P-value

In statistical significance testing, the ***p*-value** is the probability of obtaining a test statistic result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. A researcher will often "reject the null hypothesis" when the *p*-value turns out to be less than a predetermined significance level, often 0.05[1][2] or 0.01. Such a result indicates that the observed result would be highly unlikely under the null hypothesis. Many common statistical tests, such as chi-squared tests or Student's t-test, produce test statistics which can be interpreted using *p*-values.

In a statistical test, sample results are compared to possible population conditions by way of two competing hypotheses: the *null hypothesis* is a neutral or "uninteresting" statement about a population, such as "no change" in the value of a parameter from a previous known value or "no difference" between two groups; the other, the *alternative* (or *research*) *hypothesis* is the "interesting" statement that the person performing the test would like to conclude if the data will allow it. The *p*-value is the probability of obtaining the observed sample results (or a more extreme result) when the null hypothesis is actually true. If this *p*-value is very small, usually less than or equal to a threshold value previously chosen called the significance level (traditionally 5% or 1%), it suggests that the observed data is inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the other hypothesis accepted as true.

An informal interpretation of a *p*-value, based on a significance level of about 10%, might be:

- $p \leq 0.01$: very strong presumption against null hypothesis
- $0.01 < p \leq 0.05$: strong presumption against null hypothesis
- $0.05 < p \leq 0.1$: low presumption against null hypothesis
- $p > 0.1$: no presumption against the null hypothesis

A new Bayesian inference approach highlights that these threshold values are too optimistic and explain the lack of reproducibility of scientific studies, suggesting a $p < 0.001$ or 0.0053.[3] However, a follow-up article illustrates that these more stringent threshold values are not absolute, but rather arise from "the discrepancy between *p*-values and Bayes factors", and are not a complete solution to the problem of reproducibility.[4]

The *p*-value is a key concept in the approach of Ronald Fisher, where he uses it to measure the weight of the data against a specified hypothesis, and as a guideline to ignore data that does not reach a specified significance level. Fisher's approach does not involve any alternative hypothesis, which is instead a feature of the Neyman–Pearson approach. The *p*-value should not be confused with the significance level α in the Neyman–Pearson approach or the Type I error rate [false positive rate]. Fundamentally, the *p*-value does not in itself support reasoning about the probabilities of hypotheses, nor choosing between different hypotheses – it is simply a measure of how likely the data (or a more "extreme" version of it) were to have occurred, assuming the null hypothesis is true.[5]

Statistical hypothesis tests making use of *p*-values are commonly used in many fields of science and social sciences, such as economics, psychology, biology, criminal justice and criminology, and sociology.[6]

Depending on which style guide is applied, the "p" is styled either italic or not, capitalized or not, and hyphenated or not (*p*-value, *p* value, *P*-value, *P* value, p-value, p value, P-value, P value).

Basic concepts

The *p*-value is used in the context of null hypothesis testing in order to quantify the idea of statistical significance of evidence.[7] Null hypothesis testing is a reductio ad absurdum argument adapted to statistics. In essence, a claim is shown to be valid by demonstrating the improbability of the counter-claim that follows from its denial. As such, the only hypothesis which needs to be specified in this test, and which embodies the counter-claim, is referred to as the null hypothesis. A result is said to be statistically significant if it can enable the rejection of the null hypothesis. The rejection of the null hypothesis implies that the correct hypothesis lies in the logical complement of the null

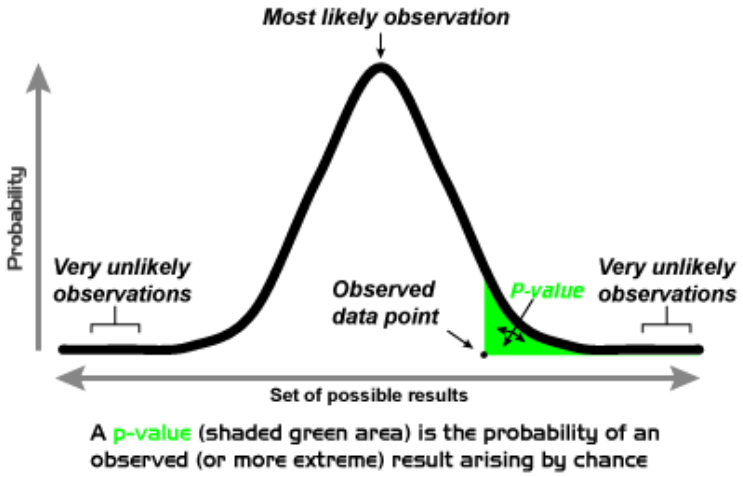
hypothesis. For instance, if the null hypothesis is assumed to be a standard normal distribution N(0,1), then the rejection of this null hypothesis can mean either (i) the mean is not zero, or (ii) the variance is not unity, or (iii) the distribution is not normal.

In statistics, a statistical hypothesis refers to a probability distribution that is assumed to govern the observed data.[8] If X is a random variable representing the observed data and H is the statistical hypothesis under consideration, then the notion of statistical significance can be naively quantified by the conditional probability $Pr(X|H)$, which gives the likelihood of the observation if the hypothesis is *assumed* to be correct. However, if X is a continuous random variable, and we observed an instance x, then $Pr(X = x|H) = 0$. Thus this naive definition is inadequate and needs to be changed so as to accommodate the continuous random variables. Nonetheless, it does help to clarify that *p*-values should not be confused with either $Pr(H|X)$, the probability of the hypothesis given the data, or $Pr(H)$, the probability of the hypothesis being true, or $Pr(X)$, the probability of observing the given data.

Definition and interpretation

The *p*-value is defined as the probability, under the assumption of hypothesis H, of obtaining a result equal to *or more extreme than what was actually observed*. Depending on how we look at it, the "more extreme than what was actually observed" can either mean $\{X \geq x\}$ (right tail event) or $\{X \leq x\}$ (left tail event) or the "smaller" of $\{X \leq x\}$ and $\{X \geq x\}$ (double tailed event). Thus the *p*-value is given by

- $Pr(X \geq x|H)$ for right tail event,
- $Pr(X \leq x|H)$ for left tail event,
- $2\min(Pr(X \leq x|H), Pr(X \geq x|H))$ for double tail event.



Example of a *p*-value computation. The vertical coordinate is the probability density of each outcome, computed under the null hypothesis. The *p*-value is the area under the curve past the observed data point.

The smaller the *p*-value, the larger the significance because it tells the investigator that the hypothesis under consideration may not adequately explain the observation. The hypothesis H is rejected if any of these probabilities is less than or equal to a small, fixed, but arbitrarily pre-defined, threshold value α, which is referred to as the level of significance. Unlike the *p*-value, the α level is not derived from any observational data nor does it depend on the underlying hypothesis; the value of α is instead determined based on the consensus of the research community that the investigator is working in.

It should be noted that since the value of x that defines the left tail or right tail event is a random variable, this makes the *p*-value a function of x and a random variable in itself defined uniformly over $[0, 1]$ interval. Thus, the *p*-value is not fixed. This implies that *p*-value cannot be given a frequency counting interpretation, since the probability has to be fixed for the frequency counting interpretation to hold. In other words, if a same test is repeated independently bearing upon the same overall null hypothesis, then it will yield different *p*-values at every repetition. Nevertheless, these different *p*-values can be combined using Fisher's combined probability test. It should further be noted that an *instantiation* of this random *p*-value can still be given a frequency counting interpretation with respect to the number of observations taken during a given test, as per the definition, as the percentage of observations more

extreme than the one observed under the assumption that the null hypothesis is true. Lastly, the fixed pre-defined α level can be interpreted as the rate of falsely rejecting the null hypothesis (or type I error), since $Pr(\text{Reject } H|H) = Pr(p \leq \alpha$

Calculation

Usually, instead of the actual observations, X is instead a test statistic. A test statistic is a scalar function of all the observations, which summarizes the data by a single number. As such, the test statistic follows a distribution determined by the function used to define that test statistic and the distribution of the observational data. For the important case where the data are hypothesized to follow the normal distribution, depending on the nature of the test statistic, and thus our underlying hypothesis of the test statistic, different null hypothesis tests have been developed. Some such tests are z-test for normal distribution, t-test for Student's t-distribution, f-test for f-distribution. When the data do not follow a normal distribution, it can still be possible to approximate the distribution of these test statistics by a normal distribution by invoking the central limit theorem for large samples, as in the case of Pearson's chi-squared test.

Thus computing a *p*-value requires a null hypothesis, a test statistic (together with deciding whether the researcher is performing a one-tailed test or a two-tailed test), and data. Even though computing the test statistic on given data may be easy, computing the sampling distribution under the null hypothesis, and then computing its CDF is often a difficult computation. Today this computation is done using statistical software, often via numeric methods (rather than exact formulas), while in the early and mid 20th century, this was instead done via tables of values, and one interpolated or extrapolated *p*-values from these discrete values. Rather than using a table of *p*-values, Fisher instead inverted the CDF, publishing a list of values of the test statistic for given fixed *p*-values; this corresponds to computing the quantile function (inverse CDF).

Examples

Here a few simple examples follow, each illustrating a potential pitfall.

One roll of a pair of dice

Suppose a researcher rolls a pair of dice once and assumes a null hypothesis that the dice are fair. The test statistic is "the sum of the rolled numbers" and is one-tailed. The researcher rolls the dice and observes that both dice show 6, yielding a test statistic of 12. The *p*-value of this outcome is 1/36, or about 0.028 (the highest test statistic out of 6×6 = 36 possible outcomes). If the researcher assumed a significance level of 0.05, he or she would deem this result significant and would reject the hypothesis that the dice are fair.

In this case, a single roll provides a very weak basis (that is, insufficient data) to draw a meaningful conclusion about the dice. This illustrates the danger with blindly applying *p*-value without considering the experiment design.

Five heads in a row

Suppose a researcher flips a coin five times in a row and assumes a null hypothesis that the coin is fair. The test statistic of "total number of heads" can be one-tailed or two-tailed: a one-tailed test corresponds to seeing if the coin is biased towards heads, while a two-tailed test corresponds to seeing if the coin is biased either way. The researcher flips the coin five times and observes heads each time (HHHHH), yielding a test statistic of 5. In a one-tailed test, this is the most extreme value out of all possible outcomes, and yields a *p*-value of $(1/2)^5 = 1/32 \approx 0.03$. If the researcher assumed a significance level of 0.05, he or she would deem this result to be significant and would reject the hypothesis that the coin is fair. In a two-tailed test, a test statistic of zero heads (TTTTT) is just as extreme, and thus the data of HHHHH would yield a *p*-value of $2\times(1/2)^5 = 1/16 \approx 0.06$, which is not significant at the 0.05 level.

This demonstrates that specifying a direction (on a symmetric test statistic) halves the *p*-value (increases the significance) and can mean the difference between data being considered significant or not.

Sample size dependence

Suppose a researcher flips a coin some arbitrary number of times (n) and assumes a null hypothesis that the coin is fair. The test statistic is the total number of heads. Suppose the researcher observes heads for each flip, yielding a test statistic of n and a p-value of $2/2^n$. If the coin was flipped only 5 times, the p-value would be $2/32 = 0.0625$, which is not significant at the 0.05 level. But if the coin was flipped 10 times, the p-value would be $2/1024 \approx 0.002$, which is significant at the 0.05 level.

In both cases the data suggest that the null hypothesis is false (that is, the coin is not fair somehow), but changing the sample size changes the p-value and significance level. In the first case the sample size is not large enough to allow the null hypothesis to be rejected at the 0.05 level (in fact, the p-value never be below 0.05).

This demonstrates that in interpreting p-values, one must also know the sample size, which complicates the analysis.

Alternating coin flips

Suppose a researcher flips a coin ten times and assumes a null hypothesis that the coin is fair. The test statistic is the total number of heads and is two-tailed. Suppose the researcher observes alternating heads and tails with every flip (HTHTHTHTHT). This yields a test statistic of 5 and a p-value of 1 (completely unexceptional), as this is the expected number of heads.

Suppose instead that test statistic for this experiment was the "number of alternations" (that is, the number of times when H followed T or T followed H), which is again two-tailed. This would yield a test statistic of 9, which is extreme, and has a p-value of $1/2^8 = 1/256 \approx 0.0039$. This would be considered extremely significant—well beyond the 0.05 level. These data indicate that, in terms of one test statistic, the data set is extremely unlikely to have occurred by chance, though it does not suggest that the coin is biased towards heads or tails.
By the first test statistic, the data yield a high p-value, suggesting that the number of heads observed is not unlikely. By the second test statistic, the data yield a low p-value, suggesting that the pattern of flips observed is very, very unlikely. There is no "alternative hypothesis," so only rejection of the null hypothesis is possible) and such data could have many causes – the data may instead be forged, or the coin flipped by a magician who intentionally alternated outcomes.

This example demonstrates that the p-value depends completely on the test statistic used, and illustrates that p-values can only help researchers to reject a null hypothesis, not consider other hypotheses.

Impossible outcome and very unlikely outcome

Suppose a researcher flips a coin two times and assumes a null hypothesis that the coin is unfair: it has two heads and no tails. The test statistic is the total number of heads (one-tailed). The researcher observes one head and one tail (HT), yielding a test statistic of 1 and a p-value of 0. In this case the data is inconsistent with the hypothesis–for a two-headed coin, a tail can never come up. In this case the outcome is not simply unlikely in the null hypothesis, but in fact impossible, and the null hypothesis can be definitely rejected as false. In practice such experiments almost never occur, as all data that could be observed would be possible in the null hypothesis (albeit unlikely).

If the null hypothesis were instead that the coin came up heads 99% of the time (otherwise the same setup), the p-value would instead be[9] $0.0199 \approx 0.02.$ In this case the null hypothesis could not definitely be ruled out – this outcome is unlikely in the null hypothesis, but not impossible – but the null hypothesis would be rejected at the 0.05 level, and in fact at the 0.02 level, since the outcome is less than 2% likely in the null hypothesis.

Coin flipping

Main article: Checking whether a coin is fair

As an example of a statistical test, an experiment is performed to determine whether a coin flip is fair (equal chance of landing heads or tails) or unfairly biased (one outcome being more likely than the other).

Suppose that the experimental results show the coin turning up heads 14 times out of 20 total flips. The null hypothesis is that the coin is fair, and the test statistic is the number of heads. If we consider a right-tailed test, the *p*-value of this result is the chance of a fair coin landing on heads *at least* 14 times out of 20 flips. This probability can be computed from binomial coefficients as

$$\begin{aligned}&\text{Prob}(14 \text{ heads}) + \text{Prob}(15 \text{ heads}) + \cdots + \text{Prob}(20 \text{ heads}) \\ &= \frac{1}{2^{20}}\left[\binom{20}{14} + \binom{20}{15} + \cdots + \binom{20}{20}\right] = \frac{60{,}460}{1{,}048{,}576} \approx 0.058\end{aligned}$$

This probability is the *p*-value, considering only extreme results which favor heads. This is called a one-tailed test. However, the deviation can be in either direction, favoring either heads or tails. We may instead calculate the two-tailed *p*-value, which considers deviations favoring either heads or tails. As the binomial distribution is symmetrical for a fair coin, the two-sided *p*-value is simply twice the above calculated single-sided *p*-value; *i.e.*, the two-sided *p*-value is 0.115.

In the above example, we thus have:

- Null hypothesis (H_0): The coin is fair, i.e. Prob(heads) = 0.5
- Test statistic: Number of heads
- Level of significance: 0.05
- Observation O: 14 heads out of 20 flips; and
- Two-tailed *p*-value of observation O given H_0 = 2*min(Prob(no. of heads ≥ 14 heads), Prob(no. of heads ≤ 14 heads))= 2*min(0.058, 0.978) = 2*0.058 = 0.115.

Note that the Prob(no. of heads ≤ 14 heads) = 1 - Prob(no. of heads ≥ 14 heads) + Prob(no. of head = 14) = 1 - 0.058 + 0.036 = 0.978; however symmetry of the binomial distribution makes this an unnecessary computation to find the smaller of the two probabilities.

Here the calculated *p*-value exceeds 0.05, so the observation is consistent with the null hypothesis, as it falls within the range of what would happen 95% of the time were the coin in fact fair. Hence, we fail to reject the null hypothesis at the 5% level. Although the coin did not fall evenly, the deviation from expected outcome is small enough to be consistent with chance.

However, had one more head been obtained, the resulting *p*-value (two-tailed) would have been 0.0414 (4.14%). This time the null hypothesis – that the observed result of 15 heads out of 20 flips can be ascribed to chance alone – is rejected when using a 5% cut-off.

History

While the modern use of *p*-values was popularized by Fisher in the 1920s, computations of *p*-values date back to the 1770s, where they were calculated by Pierre-Simon Laplace:[10]

> In the 1770s Laplace considered the statistics of almost half a million births. The statistics showed an excess of boys compared to girls. He concluded by calculation of a *p*-value that the excess was a real, but unexplained, effect.

The *p*-value was first formally introduced by Karl Pearson in his Pearson's chi-squared test,[11] using the chi-squared distribution and notated as capital P.[11] The *p*-values for the chi-squared distribution (for various values of χ^2 and degrees of freedom), now notated as *P,* was calculated in (Elderton 1902), collected in (Pearson 1914, pp. xxxi–xxxiii, 26–28, Table XII). The use of the *p*-value in statistics was popularized by Ronald Fisher,[12] and it

plays a central role in Fisher's approach to statistics.[13]

In the influential book *Statistical Methods for Research Workers* (1925), Fisher proposes the level $p = 0.05$, or a 1 in 20 chance of being exceeded by chance, as a limit for statistical significance, and applies this to a normal distribution (as a two-tailed test), thus yielding the rule of two standard deviations (on a normal distribution) for statistical significance – see 68–95–99.7 rule.[14][15][2]

He then computes a table of values, similar to Elderton, but, importantly, reverses the roles of χ^2 and p. That is, rather than computing p for different values of χ^2 (and degrees of freedom n), he computes values of χ^2 that yield specified p-values, specifically 0.99, 0.98, 0.95, 0,90, 0.80, 0.70, 0.50, 0.30, 0.20, 0.10, 0.05, 0.02, and 0.01.[16] This allowed computed values of χ^2 to be compared against cutoffs, and encouraged the use of p-values (especially 0.05, 0.02, and 0.01) as cutoffs, instead of computing and reporting p-values themselves. The same type of tables were then compiled in (Fisher & Yates 1938), which cemented the approach.[2]

As an illustration of the application of p-values to the design and interpretation of experiments, in his following book *The Design of Experiments* (1935), Fisher presented the lady tasting tea experiment,[17] which is the archetypal example of the p-value.

To evaluate a lady's claim that she (Muriel Bristol) could distinguish by taste how tea is prepared (first adding the milk to the cup, then the tea, or first tea, then milk), she was sequentially presented with 8 cups: 4 prepared one way, 4 prepared the other, and asked to determine the preparation of each cup (knowing that there were 4 of each). In this case the null hypothesis was that she had no special ability, the test was Fisher's exact test, and the p-value was $1/\binom{8}{4} = 1/70 \approx 0.014,$ so Fisher was willing to reject the null hypothesis (consider the outcome highly unlikely to be due to chance) if all were classified correctly. (In the actual experiment, Bristol correctly classified all 8 cups.)

Fisher reiterated the $p = 0.05$ threshold and explained its rationale, stating:[18]

> It is usual and convenient for experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results.

He also applies this threshold to the design of experiments, noting that had only 6 cups been presented (3 of each), a perfect classification would have only yielded a p-value of $1/\binom{6}{3} = 1/20 = 0.05,$ which would not have met this level of significance.[18] Fisher also underlined the frequentist interpretation of p, as the long-run proportion of values at least as extreme as the data, assuming the null hypothesis is true.

In later editions, Fisher explicitly contrasted the use of the p-value for statistical inference in science with the Neyman–Pearson method, which he terms "Acceptance Procedures".[19] Fisher emphasizes that while fixed levels such as 5%, 2%, and 1% are convenient, the exact p-value can be used, and the strength of evidence can and will be revised with further experimentation. In contrast, decision procedures require a clear-cut decision, yielding an irreversible action, and the procedure is based on costs of error, which he argues are inapplicable to scientific research.

Misunderstandings

Despite the ubiquity of *p*-value tests, this particular test for statistical significance has been criticized for its inherent shortcomings and the potential for misinterpretation.

The data obtained by comparing the *p*-value to a significance level will yield one of two results: either the null hypothesis is rejected, or the null hypothesis *cannot* be rejected at that significance level (which however does not imply that the null hypothesis is *true*). In Fisher's formulation, there is a disjunction: a low *p*-value means *either* that the null hypothesis is true and a highly improbable event has occurred, *or* that the null hypothesis is false.

However, people interpret the *p*-value in many incorrect ways, and try to draw other conclusions from *p*-values, which do not follow.

The *p*-value does not in itself allow reasoning about the probabilities of hypotheses; this requires multiple hypotheses or a range of hypotheses, with a prior distribution of likelihoods between them, as in Bayesian statistics, in which case one uses a likelihood function for all possible values of the prior, instead of the *p*-value for a single null hypothesis.

The *p*-value refers only to a single hypothesis, called the null hypothesis, and does not make reference to or allow conclusions about any other hypotheses, such as the alternative hypothesis in Neyman–Pearson statistical hypothesis testing. In that approach one instead has a decision function between two alternatives, often based on a test statistic, and one computes the rate of Type I and type II errors as α and β. However, the *p*-value of a test statistic cannot be directly compared to these error rates α and β – instead it is fed into a decision function.

There are several common misunderstandings about *p*-values.

1. **The *p*-value is *not* the probability that the null hypothesis is true, nor is it the probability that the alternative hypothesis is false – it is not connected to either of these.** In fact, frequentist statistics does not, and cannot, attach probabilities to hypotheses. Comparison of Bayesian and classical approaches shows that a *p*-value can be very close to zero while the posterior probability of the null is very close to unity (if there is no alternative hypothesis with a large enough *a priori* probability and which would explain the results more easily). This is Lindley's paradox. But there are also *a priori* probability distributions where the posterior probability and the *p*-value have similar or equal values.
2. **The *p*-value is *not* the probability that a finding is "merely a fluke."** Calculating the *p*-value is based on the assumption that *every* finding is a fluke, that is, the product of chance alone. Thus, the probability that the result is due to chance is in fact unity. The phrase "the results are due to chance" is used to mean that the null hypothesis is probably correct. However, that is merely a restatement of the inverse probability fallacy, since the *p*-value cannot be used to figure out the probability of a hypothesis being true.
3. **The *p*-value is *not* the probability of falsely rejecting the null hypothesis.** This error is a version of the so-called prosecutor's fallacy.
4. **The *p*-value is *not* the probability that replicating the experiment would yield the same conclusion.** Quantifying the replicability of an experiment was attempted through the concept of *p*-rep.
5. **The significance level, such as 0.05, is not determined by the *p*-value.** Rather, the significance level is decided by the person conducting the experiment (with the value 0.05 widely used by the scientific community) before the data are viewed, and is compared against the calculated *p*-value after the test has been performed. (However, reporting a *p*-value is more useful than simply saying that the results were or were not significant at a given level, and allows readers to decide for themselves whether to consider the results significant.)
6. **The *p*-value does not indicate the size or importance of the observed effect.** The two do vary together however–the larger the effect, the smaller sample size will be required to get a significant *p*-value (see effect size).

Criticisms

Main article: Statistical hypothesis testing § Criticism

Critics of *p*-values point out that the criterion used to decide "statistical significance" is based on an arbitrary choice of level (often set at 0.05). If significance testing is applied to hypotheses that are known to be false in advance, a non-significant result will simply reflect an insufficient sample size; a *p*-value depends only on the information obtained from a given experiment.

The *p*-value is incompatible with the likelihood principle, and *p*-value depends on the experiment design, or equivalently on the test statistic in question. That is, the definition of "more extreme" data depends on the sampling methodology adopted by the investigator; for example, the situation in which the investigator flips the coin 100 times yielding 50 heads has a set of extreme data that is different from the situation in which the investigator continues to flip the coin until 50 heads are achieved yielding 100 flips. This is to be expected, as the experiments are different experiments, and the sample spaces and the probability distributions for the outcomes are different even though the observed data (50 heads out of 100 flips) are the same for the two experiments.

Fisher proposed *p* as an informal measure of evidence against the null hypothesis. He called on researchers to combine *p* in the mind with other types of evidence for and against that hypothesis, such as the a priori plausibility of the hypothesis and the relative strengths of results from previous studies.[20]

Many misunderstandings concerning *p* arise because statistics classes and instructional materials ignore or at least do not emphasize the role of prior evidence in interpreting *p*; thus, the *p*-value is sometimes portrayed as the main result of statistical significance testing, rather than the acceptance or rejection of the null hypothesis at a pre-prescribed significance level. A renewed emphasis on prior evidence could encourage researchers to place *p* in the proper context, evaluating a hypothesis by weighing *p* together with all the other evidence about the hypothesis.

Related quantities

A closely related concept is the **E-value**,[21] which is the average number of times in multiple testing that one expects to obtain a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. The E-value is the product of the number of tests and the *p*-value.

The **'inflated' (or adjusted) *p*-value**,[22] is when a group of *p*-values are changed according to some multiple comparisons procedure so that each of the adjusted *p*-values can now be compared to the same threshold level of significance (α), while keeping the type I error controlled. The control is in the sense that the specific procedures controls it, it might be controlling the familywise error rate, the false discovery rate, or some other error rate.

Notes

[1] Stigler 2008.

[2] Dallal 2012, Note 31: Why P=0.05? (http://www.jerrydallal.com/LHSP/p05.htm).

[3] Valen E. Johnson and col, Revised standards for statistical evidence', Proceedings of the National Academy of Sciences of the United States of America, November 11, 2013 ()

[4] Andrew Gelmana and Christian P. Robert, 'Revised evidence for statistical standards', Proceedings of the National Academy of Sciences of the United States of America, April 23, 2014 ()

[5] Hubbard, R. (2004). Blurring the Distinctions Between p's and a's in Psychological Research, Theory Psychology June 2004 vol. 14 no. 3 295-327

[6] Babbie, E. (2007). The practice of social research 11th ed. Thomson Wadsworth: Belmont, CA.

[7] Note that the statistical significance of a result does not imply that the result is scientifically significant as well.

[8] It should be noted that a statistical hypothesis is conceptually different from a scientific hypothesis.

[9] Odds of TT is UNIQ-math-0-078b66d941de073b-QINU odds of HT and TH are UNIQ-math-1-078b66d941de073b-QINU and UNIQ-math-2-078b66d941de073b-QINU which are equal, and adding these yield UNIQ-math-3-078b66d941de073b-QINU

[10] Stigler 1986, p. 134.

[11] Pearson 1900.

[12] Inman 2004.

[13] Hubbard & Bayarri 2003, p. 1.

[14] Fisher 1925, p. 47, Chapter III. Distributions (http://psychclassics.yorku.ca/Fisher/Methods/chap3.htm).

[15] To be precise the $p = 0.05$ corresponds to about 1.96 standard deviations for a normal distribution (two-tailed test), and 2 standard deviations corresponds to about a 1 in 22 chance of being exceeded by chance, or $p \approx 0.045$; Fisher notes these approximations.

[16] Fisher 1925, pp. 78–79, 98, Chapter IV. Tests of Goodness of Fit, Independence and Homogeneity; with Table of χ^2 (http://psychclassics.yorku.ca/Fisher/Methods/chap4.htm), Table III. Table of χ^2 (http://psychclassics.yorku.ca/Fisher/Methods/tabIII.gif).

[17] Fisher 1971, II. The Principles of Experimentation, Illustrated by a Psycho-physical Experiment.

[18] Fisher 1971, Section 7. The Test of Significance.

[19] Fisher 1971, Section 12.1 Scientific Inference and Acceptance Procedures.

[20] Hubbard & Lindsay 2008.

[21] National Institutes of Health definition of E-value (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect)

[22] (page 815, second paragraph)

References

Further reading

- Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling" (http://www.economics.soton.ac.uk/staff/aldrich/1900.pdf). *Philosophical Magazine Series 5* **50** (302): 157–175. doi: 10.1080/14786440009463897 (http://dx.doi.org/10.1080/14786440009463897).
- Elderton, William Palin (1902). "Tables for Testing the Goodness of Fit of Theory to Observation". *Biometrika* **1** (2): 155–163. doi: 10.1093/biomet/1.2.155 (http://dx.doi.org/10.1093/biomet/1.2.155).
- Fisher, Ronald (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd. ISBN 0-05-002170-2.
- Fisher, Ronald A. (1971) [1935]. *The Design of Experiments* (9th ed.). Macmillan. ISBN 0-02-844690-9.
- Fisher, R. A.; Yates, F. (1938). *Statistical tables for biological, agricultural and medical research*. London.
- Stigler, Stephen M. (1986). *The history of statistics : the measurement of uncertainty before 1900*. Cambridge, Mass: Belknap Press of Harvard University Press. ISBN 0-674-40340-1.
- Hubbard, Raymond; Bayarri, M. J. (November 2003), *P Values are not Error Probabilities* (http://ftp.isds.duke.edu/WorkingPapers/03-26.pdf), a working paper that explains the difference between Fisher's evidential *p*-value and the Neyman–Pearson Type I error rate α.
- Hubbard, Raymond; Armstrong, J. Scott (2006). "Why We Don't Really Know What Statistical Significance Means: Implications for Educators" (http://repository.upenn.edu/cgi/viewcontent.cgi?article=1054&context=marketing_papers). *Journal of Marketing Education* **28** (2): 114. doi: 10.1177/0273475306288399 (http://dx.doi.org/10.1177/0273475306288399).
- Hubbard, Raymond; Lindsay, R. Murray (2008). "Why *P* Values Are Not a Useful Measure of Evidence in Statistical Significance Testing" (http://wiki.bio.dtu.dk/~agpe/papers/pval_notuseful.pdf). *Theory & Psychology* **18** (1): 69–88. doi: 10.1177/0959354307086923 (http://dx.doi.org/10.1177/0959354307086923).
- Stigler, S. (December 2008). "Fisher and the 5% level". *Chance* **21** (4): 12. doi: 10.1007/s00144-008-0033-3 (http://dx.doi.org/10.1007/s00144-008-0033-3).
- Dallal, Gerard E. (2012). *The Little Handbook of Statistical Practice* (http://www.tufts.edu/~gdallal/LHSP.HTM).

Further reading

- 12 Misconceptions, good overview given in following Article (http://xa.yimg.com/kq/groups/18751725/636586767/name/twelve+P+value+misconceptions.pdf)
- Presentation about the *p*-value (http://www.biostat.uzh.ch/aboutus/people/held/IFSPM.pdf)

External links

- Free online *p*-values calculators (http://www.danielsoper.com/statcalc/default.aspx#c14) for various specific tests (chi-square, Fisher's F-test, etc.).
- Understanding *p*-values (http://www.stat.duke.edu/~berger/p-values.html), including a Java applet that illustrates how the numerical values of *p*-values can give quite misleading impressions about the truth or falsity of the hypothesis under test.

Article Sources and Contributors

P-value *Source*: http://en.wikipedia.org/w/index.php?oldid=615841008 *Contributors*: A.M.R., ABoerma, Aaronbrick, Aedanpope, Alansohn, Alastair Haines, Algebraist, Alkarex, Amaralaw, Amkilpatrick, Anthonyhcole, Anypodetos, Arcadian, Avochelm, Avoided, Beckman16, BenFrantzDale, BetseyTrotwood, Billjefferys, Blaisorblade, Bob K31416, Bobblewik, Brandmaier, Bronstad, Btyner, CBM, CanadianLinuxUser, Cap'n Refsmmat, Capt hij, Cazort, Chefyingi, Chenmen2, Cherkash, Chezsruli, Chickenflicker, Chuckiesdad, Cola Turka, Crichto7, CrizCraig, Cru3r, Cyberbob240, Cybercobra, Cynical, DARTH SIDIOUS 2, Dan55886, DanSoper, Danko Georgiev, DavidCBryant, Dbachmann, Dcljr, Den fjättrade ankan, DerFrischmaker, Drono, DwightKingsbury, EJM86, Eliezg, Eric Kvaalen, Ethan Mitchell, Everyking, Fgnievinski, Flarity, Fnielsen, Furrykef, Giftlite, Goudzovski, Grochim, Gspr, Haein45, IdealistCynic, Ioannes Pragensis, Iskand26, JPLeRouzic, JackWasey, Jackmcbarn, Jflycn, Jimjamjak, Jo3sampl, John Quiggin, Johnmperry, Jonathan.asbell, Jovianeye, Junling, KKoolstra, Kairotic, Khahstats, Kingpin13, Kklamb, Krotera, Kwischan, L Kensington, L353a1, Labnoor, Lbertolotti, LeilaniLad, Lesath, Liangent, Libcub, Loodog, Loureiroandre, MZMcBride, Maghnus, Magister Mathematicae, Mandarax, Manoguru, Marcomio, Materialscientist, Mathstat, Mcld, Melcombe, Mgreenbe, Michael Hardy, Mike Rosoft, Mild Bill Hiccup, Miserlou, Mkdw, Mmaananda, Mogism, Mu5ti, Mycatharsis, Nbarth, Nemo bis, Neo Poz, Neonumbers, Nicholas Sund, Nouse4aname, Oleg Alexandrov, Ostracon, Ostrouchov, Paulsombart, Pchancharl, PenguiN42, Pinethicket, Pooven, Prostatguru, Quantling, Quercus solaris, Qwertyus, Qwfp, Raandrade, Ravster, Reedy, Repapetilto, Richard001, Richdiesal, Rjwilmsi, Rkb, Robma, Rod57, Rogermw, Rotring, Rrmcpj, Rstatx, Ruman, SPOBrien, Sam Blacketer, Savidan, Scorpi0n, Seren-dipper, Siroxo, Slazenger, Stangaa, Staticshakedown, SweetInfection, Talgalili, Tayste, Tdent, Tethros, Tetracube, The Anome, TheDamian, Thorwald, Timflutre, TrickyTank, TuTu522, Tucoxn, Tweenk, Ucucha, Urdutext, UsingFacts, Viraltux, Vovchyck, Wavelength, Welhaven, Whym, Widr, Wikky Horse, William Avery, Wywin, Xiaowei JIANG, Yoshigev, Yworo, Zvika, 368 anonymous edits

Image Sources, Licenses and Contributors

File:P-value Graph.png *Source*: http://en.wikipedia.org/w/index.php?title=File:P-value_Graph.png *License*: GNU Free Documentation License *Contributors*: Repapetilto

License