[BROKEN](/wiki/BROKEN" \o "BROKEN) In [frequentist statistics](/wiki/Frequentist_statistics), the ***p*-value** is a function of the observed sample results (a [test statistic](/wiki/Test_statistic)) relative to a [statistical model](/wiki/Statistical_model), which measures how extreme the observation is. Statistical hypothesis tests making use of *p*-values are commonly used in many fields of science and social sciences,[[1]](#cite_note-1) such as [economics](/wiki/Economics), [psychology](/wiki/Psychology),[[2]](#cite_note-2) [biology](/wiki/Biology), [criminal justice](/wiki/Criminal_justice) and [criminology](/wiki/Criminology), and [sociology](/wiki/Sociology).[[3]](#cite_note-3) Their misuse has been a matter of considerable controversy.

## Contents

* 1 Overview and controversy[[edit](/index.php?title=(none)&action=edit&section=1)]
* 2 Basic concepts[[edit](/index.php?title=(none)&action=edit&section=2)]
* 3 Definition and interpretation[[edit](/index.php?title=(none)&action=edit&section=3)]
* 4 Calculation[[edit](/index.php?title=(none)&action=edit&section=4)]
* 5 Examples[[edit](/index.php?title=(none)&action=edit&section=5)]
  + 5.1 One roll of a pair of dice[[edit](/index.php?title=(none)&action=edit&section=6)]
  + 5.2 Five heads in a row[[edit](/index.php?title=(none)&action=edit&section=7)]
  + 5.3 Sample size dependence[[edit](/index.php?title=(none)&action=edit&section=8)]
  + 5.4 Alternating coin flips[[edit](/index.php?title=(none)&action=edit&section=9)]
  + 5.5 Coin flipping[[edit](/index.php?title=(none)&action=edit&section=10)]
* 6 Distribution[[edit](/index.php?title=(none)&action=edit&section=11)]
* 7 History[[edit](/index.php?title=(none)&action=edit&section=12)]
* 8 Misunderstandings[[edit](/index.php?title=(none)&action=edit&section=13)]
* 9 Criticisms[[edit](/index.php?title=(none)&action=edit&section=14)]
* 10 Related quantities[[edit](/index.php?title=(none)&action=edit&section=15)]
* 11 Orthographic note[[edit](/index.php?title=(none)&action=edit&section=16)]
* 12 See also[[edit](/index.php?title=(none)&action=edit&section=17)]
* 13 Notes[[edit](/index.php?title=(none)&action=edit&section=18)]
* 14 References[[edit](/index.php?title=(none)&action=edit&section=19)]
* 15 Further reading[[edit](/index.php?title=(none)&action=edit&section=20)]
* 16 External links[[edit](/index.php?title=(none)&action=edit&section=21)]

## Overview and controversy[[edit](/index.php?title=(none)&action=edit&section=1)]

The *p*-value is defined as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the [null hypothesis](/wiki/Null_hypothesis) is true.[Template:Sfn](/wiki/Template:Sfn)[[4]](#cite_note-4) In [frequentist inference](/wiki/Frequentist_inference), the *p*-value is widely used in [statistical hypothesis testing](/wiki/Statistical_hypothesis_testing), specifically in [null hypothesis significance testing](/wiki/Null_hypothesis_significance_testing). In this method, as part of [experimental design](/wiki/Experimental_design), before performing the experiment, one first chooses a model (the [null hypothesis](/wiki/Null_hypothesis)) and a threshold value for *p*, called the [significance level](/wiki/Significance_level) of the test, traditionally 5% or 1% [[5]](#cite_note-5) and denoted as *α*. If the *p*-value is less than or equal to the chosen significance level (*α*), the test suggests that the observed data is inconsistent with the [null hypothesis](/wiki/Null_hypothesis), so the null hypothesis must be rejected. However, that does not prove that the tested hypothesis is true. When the *p*-value is calculated correctly, this test guarantees that the [Type I error rate](/wiki/Type_I_error_rate) is at most *α*. For typical analysis, using the standard *α* = 0.05 cutoff, the null hypothesis is rejected when *p* < .05 and not rejected when *p* > .05. The *p*-value does not in itself support reasoning about the probabilities of hypotheses but is only a tool for deciding whether to reject the null hypothesis.

The [American Statistical Association](/wiki/American_Statistical_Association), in a statement on the use of *p*-values,[[6]](#cite_note-6) affirmed the usefulness of properly interpreted *p*-values, but cautioned that *p*-values are "commonly misused and misinterpreted." The use of bright-line rules as cutoffs, such as p ≤ 0.05, without other supporting statistical evidence, was particularly criticized:[[6]](#cite_note-6)[Template:Quote](/wiki/Template:Quote)

While there is widespread agreement that *p*-values are often misused,[[7]](#cite_note-7)[[8]](#cite_note-8) there is no consensus on alternatives.[[9]](#cite_note-9)

## Basic concepts[[edit](/index.php?title=(none)&action=edit&section=2)]

The *p*-value is used in the context of [null hypothesis](/wiki/Null_hypothesis) testing in order to quantify the idea of [statistical significance](/wiki/Statistical_significance) of evidence.[Template:Efn](/wiki/Template:Efn) Null hypothesis testing is a [reductio ad absurdum](/wiki/Reductio_ad_absurdum) argument adapted to statistics. In essence, a claim is shown to be valid by demonstrating the improbability of the consequence that results from assuming the counter-claim to be true.

As such, the only hypothesis that needs to be specified in this test and which embodies the counter-claim is referred to as the [null hypothesis](/wiki/Null_hypothesis). A result is said to be statistically significant if it allows us to reject the null hypothesis. That is, as per the reductio ad absurdum reasoning, the statistically significant result should be highly improbable if the null hypothesis is assumed to be true. The rejection of the null hypothesis implies that the correct hypothesis lies in the logical complement of the null hypothesis. However, unless there is a single alternative to the null hypothesis, the rejection of null hypothesis does not tell us which of the alternatives might be the correct one.

For instance, if the null hypothesis is assumed to be a standard normal distribution N(0,1), the rejection of this null hypothesis can either mean (i) the mean is not zero, or (ii) the variance is not unity, or (iii) the distribution is not normal, depending on the type of test performed. However, supposing we manage to reject the zero mean hypothesis, even if we know the distribution is normal and variance is unity, the null hypothesis test does not tell us which non-zero value we should adopt as the new mean.

In statistics, a statistical hypothesis refers to a probability distribution that is assumed to govern the observed data.[Template:Efn](/wiki/Template:Efn) If <math>X</math> is a [random variable](/wiki/Random_variable) representing the observed data and <math>H</math> is the statistical hypothesis under consideration, then the notion of statistical significance can be naively quantified by the [conditional probability](/wiki/Conditional_probability) <math>\Pr(X|H)</math>, which gives the likelihood of the observation if the hypothesis is *assumed* to be correct. However, if <math>X</math> is a continuous random variable and an instance <math>x</math> is observed, <math>\Pr(X=x|H)=0.</math> Thus, this naive definition is inadequate and needs to be changed so as to accommodate the continuous random variables.

Nonetheless, it helps to clarify that *p*-values should **not** be confused with probability on hypothesis (as is done in [Bayesian Hypothesis Testing](/wiki/Bayes_factor)) such as <math>\Pr(H|X),</math> the probability of the hypothesis given the data, or <math>\Pr(H),</math> the probability of the hypothesis being true, or <math>\Pr(X),</math> the probability of observing the given data.

## Definition and interpretation[[edit](/index.php?title=(none)&action=edit&section=3)]

[thumb|right|400 px|Example of a *p*-value computation. The vertical coordinate is the](/wiki/File:P-value_in_statistical_significance_testing.svg) [probability density](/wiki/Probability_density_function) of each outcome, computed under the null hypothesis. The *p*-value is the area under the curve past the observed data point.

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

* <math>\Pr(X \geq x |H)</math> for right tail event,
* <math>\Pr(X \leq x |H)</math> for left tail event,
* <math>2\min\{\Pr(X \leq x |H),\Pr(X \geq x |H)\}</math> for double tail event.

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 <math>H</math> is rejected if any of these probabilities is less than or equal to a small, fixed but arbitrarily pre-defined threshold value <math>\alpha</math>, which is referred to as the [level of significance](/wiki/Statistical_significance). Unlike the *p*-value, the <math>\alpha</math> level is not derived from any observational data and does not depend on the underlying hypothesis; the value of <math>\alpha</math> is instead determined by the consensus of the research community that the investigator is working in.

Since the value of <math>x</math> that defines the left tail or right tail event is a random variable, this makes the *p*-value a function of <math>x</math> and a [random variable](/wiki/Random_variable) in itself defined uniformly over <math>[0,1]</math> interval, assuming <math>x</math> is continuous. 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 the same test is repeated independently bearing upon the same overall null hypothesis, it will yield different *p*-values at every repetition. Nevertheless, these different *p*-values can be combined using [Fisher's combined probability test](/wiki/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.

The fixed pre-defined <math>\alpha</math> level can be interpreted as the rate of falsely rejecting the null hypothesis (or type I error), since

<math>\Pr(\mathrm{Reject}\; H|H) = \Pr(p \leq \alpha|H) = \alpha</math>.

This also means that if we fix an instantiation of *p*-value and allow <math>\alpha</math> to vary over <math>[0,1]</math>, we can obtain an equivalent interpretation of *p*-value in terms of <math>\alpha</math> level as the lowest value of <math>\alpha</math> that can be assumed for which the null hypothesis can be rejected for a given set of observations. Obviously, assuming an <math>\alpha</math> smaller than the instantiated *p*-value will end up not rejecting the null hypothesis.

## Calculation[[edit](/index.php?title=(none)&action=edit&section=4)]

Usually, instead of the actual observations, <math>X</math> is instead a [test statistic](/wiki/Test_statistic). A test statistic is a [scalar](/wiki/Scalar_(mathematics)) function of all the observations, such as the average or the correlation coefficient, which summarizes the characteristics of the data by a single number, relevant to a particular inquiry. As such, the test statistic follows a distribution determined by the function used to define that test statistic and the distribution of the input observational data.

For the important case in which the data are hypothesized to follow the normal distribution, depending on the nature of the test statistic and thus the underlying hypothesis of the test statistic, different null hypothesis tests have been developed. Some such tests are [z-test](/wiki/Z-test) for [normal distribution](/wiki/Normal_distribution), [t-test](/wiki/T-test) for [Student's t-distribution](/wiki/Student's_t-distribution), [f-test](/wiki/F-test) for [f-distribution](/wiki/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](/wiki/Central_limit_theorem) for large samples, as in the case of [Pearson's chi-squared test](/wiki/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](/wiki/One-tailed_test) or a [two-tailed test](/wiki/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 [cumulative distribution function](/wiki/Cumulative_distribution_function) (CDF) is often a difficult computation. Today, this computation is done using statistical software, often via numeric methods (rather than exact formulae), but 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](/wiki/Quantile_function) (inverse CDF).

## Examples[[edit](/index.php?title=(none)&action=edit&section=5)]

Here a few simple examples follow, each illustrating a potential pitfall. [Template:Original research](/wiki/Template:Original_research)

### One roll of a pair of dice[[edit](/index.php?title=(none)&action=edit&section=6)]

Suppose a researcher rolls a pair of dice once and assumes a null hypothesis that the dice are fair, not loaded or weighted toward any specific number/roll/result; uniform. 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 (because under the assumption of the null hypothesis, the test statistic is uniformly distributed) 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, this result would be deemed significant and the hypothesis that the dice are fair would be rejected.

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](/wiki/Experiment_design).

### Five heads in a row[[edit](/index.php?title=(none)&action=edit&section=7)]

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, but 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 ≈ 0.03. If the researcher assumed a significance level of 0.05, this result would be deemed significant and the hypothesis that the coin is fair would be rejected. 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×(1/2)5 = 1/16 ≈ 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[[edit](/index.php?title=(none)&action=edit&section=8)]

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 and is a two-tailed test. Suppose the researcher observes heads for each flip, yielding a test statistic of *n* and a *p*-value of 2/2n. 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 ≈ 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. 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 can never be below 0.05 for the coin example).

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

### Alternating coin flips[[edit](/index.php?title=(none)&action=edit&section=9)]

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 that is the expected number of heads.

Suppose instead that the 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. That would yield a test statistic of 9, which is extreme and has a *p*-value of <math>1/2^9 = 1/512 \approx 0.0039</math>. That 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, but 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 may be 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.

### Coin flipping[[edit](/index.php?title=(none)&action=edit&section=10)]

[Template:Main](/wiki/Template:Main) As an example of a statistical test, an experiment is performed to determine whether a [coin flip](/wiki/Coin_flipping) is [fair](/wiki/Fair_coin) (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 a right-tailed test is considered, the *p*-value of this result is the chance of a fair coin landing on heads *at least* 14 times out of 20 flips. That probability can be computed from [binomial coefficients](/wiki/Binomial_coefficient) as

<math>

\begin{align} & \operatorname{Prob}(14\text{ heads}) + \operatorname{Prob}(15\text{ heads}) + \cdots + \operatorname{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{align} </math>

This probability is the *p*-value, considering only extreme results that favor heads. This is called a [one-tailed test](/wiki/One-_and_two-tailed_tests). However, the deviation can be in either direction, favoring either heads or tails. The two-tailed *p*-value, which considers deviations favoring either heads or tails, may instead be calculated. As the [binomial distribution](/wiki/Binomial_distribution) is symmetrical for a fair coin, the two-sided *p*-value is simply twice the above calculated single-sided *p*-value: the two-sided *p*-value is 0.115.

In the above example:

* Null hypothesis (H0): The coin is fair, with 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 H0 = 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 that 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 is in fact fair. Hence, the null hypothesis at the 5% level is not rejected. 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%). The null hypothesis is rejected when a 5% cut-off is used.

## Distribution[[edit](/index.php?title=(none)&action=edit&section=11)]

When the null hypothesis is true, the [probability distribution](/wiki/Probability_distribution) of the *p*-value is [uniform](/wiki/Uniform_distribution_(continuous)) on the interval [0,1]. By contrast, if the alternative hypothesis is true, the distribution is dependent on sample size and the true value of the parameter being studied.[[1]](#cite_note-1)[[10]](#cite_note-10) The distribution of *p*-values for a group of studies is called a *p*-curve.[[11]](#cite_note-11) The curve is affected by four factors: the probability that a study is examining a true hypothesis rather than a false hypothesis, the [power](/wiki/Statistical_power) of the studies investigating true hypotheses, the Type 1 error rates, and [publication bias](/wiki/Publication_bias).[[12]](#cite_note-12) A *p*-curve can be used to assess the reliability of scientific literature, such as by detecting publication bias or *p*-hacking.[[11]](#cite_note-11)[[13]](#cite_note-13)

## History[[edit](/index.php?title=(none)&action=edit&section=12)]

[thumb|right|Pierre-Simon Laplace](/wiki/File:Pierre-Simon-Laplace_(1749-1827).jpg) [thumb|right|200px|Biologist and statistician Ronald Fisher](/wiki/File:Biologist_and_statistician_Ronald_Fisher.jpg)

Computations of *p*-values date back to the 1770s, where they were calculated by [Pierre-Simon Laplace](/wiki/Pierre-Simon_Laplace):[Template:Sfn](/wiki/Template:Sfn) [Template:Quotation](/wiki/Template:Quotation)

The *p*-value was first formally introduced by [Karl Pearson](/wiki/Karl_Pearson), in his [Pearson's chi-squared test](/wiki/Pearson's_chi-squared_test),[Template:Sfn](/wiki/Template:Sfn) using the [chi-squared distribution](/wiki/Chi-squared_distribution) and notated as capital P.[Template:Sfn](/wiki/Template:Sfn) The *p*-values for the [chi-squared distribution](/wiki/Chi-squared_distribution) (for various values of *χ*2 and degrees of freedom), now notated as *P,* was calculated in [Template:Harv](/wiki/Template:Harv), collected in [Template:Harv](/wiki/Template:Harv).

The use of the *p*-value in statistics was popularized by [Ronald Fisher](/wiki/Ronald_Fisher),[Template:Sfn](/wiki/Template:Sfn) and it plays a central role in his approach to the subject.[Template:Sfn](/wiki/Template:Sfn) In his influential book [*Statistical Methods for Research Workers*](/wiki/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](/wiki/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](/wiki/68–95–99.7_rule)).[Template:Sfn](/wiki/Template:Sfn)[Template:Efn](/wiki/Template:Efn)[Template:Sfn](/wiki/Template:Sfn)

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.[Template:Sfn](/wiki/Template:Sfn) That 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 [Template:Harv](/wiki/Template:Harv), which cemented the approach.[Template:Sfn](/wiki/Template:Sfn)

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

To evaluate a lady's claim that she ([Muriel Bristol](/wiki/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 that case, the null hypothesis was that she had no special ability, the test was [Fisher's exact test](/wiki/Fisher's_exact_test), and the *p*-value was <math>1/\binom{8}{4} = 1/70 \approx 0.014,</math> 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:[Template:Sfn](/wiki/Template:Sfn) [Template:Quotation](/wiki/Template:Quotation) 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 <math>1/\binom{6}{3} = 1/20 = 0.05,</math> which would not have met this level of significance.[Template:Sfn](/wiki/Template:Sfn) 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".[Template:Sfn](/wiki/Template:Sfn) 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[[edit](/index.php?title=(none)&action=edit&section=13)]

[Template:Main](/wiki/Template:Main) [Template:See also](/wiki/Template:See_also) [Template:Merge from](/wiki/Template:Merge_from)

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.

The *p*-value does not in itself allow reasoning about the probabilities of hypotheses, which requires multiple hypotheses or a range of hypotheses, with a [prior distribution](/wiki/Prior_distribution) of likelihoods between them, as in [Bayesian statistics](/wiki/Bayesian_statistics). There, one uses a [likelihood function](/wiki/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](/wiki/Alternative_hypothesis) in Neyman–Pearson [statistical hypothesis testing](/wiki/Statistical_hypothesis_testing). In that approach, one instead has a decision function between two alternatives, often based on a [test statistic](/wiki/Test_statistic), and computes the rate of [type I and type II errors](/wiki/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.

## Criticisms[[edit](/index.php?title=(none)&action=edit&section=14)]

[Template:Main](/wiki/Template:Main)

Critics of *p*-values point out that the criterion used to decide "statistical significance" is based on an arbitrary choice of alpha level (often set at 0.05), and that this criterion leads to an alarming number of false positive tests. The fraction of all “statistically significant” tests in which the null hypothesis is true may be considerably higher than the alpha level, depending on how many of the null hypotheses were false and how much statistical power was used to test them.<ref name=SBB2001>[Template:Cite journal](/wiki/Template:Cite_journal)</ref><ref name=VJ2013>[Template:Cite journal](/wiki/Template:Cite_journal)</ref><ref name=DC2015>[Template:Cite journal](/wiki/Template:Cite_journal)</ref>

Dividing results into significant and non-significant effects can be highly misleading.[[8][8]](#cite_note-8)[[14]](#cite_note-14) For instance, analysis of nearly identical datasets can result in *p*-values that differ greatly in significance.[[14]](#cite_note-14) In [medical research](/wiki/Medical_research), *p*-values were a considerable improvement over previous approaches, but misunderstandings of *p*-values have become more important for reasons such as the increased statistical complexity of published research.[[8]](#cite_note-8) It has been suggested that in fields such as [psychology](/wiki/Psychology), where studies typically have low [statistical power](/wiki/Statistical_power), using significance testing can lead to increased error rates.[[14]](#cite_note-14)[[15]](#cite_note-15) The use of significance testing as the basis for decisions has also been criticized because of the widespread misunderstandings about the process.[[8]](#cite_note-8)[[16]](#cite_note-16)[[17]](#cite_note-17) For example, *p*-values do not address the probability of the null hypothesis being true or false, and the choice of significance threshold should not be arbitrary but instead informed by the consequences of a false positive.[[14]](#cite_note-14) It is possible to use [Bayes factors](/wiki/Bayes_factors) for calibration, which allows the use of *p*-values while reducing the impact of the *p*-value fallacy, although these approaches introduce other biases as well.[[18]](#cite_note-18) The *p*-value is incompatible with the [likelihood principle](/wiki/Likelihood_principle) and depends on the [experiment design](/wiki/Experiment_design), the test statistic in question. That is, the definition of "more extreme" data depends on the sampling methodology adopted by the investigator;[[19]](#cite_note-19) 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.[[20]](#cite_note-20) The incompatibility of the *p*-value with the likelihood principle indicates a lack of conceptual integrity in this methodology since the intentions of the analyst to analyze the data along differing prescriptions (experimental designs, choice of statistic) yield different results for the same data concerning the same underlying question about the coin's fairness. This concern over the issue of consistent reasoning concerning probable inference led [Richard Cox](/wiki/Richard_Threlkeld_Cox) to develop an axiomatic basis for probability conditioned on an essential consistency requirement that leads ultimately to [Bayes Rule](/wiki/Bayes_Rule).[[21]](#cite_note-21)

## Related quantities[[edit](/index.php?title=(none)&action=edit&section=15)]

A closely related concept is the **E-value**,[[22]](#cite_note-22) which is the [expected](/wiki/Conditional_expectation) number of times in [multiple testing](/wiki/Multiple_comparisons) that one expects to obtain a test statistic at least as extreme as the one that was actually observed if one assumes that the null hypothesis is true. The E-value is the product of the number of tests and the *p*-value.

## Orthographic note[[edit](/index.php?title=(none)&action=edit&section=16)]

Depending on which [style guide](/wiki/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). For example, [AMA style](/wiki/AMA_style) uses *P* value but [APA style](/wiki/APA_style) uses *p* value.

## See also[[edit](/index.php?title=(none)&action=edit&section=17)]

* [Bonferroni correction](/wiki/Bonferroni_correction)
* [Confidence interval](/wiki/Confidence_interval)
* [Counternull](/wiki/Counternull)
* [False discovery rate](/wiki/False_discovery_rate)
* [Fisher's method of combining *p*-values](/wiki/Fisher's_method)
* [Generalized *p*-value](/wiki/Generalized_p-value)
* [Holm–Bonferroni method](/wiki/Holm–Bonferroni_method)
* [Multiple comparisons](/wiki/Multiple_comparisons)
* [Null hypothesis](/wiki/Null_hypothesis)
* [*p*-rep](/wiki/P-rep)
* [*p*-value fallacy](/wiki/P-value_fallacy)
* [Statistical hypothesis testing](/wiki/Statistical_hypothesis_testing)

## Notes[[edit](/index.php?title=(none)&action=edit&section=18)]

[Template:Notelist](/wiki/Template:Notelist)

## References[[edit](/index.php?title=(none)&action=edit&section=19)]

[Template:Reflist](/wiki/Template:Reflist) [Template:Refbegin](/wiki/Template:Refbegin)

## Further reading[[edit](/index.php?title=(none)&action=edit&section=20)]

[Template:More footnotes](/wiki/Template:More_footnotes)

* [Template:Cite journal](/wiki/Template:Cite_journal)
* [Template:Cite journal](/wiki/Template:Cite_journal)
* [Template:Cite book](/wiki/Template:Cite_book)
* [Template:Cite book](/wiki/Template:Cite_book)
* [Template:Cite book](/wiki/Template:Cite_book)
* [Template:Cite book](/wiki/Template:Cite_book)
* [Template:Citation](/wiki/Template:Citation)
* [Template:Cite journal](/wiki/Template:Cite_journal)
* [Template:Cite journal](/wiki/Template:Cite_journal)
* [Template:Cite journal](/wiki/Template:Cite_journal)
* [Template:Cite book](/wiki/Template:Cite_book)
* [Template:Cite journal](/wiki/Template:Cite_journal)
* [Template:Cite book](/wiki/Template:Cite_book)

[Template:Refend](/wiki/Template:Refend)

## External links[[edit](/index.php?title=(none)&action=edit&section=21)]

[Template:Commons category](/wiki/Template:Commons_category)

* [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/%7Eberger/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.

[Template:Statistics](/wiki/Template:Statistics)

[Category:Hypothesis testing](/wiki/Category:Hypothesis_testing)

[ja:有意#p値](/wiki/Ja:有意#p値)