From Reid, N., 2015. Significance, Tests of. In: James D. Wright (editor-in-chief), International Encyclopedia of the Social & Behavioral Sciences, 2nd edition, Vol 21. Oxford: Elsevier. pp. 957–962.
ISBN: 9780080970868

Significance, Tests of

Nancy Reid, University of Toronto, Toronto, ON, Canada

Abstract

Tests of significance are used in statistics to assess the agreement between data and models. This article describes how such tests are formulated and used, illustrates this with several examples, and discusses some difficulties in interpretation that have been raised.

A test of significance assesses the agreement between the data and a hypothesized statistical model. The magnitude of the agreement is expressed as an observed level of significance, or *p*-value, which is the probability of obtaining data as or more extreme than the observed data, if the hypothesized model were true. A very small *p*-value suggests that either the observed data are not compatible with the hypothesized model or an event of very small probability has been observed. A large *p*-value indicates that the observed data are compatible with the hypothesized model. As the *p*-value is a probability, it must be between 0 and 1, and a very common convention is to declare values smaller than 0.05 as 'small' or 'statistically significant' and values larger than 0.05 as 'not statistically significant.' The historical rationale for this very arbitrary cut-off point is that the calculation of a *p*-value was difficult, and tables useful for common statistical models were prepared for general use.

To make this more concrete, it is necessary to consider statistical models and the notion of a (simplifying) hypothesis within that model. The theory of this is outlined in Section Model-Based Inference, with some more specific points considered in Section Further Topics. Section Difficulties with Significance Tests provides some brief comments on criticisms of significance tests. This introduction concludes with a highly idealized example to convey the idea of data being inconsistent with a hypothesized model.

Example 1. Students in a statistics class partake in an activity to assess their ability to distinguish between two competing brands of cola, and to identify from taste alone their preferred brand. Each of the 20 students expresses a preference for one brand or the other, but just one student claims to be able to discriminate perfectly between the two. Twenty cups of each brand are prepared by the instructor and labeled '1' and '2.' Each student is presented with a pair of cups and asked to record which label corresponds to Brand A. The result is that 12 students correctly identify the competing brands, although the student who claimed a perfect ability to discriminate identified the brands incorrectly.

The labeling of the cups as 1 or 2 by the instructor was completely random, i.e., cup 1 was equally likely to contain Brand *A* or *B*. The students did not discuss their opinions with their classmates, and the taste testing was completed fairly quickly. Under these conditions, it is plausible that each student has a probability of $\frac{1}{2}$ of identifying the brands correctly simply by guessing, so that about 10 students would correctly identify the brands with no discriminatory ability at all. That 12 students did does not seem inconsistent with guess work, and the *p*-value helps to quantify this. The probability of observing 12 or more correct results if one correct result has probability $\frac{1}{2}$ and the guesses are independent can be computed by the binomial formula as $\left\{\binom{20}{12}+\binom{20}{13}+\cdots+\binom{20}{20}\right\}\left(\frac{1}{2}\right)^{20}=0.34$: there is no evidence from these data that the number of correct answers could not have been obtained by guessing: in more statistical language assuming a binomial model, the observed data is consistent with probability of success $\frac{1}{2}$.

The student who claimed to have perfect discrimination, but actually guessed incorrectly, argued that her abilities should not be dismissed on the basis of one mistake, so the class carried out some computations to see what the *p*-value for the same observed data would be if the number of pairs of cups was increased. The probability of one or zero mistakes in a set of *n* trials for various values of *n*, is given in Table 1. From this we see that, for example, one or no mistakes in five trials is consistent with guess work but the same result in 10 trials is much less so.

In both parts of this example we assumed a model of independent trials, each of which could result in a success or failure, with constant probability of success. Our calculations also assumed this constant probability of success was 0.5. This latter restriction on the model is often called a 'null hypothesis' and the test of significance is a test of this null hypothesis; the *p*-value measures the consistency of the data with this null hypothesis. In many applications the null hypothesis plays the role of a conservative position that the experimenter hopes to disprove, and one reason for requiring rather small *p*-values

Table 1 Probability of zero or one mistakes in *n* independent Bernoulli trials with probability of a mistake = 0.5

n	*Probability*
5	0.1875
6	0.1094
7	0.0625
8	0.0352
9	0.0195
10	0.0107
11	0.0059
12	0.0032
13	0.0017
14	0.0009
15	0.0002

International Encyclopedia of the Social & Behavioral Sciences, Second Edition, 2015, 957–962

before declaring statistical significance is to raise the standard of proof required to replace a relatively simple working hypothesis by one that is possibly more complex and less well understood.

As formulated here the hypothesis being tested is that the probability of a correct choice is 0.5, and not the other aspects of the model, such as independence of the trials, and unchanging probability of success. The number of observed successes does not measure such model features, it provides information only on the probability of success. Functions of the data that do measure such model features can be constructed, and from these significance tests that assess the fitness of an assumed model; these play an important role in statistical inference as well.

Model-Based Inference

Models and Null Hypothesis

We assume that we have a statistical model for a random variable Y taking values in a sample space $\mathcal{Y}$, described by a parametric family of densities $\{f(y;\theta);\theta\in\Theta\}$. Tests of significance can in fact be constructed in more general settings but this framework is useful for defining the main ideas. If Y is the total number of successes in n independent Bernoulli trials with constant probability of success, then

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \qquad [1]$$

$\Theta = [0,1]$, and $\mathcal{Y} = \{0,1,\ldots,n\}$. If Y is a continuous random variable following a normal or bell curve distribution with mean θ_1 and variance θ_2^2, then

$$f(y;\theta) = \frac{1}{\sqrt{2\pi}\theta_2}\exp\left\{-\frac{1}{2\theta_2^2}(y-\theta_1)^2\right\} \qquad [2]$$

$\Theta = \mathbb{R}\times\mathbb{R}^+$ (R - real line; R - positive real line), and $\mathcal{Y} = \mathbb{R}$. The model for n independent observations from this distribution is

$$f(y_1,\ldots,y_n;\theta) = \frac{1}{\left(\sqrt{2\pi}\right)^n\theta_2^n}\exp\left\{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(y_i-\theta_1)^2\right\} \qquad [3]$$

$\Theta = \mathbb{R}\times\mathbb{R}^+$, and $\mathcal{Y} = \mathbb{R}^n$. For further discussion of statistical models, *see* Statistical Sufficiency; Distributions, Statistical: Special and Discrete; Distributions, Statistical: Approximations; Statistical: Special and Continuous.

As noted above, we assume the model is given, and our interest is in inference about the parameter θ. While this could take various forms, a test of significance starts with a so-called *null hypothesis* about θ, of the form

$$H_0 : \theta = \theta_0 \qquad [4]$$

or

$$H_0 : \theta\in\Theta_0 \qquad [5]$$

In [4] the parameter θ is fully specified, and H_0 is called a *point* null hypothesis or a *simple* null hypothesis. If θ is not fully specified, as in [5], H_0 is called a *composite* null hypothesis. In the taste-testing examples the simple null hypothesis was $\theta = 0.5$. In the normal model, [2], a hypothesis about the mean, such as $H_0 : \theta_1 = 0$, is composite, as the variance is left unspecified. Another composite null hypothesis is $H_0 : \theta_2 = \theta_1$, which restricts the full parameter space to a one-dimensional curve in $\mathbb{R}\times\mathbb{R}^+$.

A test is constructed by choosing a *test statistic* which is a function of the data that in some natural way measures departure from what is expected under the null hypothesis, and which has been standardized so that its distribution is known either exactly or to a good approximation under the null hypothesis. Test statistics are usually constructed so that large values indicate a discrepancy from the hypothesis.

Example 2. In the binomial model [1], the distribution of Y is completely specified by the null hypothesis $\theta = 0.5$ as

$$f(y) = \binom{n}{y}2^{-n}$$

and consistency of a given observed value y_0 of y, is measured by the p-value $\sum_{y=y_0}^{n}\binom{n}{y}2^{-n}$, the probability of observing a value as or more extreme than y_0. If y_0 is quite a bit smaller than expected, then it would be more usual to compute the p-value as $\sum_{y=0}^{y_0}\binom{n}{y}2^{-n}$. Each of these calculations was carried out in the discussion of taste testing in the Introduction.

Example 3. In independent sampling from the normal distribution, given at [3], we usually test the composite null hypothesis $H_0 : \theta_1 = \theta_{10}$, by constructing the *t-statistic*

$$T = \sqrt{n}\left(\overline{Y}-\theta_{10}\right)/S$$

where $\overline{Y} = n^{-1}\sum_{i=1}^{n}Y_i$ and $S^2 = (n-1)^{-1}\sum(Y_i-\overline{Y})^2$. Under H_0, T follows a t-distribution on $n-1$ degrees of freedom, and if large values of T are considered evidence against H_0, the p-value is

$$\Pr\left\{T\geq\sqrt{n}\left(\overline{y}-\theta_{10}\right)/s\right\}$$

where $\overline{y}$ and s are the values observed in the sample. (If small values of T are considered evidence against H_0, we would use the other tail of the distribution; see Section Two-Sided Testing for comments on two-sided tests of significance.) This probability needs to be computed numerically from an expression for the cumulative distribution function of the t-distribution. Historically tables of this distribution were provided for ready reference, typically by identifying a few critical values, such as $t_{0.10}$, $t_{0.05}$, and $t_{0.01}$ satisfying $\Pr\{T_\nu\geq t_\alpha\} = \alpha$, where T_ν is a random variable following a t-distribution on ν degrees of freedom. It was arguably the publication of these tables that led to a focus on the use of particular fixed levels for testing in applied work.

Example 4. Assume the model specifies that $Y_1,\ldots,Y_n$ are independent, identically distributed from a distribution with density $f(\cdot)$ on $\mathbb{R}$ and that we are interested in testing whether or not $f(\cdot)$ is a normal density:

$$H_0 : f(y) = \left(\sqrt{2\pi}\right)^{-1}e^{-\frac{1}{2}y^2} \qquad [6]$$

or

$$H_0 : f(y) = \left(\sqrt{2\pi}\right)^{-1}\theta_2^{-1}\exp\left\{-(y-\theta_1)^2/\left(2\theta_2^2\right)\right\} \qquad [7]$$

the former is a simple and the latter is a composite null hypothesis. For this problem it is less obvious how to construct

a test statistic or how to choose among alternative test statistics. Under [6] we know the distribution of each observation (standard normal) and thus of any function of the observations. The ordered values of Y could be compared to their expected values under [6], for example by plotting one against the other, and deviation of this plot from a line with intercept 0 and slope 1 could be measured in various ways. In the case of the composite null hypothesis [7], we could make use of the result that under H_0, $(Y_i - \overline{Y})/S$ has a distribution free of θ_1 and θ_2, and the vector of these residuals is independent of the pair $(\overline{Y}, S^2)$, and then for example compare the sample skewness $n^{-1}\sum_{i=1}^{n}\{(Y_i - \overline{Y})/S\}^3$ with that expected under normality.

Example 5. Suppose we have a sample of independent observations $Y_1,\ldots,Y_n$ on a circle of radius 1 and our null hypothesis is that the observations are uniformly distributed on the circle. One choice of a test statistic is $T_1 = \sum_{i=1}^{n}\cos(Y_i)$, very large positive (or negative) values indicating a concentration of observations at angle 0 (or π). If we instead wish to detect clumps of observations at two angles differing by π then $T_2 = \sum_{i=1}^{n}\{\cos(2Y_i) - 1\}$ would be more appropriate. The exact distribution of T_1 under H_0 is not available in closed form, but the mean and variance are readily computed as 0 and n, so a normal approximation might be used to compute the p-value.

In the examples described above the test statistics are ad hoc choices likely to be large if the null hypothesis is not true; these are called pure tests of significance, and are treated in detail in Cox and Hinkley (1974: Chapter 3); Examples 4 and 5 above are drawn from that chapter. A more sensitive test can be constructed if we have more specific knowledge of the likely form of departures from the null hypothesis. The theory of hypothesis testing formalizes this by setting up a null hypothesis and alternative hypothesis, and seeking to construct an optimal test for discriminating between them (*see* Hypothesis Testing, in Statistics). In the remainder of this section we consider an approach based on the likelihood function.

Significance Tests Based on Likelihood

In parametric models tests of significance are often constructed by using the likelihood function, and the p-value is computed by using an established approximation to the distribution of the test statistic. The likelihood function is proportional to the joint density of the data:

$$L(\theta; y) = c(y)f(y; \theta) \quad [8]$$

see Likelihood, Methods of Statistical Inference.

We first suppose that we are testing the simple null hypothesis $H_0 : \theta = \theta_0$ in the parametric model $f(y; \theta)$. Three test statistics often constructed from the likelihood function are the Wald or maximum likelihood statistic:

$$w_e = \left(\widehat{\theta} - \theta_0\right)^T j\left(\widehat{\theta}\right)\left(\widehat{\theta} - \theta_0\right) \quad [9]$$

the Rao or score statistic:

$$w_u = U(\theta_0)^T\left\{j\left(\widehat{\theta}\right)\right\}^{-1}U(\theta_0) \quad [10]$$

and the likelihood ratio statistic:

$$w = 2\left\{\ell\left(\widehat{\theta}\right) - \ell(\theta_0)\right\} \quad [11]$$

where in [8], [9], and [10] the following notation is used:

$$\sup_{\theta} L(\theta; y) = L\left(\widehat{\theta}; y\right) \quad [12]$$

$$\begin{aligned}\ell(\theta) &= \log L(\theta)\\ U(\theta) &= \ell'(\theta)\end{aligned} \quad [13]$$

$$j(\theta) = -\ell''(\theta) \quad [14]$$

The distributions of each of the statistics [8], [9], and [10] can be approximated by a χ^2_k distribution, under the model $f(y; \theta_0)$, with k the dimension of θ in the model. This relies on being able to apply a central limit theorem to $U(\theta)$, and to identifying the maximum likelihood estimator $\widehat{\theta}$ with the root of the equation $U(\theta) = 0$. The precise regularity conditions needed are somewhat elaborate; see for example Lehmann and Casella (1998: Chapter 6) and references therein. The important point is that under the simple null hypothesis the approximate distributions of each of these test statistics is known, and p-values readily computed.

In the case that the hypothesis is composite, a similar triple of test statistics computed from the likelihood function is available, but the notation needed to define them is more elaborate. The details can be found for example in Cox and Hinkley (1974: Chapter 9.3) and the notation above follows theirs.

If θ is a one-dimensional parameter, then a one-sided version of the test statistics given at [8], [9], and [10] can be used instead, as the signed square root of w_e, w_u, or w follows approximately a standard normal distribution.

It is rare that the exact distribution of test statistics can be computed, but the normal or chi-squared approximation can often be improved. These improvements are discussed in Barndorff-Nielsen and Cox (1994), Pace and Salvan (1997), Severini (2000) and Brazzale et al. (2007). One conclusion of this work is that among the three test statistics above, the signed square root of the likelihood ratio statistic w is generally preferred on a number of grounds, including the accuracy of the normal approximation to its exact distribution. This is true for both simple and composite tests of a scalar parameter.

Significance Functions and Posterior Probabilities

We can also use a test of significance to consider the whole set or interval of values of θ that are consistent with the data. If θ is scalar one of the simplest ways to do this is to compute $r(\theta) = \pm\sqrt{w(\theta)}$ as a function of θ, and tabulate or plot $\Phi\{r(\theta)\}$ against θ, choosing the negative root for $\widehat{\theta} < \theta$, and the positive square root otherwise. This significance function will in regular models decrease from one to zero as θ ranges over an interval of values. The θ values for which $\Phi(r)$ is 0.975 and 0.025 provide the endpoints of an approximate 95% confidence interval for θ. This approach is emphasized in Fraser (1991).

In a Bayesian approach to inference it is possible to make probability statements about the parameter or parameters in the model by constructing a posterior probability distribution for them. In a model with a scalar parameter θ based on a prior $\pi(\theta)$ and model $f(y; \theta)$ we compute a posterior density for θ as

$$\pi(\theta|y) \propto f(y; \theta)\pi(\theta) \quad [15]$$

and can assess any particular value θ_0 by computing

$$\int_{\theta_0}^{\infty} \pi(\theta|y)d\theta$$

the posterior probability of θ being larger than θ_0. This posterior probability is different from a p-value: a p-value assesses the data in light of a fixed value of θ, and the posterior probability assesses a fixed value of θ in light of the probability distribution ascribed to the parameter. Many people find a posterior probability easier to understand, and indeed often interpret the p-value in this way. There is a literature on choosing priors called 'matching priors' to reconcile these two approaches to inference; see Kass and Wasserman (1996), and Datta and Mukerjee (2004); see also Bayesian Statistics. Unfortunately it is not possible to find so-called matching priors that are simultaneously matching for all components of a vector parameter.

Hypothesis Testing

Little has been said here about the choice of a test statistic for carrying out a test of significance. The difficulty is that the theory of significance testing provides no guidance on this choice. The likelihood-based test statistics described above have proved to be reasonably effective in parametric models, but in a more complicated problem, such as testing the goodness of fit of a hypothesized model, this approach is often not available. To make further progress in the choice of test statistics, the classical approach is to formulate a notion of a 'powerful' test statistic, i.e., one that will reliably give small p-values when the null hypothesis is not correct. To do this in a systematic way requires specifying what model might hold if in fact the null hypothesis is incorrect. In parametric models where the null hypothesis is $H_0 : \theta = \theta_0$ the alternative may well be $H_a : \theta \neq \theta_0$. In more general settings the null hypothesis might be H_0 : 'the response variable follows a normal distribution' and the alternative H_a : 'the response variable does not follow a normal distribution.' Even in the parametric setting if θ is a vector parameter it may be necessary to consider what direction in the parameter space away from θ_0 is of interest. The formalization of these ideas is the theory of hypothesis testing, which considers both null and alternative hypotheses and optimal choices of test statistics (*see* Hypothesis Testing, in Statistics; Goodness of Fit: Overview; Methods and Models).

Further Topics

Fixed Level Testing

The problem of focusing on one or two so-called critical p-values is sometimes referred to as fixed-level testing. This was useful when computation of p-values was a very lengthy exercise, and it was usual to provide tables of critical values. It is now usually a very routine matter to compute the exact p-value, which is usually (and should be) reported along with other details such as sample size, estimated effect size, and details of the study design. There is still in some quarters a reliance on fixed level test, with the result that studies for which the p-value is judged 'not statistically significant' may not be published. This is sometimes called the 'file drawer problem,' and a quantitative analysis was considered in Dawid and Dickey (1977).

In several fields it is now standard to make the results of all studies on a given topic, including inconclusive studies, available online. In healthcare the Cochrane collaboration is perhaps the best-known of these. This issue is particularly important for *meta-analysis*, see Section Combining Tests of Significance.

Achievable p-values

In some problems where the distribution of the test statistic under the null hypothesis is concentrated on a discrete set, the number of available p-values will be relatively small. This happens with categorical data, especially if the sample size is small. Some authors have argued that for such highly discrete situations a better assessment of the null hypothesis can be achieved by the use of Barnard's mid p-value, which replaces $\Pr\{t(Y) \geq t^0\}$ with $(1/2)\Pr\{t(Y) = t^0\} + \Pr\{t(y) > t^0\}$, where $t(Y)$ is the statistic on which the significance test is based, and t^0 is the observed value in the data; see Agresti (1992) and references therein.

Combining Tests of Significance

The p-value is a function of the data, taking small values when the data are incompatible with the null hypothesis, and vice versa. As a function of y the p-value itself has a distribution under the model $f(y; \theta)$ and in particular under the null hypothesis H_0 has the uniform distribution on the interval $(0,1)$. In principle then if we have computed p-values from a number of different datasets the p-values can be compared to observations from a $U(0,1)$ distribution with the objective of obtaining evidence of failure of the null hypothesis across the collection of datasets. This is one of the ideas behind *meta-analysis*; *see* Meta-analysis: Overview; Meta Analysis: Tools. One difficulty is that the studies will nearly always differ in a number of respects that may mean they are not all measuring the same parameter, or measuring it in the same way. Another difficulty is that studies for which the p-value is not 'statistically significant' will not have been published, and thus are unavailable to be included in a *meta-analysis*. This selection effect may seriously bias the results of the *metametaanalysis*. In some fields this selection effect has been lessened by the practice of registering all trials in a given subject area on the internet.

Two-Sided Testing

A point of confusion in the evaluation of p-values for testing scalar parameters is the distinction sometimes made between one-sided and two-sided tests of significance. A reliable procedure is to compute the p-value as the twice the smaller of the probabilities that the test statistic is larger than or smaller than the observed value, under the null hypothesis. This so-called two-sided p-value measures disagreement with the null hypothesis in two directions away from the null hypothesis, toward the alternative that the, say, new treatment is worse than the old treatment as well as better. In the absence of very

concrete a priori evidence that the alternative hypothesis is genuinely one-sided this p-value is preferable.

In testing hypotheses about parameters of dimension larger than one, it can be difficult to decide on the relevant direction away from the null hypothesis. One solution is to identify the parameters of interest individually, and carry out separate tests on each of these parameters in turn. This will usually be effective if there are a relatively small number of parameters of interest. In applications involving computation of a very large number of p-values, new techniques are needed; these are briefly discussed in the next section.

Difficulties with Significance Tests

Sample Size

In Figure 1 we show the p-value for a one-sided t-test with an observed value of the t-statistic equal to 2.0, as a function of the sample size. In this example, and quite generally, the p-value is a decreasing function of the size of the sample, so that a very large study is more likely to show 'statistical significance' than a smaller study. This has led to considerable criticism of the p-value as a summary measure. The p-value is also sometimes misinterpreted, especially when it is small, as the probability that the null hypothesis is false. Some statisticians have argued that for this reason posterior probabilities are a better measure of disagreement with the null hypothesis; see for example Berger and Sellke (1987) and Schervish (1996).

To some extent the criticism can be countered by noting that the p-value is just one summary measure of a set of data, and excessive reliance on one measure is inappropriate. In a parametric setting it is nearly always advisable to provide, along with the p-value for testing a particular value of the parameter of interest, an estimate of the effect size, or some relevant parameter of the model, along with an indication of the precision of this estimate. This can be accomplished by reporting a significance function, if the parameter of interest is one-dimensional. At a more practical level, it should always be noted that a small p-value should be interpreted in the context of other aspects of the study. For example a p-value of less than 0.05 could be based on a very small difference in a study of 10 000 cases or a relatively large difference in a study of 1000 cases. A 1% reduction in an average response may be of substantial importance for one scientific context, and meaningless for others; this needs to be evaluated in that context, and not by relying on the fact that it is 'statistically significant.' Unfortunately, the notion that a study report is complete if and only if the p-value is found to be less than 0.05 is fairly widely ingrained in some disciplines, and indeed forms a part of the requirements of some government agencies for approving new treatments.

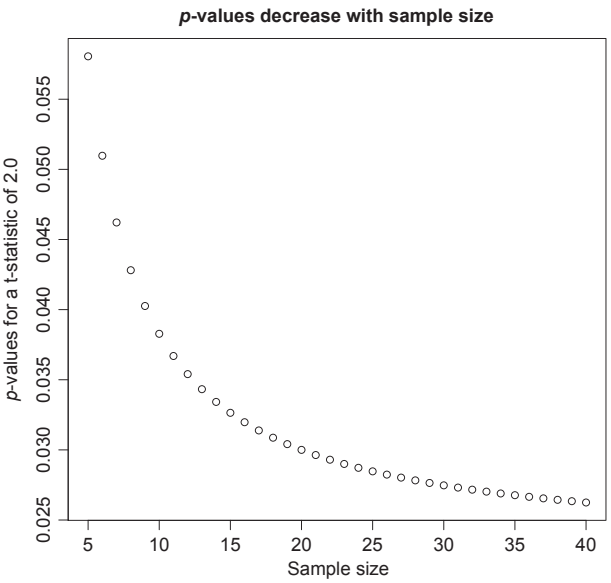

Figure 1 The p-value associated with an observed t-statistic equal to 2.0, as a function of the sample size.

Multiple Testing

If a number of significance tests are carried out on the same set of data, but the significance level, or p-value, that is reported is the smallest of these, then a different analysis is needed. This smallest p-value will not have the interpretation of a p-value from a single test; for example if we regard p-values less that 0.05 as 'significant,' then we would expect to find 5 spuriously significant results in 100 tests, on average.

In some applications there might be two or three thousand tests carried out, all of a similar type. One example arises in image analysis, for example comparing the blood flow in each of several thousand pixels or voxels of a brain scan, under two (or more) conditions. Large numbers of t-tests are often conducted in genomic analysis of expression arrays, with again the goal of comparing two conditions. The p-value for a single such comparison is no longer a reliable measure of the consistency of the data with the null hypothesis; several p-values will be small even when the null hypothesis is true, simply by chance.

There is a large literature on assessments of hypotheses under multiple testing, that has become particularly prominent in biological applications, but also has applications to many other sciences. *See* False Discovery Rate; Multiple Comparisons A good statistical theory reference is Efron (2010), and there are a number of more specialized books, for particular scientific fields, such as Dudoit and van der Laan (2008). In high-energy physics the search for new particles involves looking through a great many energy bands, sometimes called in that context the 'look elsewhere' effect, and it has become conventional to require a so-called '5-sigma' result to claim discovery of a new particle. This relates the observed level of significance to the probability that a normal random variable exceeds its mean by five standard deviations; this probability is 3×10^{-7}.

In many areas of research the delineation of the number of significance tests that have been carried out, but are not reported, is less clear. For example, in a new epidemiological study of the health effects of some environmental agent, several different models may be fitted to the data, including perhaps various transformations of the exposure measurements, different levels of control for confounding, and so on. Again the evidence provided by any single reported significance test needs to be considered in the light of the many other

tests that may have been formally or informally applied to the same data.

The reliance on multiple significance tests forms a part of the increasing concern, in the academic and in the popular press, about the replicability of published scientific research; see for example, Ioannidis (2005), or Economist (2013). Research in this area is ongoing; for a recent discussion in the medical context see Jager and Leek (2013) and the accompanying discussion.

Conclusion

A test of statistical significance is a mathematical calculation based on a test statistic, a null hypothesis, and the distribution of the test statistic under the null hypothesis. The result of the test is to indicate whether the data are consistent with the null hypothesis: if they are not, then either we have observed an event of low probability, or the null hypothesis is not correct.

The choice of test statistic is in principle arbitrary, but in practice might be determined by convention in the field of application, by intuition in a relatively new setting, or by one or more considerations developed in statistical theory. It is convenient to use test statistics whose distributions can be easily calculated exactly or to a good approximation. It is useful to use a test statistic that is sensitive to the particular departures from the null hypothesis that are of particular interest in the application.

A test of statistical significance is just one component of the analysis of a set of data, and should be supplemented by estimates of effects of interest, considerations related to sample size, and a discussion of the validity of any assumptions of independence or underlying models that have been made in the analysis. A statistically significant result is not necessarily an important result in any particular analysis, but needs to be considered in the context of research in that field.

An eloquent introduction to tests of significance is given in Fisher (1935: Chapter II). Kalbfleisch (1979: Chapter 12) is a good text book reference at an undergraduate level. The discussion here draw considerably from Cox and Hinkley (1974: Chapter 3), which is a good reference at a more advanced level. An excellent overview is given in Cox (1977).

See also: Bayesian Statistics; Distributions, Statistical: Approximations; Distributions, Statistical: Special and Discrete; Hypothesis Testing in Statistics; Multiple Comparisons, Statistics of; Statistical Sufficiency.

Bibliography

Agresti, A., 1992. A survey of exact inference for contingency tables. Statistical Science 7, 131–153.
Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Chapman and Hall, London.
Berger, J.O., Sellke, T., 1987. Testing a point null hypothesis: the irreconcilability of p-values and evidence. Journal of the American Statistical Association 82, 112–122.
Brazzale, A.R., Davison, A.C., Reid, N., 2007. Applied Asymptotics: Case Studies in Small Sample Statistics. Cambridge University Press, Cambridge.
Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman and Hall, London.
Cox, D.R., 1977. The role of significance tests. Scandinavian Journal of Statistics 4, 49–70.
Datta, G., Mukerjee, R., 2004. Probability Matching Priors: Higher Order Asymptotics. Springer-Verlag, New York.
Dawid, A.P., Dickey, J.M., 1977. Likelihood and Bayesian inference from selectively reported data. Journal of the American Statistical Association 72, 845–850.
Dudoit, S., van der Laan, M.J., 2008. Multiple Testing Procedures with Applications to Genomics. Springer-Verlag, New York.
Economist, October 11, 2013. Unreliable Research: Trouble at the Lab. Print Edition.
Efron, B., 2010. Large Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. (Institute of Mathematical Statistics Monographs). Cambridge University Press, Cambridge.
Fisher, R.A., 1935. The Design of Experiments. Oliver & Boyd, Edinburgh.
Fraser, D.A.S., 1991. Statistical inference: likelihood to significance. Journal of the American Statistical Association 86, 258–265.
Ioannidis, J., 2005. Why most published research findings are false. PLOS Medicine 2, 696–701.
Jager, L.R., Leek, J., 2013. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics 15, 1–12.
Kalbfleisch, J.G., 1979. Probability and Statistical Inference, vol. 2. Springer-Verlag, New York.
Kass, R.E., Wasserman, L., 1996. Formal rules for selecting prior distributions: a review and annotated bibliography. Journal of the American Statistical Association 91, 1343–1370.
Lehmann, E.L., Casella, G., 1998. Theory of Point Estimation. Springer-Verlag, New York.
Pace, L., Salvan, A., 1997. Principles of Statistical Inference from a Neo-Fisherian Perspective. World Scientific, Singapore.
Schervish, M.J., 1996. P values: what they are and what they are not. American Statistician 96, 203–206.
Severini, T.A., 2000. Likelihood Methods in Statistics. Oxford University Press, Oxford.