OPEN EDUCATIONAL RESOURCE

The Meaning of a p-Value: Definition, Interpretation and Significance

Derek Jansen, MBA
Grad Coach
Reviewed by Dr. Eunice Rautenbach, DTech · November 2026

Abstract

The *p*-value is among the most widely reported numbers in quantitative research and among the most often misread, to the point that the American Statistical Association issued a formal statement on its use. This article explains what a *p*-value measures: the probability, computed under a statistical model that includes the null hypothesis, of a result at least as extreme as the one observed. It sets out two common misreadings: that a *p*-value is the probability that the null hypothesis is true, and that it is the probability that the result arose by chance alone. It then distinguishes the *p*-value from the significance level fixed before the analysis, explains why statistical significance says nothing about the size of an effect, and summarizes Cohen's conventional benchmarks for effect sizes. An extract from a master's thesis shows a test statistic, a Bonferroni correction and effect sizes reported together, and a recalculation shows one of its results narrowly missing the corrected threshold. The article closes with the main errors of interpretation and the APA Style rules for reporting *p*-values. It is intended for students and early-career researchers interpreting or reporting statistical tests, and for the educators who teach statistics and research methods.

Keywords: p-value meaning; p-value definition; interpreting p-values; p-value significance; significance level; statistical significance; effect size; open educational resource

Suggested citation: Jansen, D. (2026). *The meaning of a p-value: Definition, interpretation and significance*. Grad Coach. https://doi.org/10.5281/zenodo.24198069

About this article: This is the formal edition of an article first published at gradcoach.com/what-does-p-value-actually-mean/. Further open educational resources on statistics and research methods are listed at gradcoach.com/open-educational-resources/.

© 2026 Grad Coach. Free to share and adapt under CC BY 4.0 with credit to Grad Coach and a link to gradcoach.com/what-does-p-value-actually-mean/. The quoted thesis extract remains its author's; the thesis keeps its own license.

1. Introduction

Few numbers in quantitative research are reported as often as the *p*-value, and few are so frequently misread. The misreading is common enough that the American Statistical Association (ASA) published a formal statement on *p*-values to correct it (Wasserstein & Lazar, 2016), and the shorthand definitions that many students learn are part of the problem (Greenland et al., 2016, p. 337).

The confusion usually concerns three questions: what a *p*-value actually measures, what counts as statistically significant, and what a *p*-value cannot show. This article is written for students and researchers who interpret and report statistical tests, and for those who teach them. Section 2 defines the *p*-value and two common misreadings of it. Section 3 distinguishes the *p*-value from the significance level. Section 4 explains why a *p*-value does not measure the size of an effect. Section 5 examines how a master's thesis reports its *p*-values, Section 6 describes two common errors, and Section 7 sets out how *p*-values are reported in APA Style.

2. What a p-Value Measures

A statistical test usually begins from a **null hypothesis**: the assumption that there is no effect, such as no difference between two groups or no relationship between two variables. The *p*-value answers one question about that hypothesis: **if the null hypothesis were true, how likely would a result at least as extreme as the observed one be?** The ASA statement defines it informally as "the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value" (Wasserstein & Lazar, 2016, p. 131). A small *p*-value therefore indicates that the data would be unusual if there were no real effect.

The phrase "specified statistical model" matters. The model contains the null hypothesis, but also every other assumption used to compute the *p*-value, such as random sampling, the distribution of the data and the way the analyses were chosen. Greenland et al. (2016) stress that the *p*-value tests the whole model, not only the null hypothesis: a very small value shows that the data would be unusual if every assumption were correct, but it does not show which assumption is incorrect (p. 339).

A *p*-value shows nothing more than this. The ASA statement names two common misreadings directly: *p*-values "do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone" (Wasserstein & Lazar, 2016, p. 131). The second misreading reverses the logic. A *p*-value is calculated on the assumption that chance alone is operating, so it cannot also be the probability that chance alone produced the result (Greenland et al., 2016, p. 340).

A study testing whether a new learning technique improves examination scores illustrates the point. The null hypothesis is that the technique makes no difference. A two-sided *p*-value of .03 means that, if the technique had no effect and the test's assumptions held, a difference at least as

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

large as the one observed, in either direction, would arise in about 3% of studies like this one. It does not mean that there is a 97% probability that the technique works, and it does not mean that there is a 3% probability that the result arose by chance alone.

3. The Significance Level and Statistical Significance

Before the analysis, the researcher sets a threshold called the **significance level**, or alpha (α). A result is described as statistically significant when the *p*-value falls at or below it. The significance level is usually .05 (Greenland et al., 2016, p. 339). The significance level and the *p*-value are easily confused but are different in kind: the significance level is fixed in advance as part of the study design and is not changed in light of the data, whereas the *p*-value is computed from the data and is a result of the analysis (Greenland et al., 2016, pp. 339–340).

Table 1 shows how ranges of *p*-values are conventionally read against a .05 threshold. One term in the table, **effect size**, is a measure of how large a difference or relationship is; Section 4 returns to it.

Table 1. Conventional readings of *p*-values against a significance level of .05.

p-value	At a .05 level	What it indicates about the data
Below .001	Significant	A result at least this extreme would occur less than 0.1% of the time if there were no real effect
.001 to under .01	Significant	Very unlikely under the null hypothesis: less than 1% of the time
.01 to .05	Significant	Unlikely under the null hypothesis, but the effect size and the number of tests run should be checked
Above .05 to .10	Not significant	Somewhat unusual under the null hypothesis; reported as not significant rather than "nearly" significant
Above .10	Not significant	Not surprising under the null hypothesis, which is not the same as showing that there is no effect

The .05 threshold is a convention rather than a fixed rule. Some studies set a stricter level, such as .01 (Cohen, 1992), and some researchers have proposed lowering the default threshold for claims of new discoveries to .005 (Benjamin et al., 2018).

The ASA statement goes further: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold" (Wasserstein & Lazar, 2016, p. 131). Instead, researchers should take contextual factors into account, "including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis" (p. 131). In practice, a *p*-value is read alongside:

- **the effect size**: how large the difference or relationship is (Section 4);
- **the design of the study**: the sample size, how the sample was drawn and what was controlled for;
- **the quality of the measures**: whether the instruments are validated and reliable;

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

- **the assumptions of the test**: for example, normality or equal variances, where the test relies on them;
- **the findings of other studies**: whether the result is consistent with the existing literature or contradicts it.

Many researchers also report a **confidence interval**. Instead of a yes-or-no verdict, a confidence interval gives the range of effect sizes that are more compatible with the data, under the statistical model, than values outside the range (Greenland et al., 2016, p. 340).

A researcher should find out which threshold the field uses and why, set it before examining the data, and state it in the methods section. Changing the threshold after the results are known defeats the purpose of fixing it in advance (Greenland et al., 2016, pp. 339–340).

4. p-Values Do Not Measure Effect Size

A small p-value does not mean that an important effect has been found. A p-value concerns statistical significance; it **says nothing about practical significance**, which is judged from the effect size. The ASA statement is explicit: a p-value "does not measure the size of an effect or the importance of a result" (Wasserstein & Lazar, 2016, p. 132). With a large enough sample, even a tiny and practically meaningless difference can produce a very small p-value; with a small sample, a large and important difference can fail to reach significance (p. 132).

For this reason, an effect size should be reported alongside every p-value. Effect sizes tell the reader how large a difference or relationship is, and Cohen's conventional benchmarks are the usual starting point for interpreting them (Table 2).

Table 2. Cohen's benchmarks for small, medium and large effects.

Effect size measure	Small	Medium	Large
Cohen's d (difference between two group means)	0.2	0.5	0.8
Eta squared, η^2 (share of variance explained)	.01	.06	.14
Pearson's r (correlation)	.1	.3	.5

Note. The d and r values are from Cohen (1992). The η^2 values are Cohen's (1988), as cited in Lakens (2013, p. 7); Cohen (1992) gives the equivalent values of Cohen's f, his effect size for analysis of variance (the spread of the group means relative to the spread within groups): .10, .25 and .40.

Two cautions apply. First, statistical software such as SPSS reports **partial eta squared**. In a one-way between-subjects design (one factor, with each participant in only one group), partial eta squared equals eta squared. In designs with further factors, covariates or repeated measures, the two differ, because partial eta squared leaves the variance from those other sources out of the total it divides by (Lakens, 2013, pp. 1, 5–7). A reader should check which of the two a study reports before comparing it with these benchmarks. Second, the benchmarks are rough guides rather than rules. The values for d are arbitrary and should not be interpreted rigidly (Thompson, 2007, as cited in Lakens, 2013, p. 3). These benchmarks are best used only when findings are so novel

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

that they cannot be compared with related findings in the literature (Cohen, 1988, as cited in Lakens, 2013, p. 3). Where published studies in the field report typical effect sizes, an effect size is better compared with those (Lakens, 2013, p. 7).

5. A Worked Example from a Master's Thesis

A Master of Arts thesis from Purdue University's Department of Languages and Cultures shows these ideas in practice. Osumi (2019) surveyed 100 university students of Japanese and compared their motivation across four academic reasons for taking the course: as a major, as a minor, as a requirement, or for fun. The following passage reports three separate one-way analyses of variance (ANOVAs), each comparing the mean scores of the four groups on one motivational factor. Each result gives the test statistic F, the ratio of the variance between the groups to the variance within the groups. The statistic is followed in brackets by its two degrees of freedom (the number of groups minus one, and the number of participants included in the test minus the number of groups), and then the *p*-value. The final sentence gives an effect size for each test, eta squared (η^2), the share of the variance in scores explained by group membership. Table 3 sets out what the passage does well:

> To explore these factors more in-depth, three one-way ANOVAs were performed with academic reasons (major, minor, requirement, and for fun) as the independent variable and these three motivational factors (Instrumentality-promotion, Instrumentality-prevention, and the ideal L2 self) as the dependent variable, and the results are shown in Table 8 [in the original]. Significant differences were found in all three motivational factors at the 5% level ($\alpha = 0.017$ Bonferroni corrected): Instrumentality-promotion ($F [3, 94] = 4.821$, $p < 0.004$), Instrumentality-prevention ($F [3, 95] = 4.21$, $p = 0.008$), and the ideal L2 self ($F [3, 94] = 3.549$, $p = 0.017$). The effect sizes were all medium (Instrumentality-promotion, $\eta^2 = 0.13$; Instrumentality-prevention, $\eta^2 = 0.12$; the ideal L2 self, $\eta^2 = 0.10$).
>
> Osumi (2019, pp. 64–65)

The passage makes three sound reporting moves (Table 3).

Table 3. Three reporting moves in the extract from Osumi (2019).

The move	Where it occurs, and why it matters
Reports the test, not only the *p*-value	"F [3, 94] = 4.821". The test statistic and its degrees of freedom allow a reader to check the result rather than accept the *p*-value on trust.
Adjusts for running three tests	"α = 0.017 Bonferroni corrected". The .05 threshold is divided by the number of tests (.05/3), so that the probability of at least one false positive (a significant result when the null hypothesis is true) across all three stays at or below 5% (Bland & Altman, 1995).
Reports effect sizes	"The effect sizes were all medium". Against Cohen's benchmarks, values of .10 to .13 lie between medium (.06) and large (.14).

The third result needs a closer look. The exact corrected threshold is .05/3, about .0167; the thesis rounds it to .017 and reports the *p*-value for this test as .017 as well, so the two numbers look identical. Recalculated from the reported $F(3, 94) = 3.549$, however, the exact *p*-value is about .0174,

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

which lies just above both .0167 and .017. Strictly, this result does not fall at or below the corrected threshold, although the thesis counts it as significant. The other two results are unaffected: recalculated from their F statistics, they are about .0036 and .0077.

Here **the effect size is useful**. An η^2 of .10 tells an examiner how large the difference is, on whichever side of the threshold the p-value falls. A borderline result is best reported with the exact p-value and the exact threshold, and interpreted through the effect size, rather than rounded until it appears significant.

6. Two Common Errors

The first error is **treating a p-value above .05 as proof that there is no effect**. A p-value of .06 shows only that the data do not clear the threshold that was set. A large p-value indicates only that the data would not be unusual if the null hypothesis and the model's other assumptions were correct, and that the same data would also not be unusual under many other hypotheses (Greenland et al., 2016, p. 341). The study may have had too little statistical power, which is the pre-study probability that the test will reject the null hypothesis when a real effect of a specified size exists (Greenland et al., 2016, p. 345). In a small study especially, even large effects may be "drowned in noise" and fail to reach statistical significance (Greenland et al., 2016, p. 341).

The second error is **running many tests without adjusting the significance level**. Each test at .05 carries a 5% risk of a false positive when there is no real effect, and across many tests these risks accumulate. Across 20 independent tests of true null hypotheses, the probability that none is significant is $.95^{20} \approx .36$, so the probability of at least one "significant" result is about .64 (Bland & Altman, 1995). A correction such as the Bonferroni method, used in the example in Section 5, limits that risk. Statistical tests should also be planned in advance and all of them reported, not only those that reached significance: reporting only selected analyses renders the reported p-values "essentially uninterpretable" (Wasserstein & Lazar, 2016, p. 132).

7. Reporting p-Values in APA Style

Many degree programs, especially in the social sciences, follow APA Style for statistics, and its rules for p-values are short (American Psychological Association, 2024):

- **Report the exact value** to two or three decimal places, such as $p = .032$.
- **Below .001, write $p < .001$** rather than the exact value.
- **Omit the leading zero**, because a p-value cannot be greater than 1: $p = .008$, not $p = 0.008$.
- **Italicize the p**, as with other letters used as statistical symbols.

It follows that $p = .000$ should never be reported, even when software displays it. A p-value can be extremely small but not zero; the display has been rounded, and the value is reported as $p < .001$.

The p-value should also be reported with its context: the test statistic, its degrees of freedom and an effect size. Written in APA Style, the second result in the example in Section 5 would read $F(3, 95) = 4.21$, $p = .008$, $\eta^2 = .12$.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

8. Conclusion

A p-value indicates how unusual the data would be if there were no real effect and the other assumptions of the statistical model held. It is not the probability that the null hypothesis is true, nor the probability that the result arose by chance alone. The significance level against which it is judged is a convention, to be chosen and justified before the analysis. Statistical significance says nothing about the size or importance of an effect, so an effect size belongs beside every p-value.

The interpretation of any p-value also depends on its research context. The sample size, the design and the field all shape how much a given result shows. For example, the same p-value can reflect a trivial effect in a very large study and a substantial one in a small pilot. Furthermore, a result from a well-designed, preregistered trial deserves more trust than the same p-value from an exploratory analysis. A preregistered trial is one whose hypotheses and analysis plan were registered publicly before the data were collected or analyzed. Published studies and completed theses in the same area show how researchers in that field report their p-values, which thresholds they use and which effect sizes they treat as meaningful. Reporting exact values, correcting for multiple tests and reporting every test conducted allow a reader to judge the evidence on its merits.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

References

American Psychological Association. (2024). *APA Style numbers and statistics guide*. https://apastyle.apa.org/instructional-aids/numbers-statistics-guide.pdf

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ*, 310(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Osumi, A. (2019). *Exploring motivation for learning Japanese as a foreign language* [Master's thesis, Purdue University]. Purdue University Graduate School. https://doi.org/10.25394/PGS.8023421.v1

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com