of two decisions: 'reject H_0' or 'do not reject H_0.' It is more realistic to see the data at hand as part of the evidence that was and will be accumulated about these hypotheses. This is a major reason why presenting the result of the test as a p-value is more helpful to other researchers than presenting only the 'significant–nonsignificant' dichotomy (also see Wilkinson and TFSI 1999).

There are several ways in which the results obtained from the data at hand, perhaps summarized in the p-value, can be combined with other evidence. In the first place, an informal combination can be and often will be made by a sensible, nonformalized weighing of the various available pieces of evidence. One formal way of combining evidence is provided by the Bayesian paradigm (e.g., Gelman et al. 1995). This paradigm presupposes, however, that the prior and current ideas of the researcher about the plausibility of all possible values of the statistical parameter under study be represented in a probability distribution for this parameter, a very strong requirement indeed. Other formal ways of combining evidence have been developed and are known collectively by the name of *meta-analysis* (e.g., Cook et al. 1992, Schmidt 1996). These methods require that the several to-be-combined hypothesis tests address the same substantive hypothesis, or that they can be regarded as tests of the same substantive parameter.

5. *Conclusion*

The statistical hypothesis test is one of the basic elements of the toolbox of the empirical researcher in the social and behavioral sciences. The difficulty of reasoning on the basis of uncertain data has led, however, to many misunderstandings and unfortunate applications of this tool. The Neyman–Pearson approach is a mathematically elegant formulation of what a hypothesis test could be. Its precise formulation rarely applies literally to the daily work of social and behavioral scientists, very often it is a useful approximation to a part of the research question, sometimes it is inappropriate—but sacrosanct it is not. A hypothesis test evaluates the data using a test statistic set up to contrast the null hypothesis with the alternative hypothesis, and the p-value is the probability to obtain, *if* the null hypothesis is true, outcomes of the test statistic that are at least as high as the outcome actually observed. Low values therefore are evidence against the null hypothesis, contrasted with the alternative hypothesis. It is more conducive to the advance of science to report p-values than merely whether the hypothesis was rejected at the conventional 0.05 level of significance.

In the interpretation of test outcomes one should be aware that these are subject to random variability; and that the probability calculations are based on assumptions which may be in error. Incorrect assumptions can invalidate the conclusions seriously. Nonparametric and robust statistical procedures have been developed which depend less on these kinds of assumption, and diagnostics have been developed for checking some of the assumptions. When reporting empirical results, it is usually helpful to report not only tests but also estimated effect sizes and/or confidence intervals for important parameters. To sum up: 'It seems there is no escaping the use of judgment in the use and interpretation of statistical significance tests' (Nickerson 2000, p. 256).

See also: Hypothesis Testing in Statistics; Hypothesis Tests, Multiplicity of; Model Testing and Selection, Theory of; Significance, Tests of

Bibliography

Cook R D, Weisberg S 1999 *Applied Regression Including Computing and Graphics*. Wiley, New York

Cook T D, Cooper H, Cordray D S, Hartmann H, Hedges L V, Light R J, Louis T A, Mosteller F 1992 *Meta-analysis for Explanation. A Casebook*. Russell Sage Foundation, New York

Fisher R A 1925 *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, UK

Fox J 1991 *Regression Diagnostics*. Sage, Newbury Park, CA

Gelman A, Carlin J B, Stern H S, Rubin D B 1995 *Bayesian Data Analysis*. Chapman & Hall, London

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L 1989 *The Empire of Chance*. Cambridge University Press, New York

Hacking I 1965 *Logic of Statistical Inference*. Cambridge University Press, Cambridge, UK

Harlow L L, Mulaik S A, Steiger J H (eds.) 1997 *What if there were no Significance Tests?* Erlbaum, Mahwah, NJ

Lehmann E L 1993 The Fisher, Neymann–Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* **88**: 1242–9

Leventhal L 1999 Updating the debate on one- versus two-tailed tests with the directional two-tailed test. *Psychological Reports* **84**: 707–18

Neyman J, Pearson E S 1928 On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A**: 175–240, 263–94

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

Schmidt F L 1996 Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods* **1**: 115–29

Wilcox R R 1998 The goals and strategies of robust methods (with discussion). *British Journal of Mathematical and Statistical Psychology* **51**: 1–64

Wilkinson L, Task force on statistical inference 1999 Statistical methods in psychology journals. *American Psychologist* **54**: 594–604

T. A. B. Snijders

Hypothesis Tests, Multiplicity of

Standard procedures for controlling the probability of erroneous statistical inferences apply only for a single comparison or determination from a given set of data. Multiplicity is encountered whenever a data analyst draws more than one inference from a data set. With multiplicity, standard procedures must be adjusted for the number of comparisons or determinations that actually or potentially are performed. If some adjustment is not made, the probability of erroneous inferences can be expected to increase with the degree of multiplicity. In other words, the data analyst loses control of a critical feature of the analysis.

Questions arising from problems of multiplicity raise a diversity of issues, issues that are important, difficult, and not fully resolved. This account empha-

International Encyclopedia of the Social & Behavioral Sciences ISBN: 0-08-043076-7