OPEN EDUCATIONAL RESOURCE

# Inferential Statistics: Definition, Types and Examples

**Derek Jansen, MBA**
Grad Coach
Reviewed by Kerryn Warren, PhD · November 2026

## Abstract

Inferential statistics allow a researcher to draw conclusions about a population from a sample, but the logic behind them is misread more often than the arithmetic. A result can be overstated, or a real effect dismissed because a *p*-value fell just above .05. This article defines inferential statistics and sets out how they differ from descriptive statistics. It explains the idea on which all inference rests, sampling error and the central limit theorem, and the two tasks that inference performs: estimation with confidence intervals, and hypothesis testing with *p*-values, including the two misreadings of *p*-values that are most common. It then describes five common inferential tests (the *t*-test, analysis of variance, the chi-square test, correlation and regression), each with a worked or simulated example and, for three of them, an extract from a doctoral study showing how the result was reported. Throughout, it distinguishes what a test can show from what only the research design can establish. It is intended for students and early-career researchers analyzing or reporting quantitative data, and for the educators who teach research methods and statistics.

**Keywords:** inferential statistics; inferential statistics examples; types of inferential statistics; descriptive vs inferential statistics; confidence intervals; hypothesis testing; statistical tests; open educational resource

**Suggested citation:** Jansen, D. (2026). *Inferential statistics: Definition, types and examples.* Grad Coach. https://doi.org/10.5281/zenodo.24197814

**About this article:** This is the formal edition of an article first published at gradcoach.com/inferential-statistics/, where the figures can also be downloaded for reuse. Further open educational resources on statistics and analysis 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/inferential-statistics/. Quoted extracts remain their authors'; each source keeps its own license.

## 1. Introduction

Inferential statistics are used when a sample must represent a population it only partly covers. The calculations are seldom the difficulty, since software performs them. The difficulty lies in the reasoning behind them. Misinterpreted results can lead a study to claim a finding its data cannot support, or to discard a real effect because its $p$-value was .06.

This article is written for students and researchers who run and report inferential tests, and for those who teach them. Section 2 defines inferential statistics and the two tasks they perform. Section 3 sets out how they differ from descriptive statistics. Section 4 explains how inference works: sampling error, the central limit theorem, confidence intervals and hypothesis testing. Section 5 describes five common tests, three of them with an extract from a doctoral study showing how the result was reported and interpreted.

## 2. What Inferential Statistics Are

**Inferential statistics** are the methods used to draw conclusions about a population from a sample, together with a statement of how much uncertainty those conclusions carry. The Purdue Online Writing Lab (n.d.) describes their goal as discovering a property or general pattern in a large group by studying a smaller group, in the expectation that the results will generalize to the larger one. In practice, this involves two tasks:

1. **Estimation:** using a sample statistic to estimate a population value, with a range that shows how precise the estimate is.
2. **Hypothesis testing:** assessing whether a pattern in the sample is likely to reflect a real pattern in the population, or would be unsurprising even if no such pattern existed there.

The second task, asking whether a pattern in the sample would be surprising if there were no such pattern in the population, is the question of **statistical significance**, and it is how inferential statistics are used to test hypotheses.

Consider an illustrative survey in which 100 people in a city are asked about their favorite type of food and 70 of them choose pizza. The figure of 70% describes the sample. Inferential statistics allow the analysis to go further in two ways. First, the share of the whole city (the population) that prefers pizza can be estimated, with a margin of uncertainty around it. Second, a test can assess whether pizza's share of the city really differs from one half, or whether the 70% could reflect only which 100 people happened to be asked. Figure 1 applies both tasks to the same survey; the test shown is a chi-square test, described in Section 5.3.

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

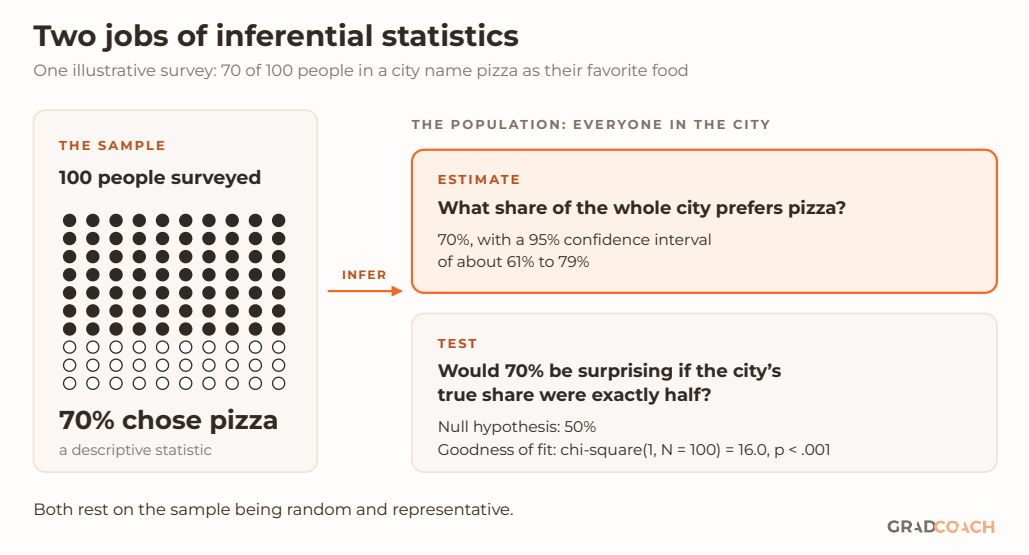



**Figure 1.** One sample, two questions: how large is the share, and would 70% be surprising if it were half?

Both answers depend on one condition. Inferential statistics work best when the sample is **drawn at random** from the population and represents it (Purdue Online Writing Lab, n.d.). If the 100 people were surveyed outside a pizzeria, no test could rescue a conclusion about the city. The sampling method therefore matters as much as the statistics that follow it.

## 3. Inferential and Descriptive Statistics

**Descriptive statistics** are the other branch of statistics. They summarize the data that were collected (the sample), whereas inferential statistics use that sample to say something about the wider population, and potentially to make predictions about it.

Consider a study of shoe brand preferences among men and women. Reporting the proportion of each group in the sample that prefers each brand requires only descriptive statistics. Assessing whether those proportions differ between men and women in the wider population, beyond what sampling variation alone would be likely to produce, requires inferential statistics. Table 1 compares the two branches across five areas.

**Table 1.** Descriptive and inferential statistics compared.

| | Descriptive statistics | Inferential statistics |
|---|---|---|
| What they cover | Only the people or things measured | The wider population from which the sample came |
| The question they answer | What do the data look like? | What do the data suggest beyond the sample? |
| Typical outputs | Means, medians, standard deviations, percentages | Confidence intervals, test statistics, *p*-values |
| Uncertainty | None from sampling | Built in, because a different sample would give different numbers |
| Pizza survey example | 70% of the 100 people chose pizza | A 95% confidence interval of about 61% to 79% for the whole city |

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

Many studies need both, depending on their aims. In a results chapter or section, **descriptive statistics come first**. They also indicate which inferential tests the data can support, because many tests assume that the data have a particular shape.

## 4. How Inference Works

### 4.1 Sampling error and the central limit theorem

Every inference starts from one fact: **a different sample would give a different answer**. A second survey of 100 people in the same city might find 66% or 73% rather than 70%. A value calculated from a sample is called a *statistic*, and the true value in the population is called a *parameter*. The part of the difference between the two that arises from sampling itself is called **sampling error**. It occurs even with a perfectly random sample, because no sample represents its population exactly, and it shrinks as the sample grows (Illowsky & Dean, 2023, Sections 1.1–1.2).

Inference is possible because sampling error behaves predictably. If samples of the same size were drawn repeatedly and the mean of each calculated, those means would cluster around the population mean and would tend to follow a normal distribution, more closely as the sample size increases. This pattern is the **central limit theorem**. It holds for almost any shape of the original distribution (strictly, any with a finite variance), provided the samples are large enough, and so applies even when the data are skewed, that is, bunched at one end with a long tail of less common values toward the other (Illowsky & Dean, 2023, Section 7.1).

Figure 2 illustrates the theorem with waiting times, which are strongly skewed: most waits are short and a few are long.

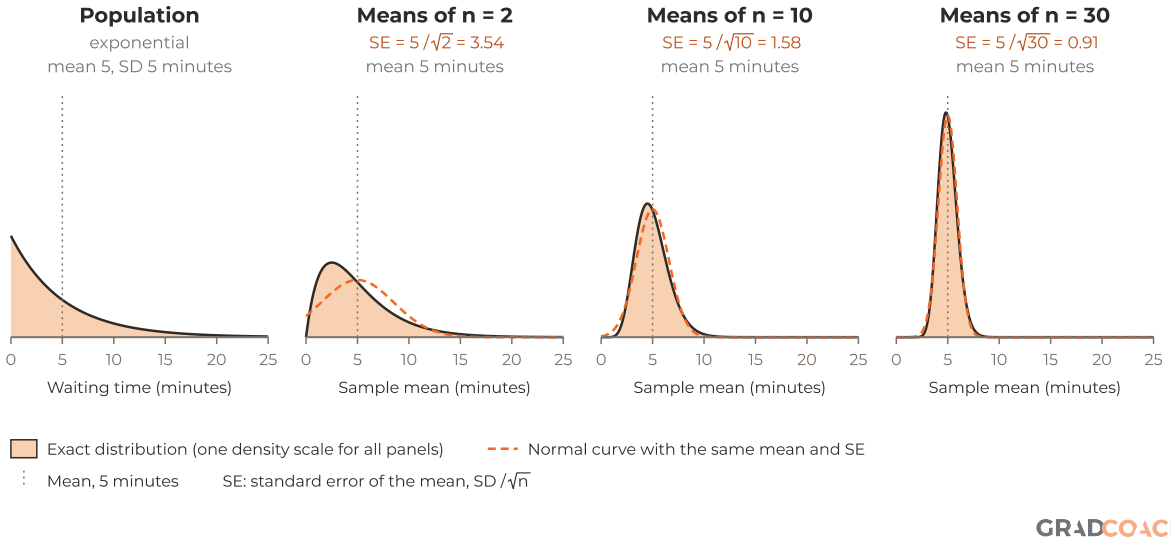



**Figure 2.** As the sample grows, sample means narrow into a bell shape. From Jansen (2026a), CC BY 4.0.

Two features of the figure matter. Every panel has the same mean, 5 minutes, and the spread of the sample means shrinks as the sample grows, from 3.54 minutes with samples of two to 0.91 minutes with samples of 30. That spread is the **standard error of the mean**: the population standard deviation divided by the square root of the sample size (Illowsky & Dean, 2023, Section 7.1).

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

In practice the population standard deviation is unknown, so the sample standard deviation is used instead. Almost every confidence interval and test described below is built on a standard error.

### 4.2 Estimation and confidence intervals

The first task of inferential statistics is estimation. The sample statistic (70% chose pizza) is the best single estimate of the population value, and a **confidence interval** places a range around it that reflects sampling error. For the pizza survey, a 95% confidence interval, calculated with the normal approximation (which treats the sampling distribution of the sample percentage as normal; Illowsky & Dean, 2023, Section 8.3), runs from about 61% to 79%.

This range is frequently misread. It does not mean that there is a 95% probability that the true figure lies between 61% and 79%. A widely cited guide to statistical misinterpretations explains that a particular interval either contains the true value or does not, and states what the 95% refers to:

> The frequency with which an observed interval (e.g., 0.72–2.88) contains the true effect is either 100 % if the true effect is within the interval or 0 % if not; the 95 % refers only to how often 95 % confidence intervals computed from very many studies would contain the true size if all the assumptions used to compute the intervals were correct.
>
> Greenland et al. (2016, p. 343)

The 95%, in other words, describes the method, not any particular interval. What a confidence interval does show, at a glance, is **precision**. A narrow interval indicates a precise estimate; a wide interval indicates that the sample left considerable room for doubt.

### 4.3 Hypothesis testing and *p*-values

The second task is testing. A test starts from a **null hypothesis**, the claim against which the test looks for evidence. Most often the null hypothesis states that there is no effect, no difference or no relationship in the population, but it can also specify another value of interest (Greenland et al., 2016). For the pizza survey, a natural benchmark is one half, the line between a majority favorite and a minority one, so the null hypothesis is that pizza is the favorite of exactly 50% of the city.

The test then produces a ***p*-value**: the probability of obtaining a test statistic at least as extreme as the one observed if the null hypothesis and every other assumption of the test were correct (Greenland et al., 2016). It measures **how surprising the result would be** in a world in which the null hypothesis and the test's assumptions hold, and the smaller the $p$-value, the more surprising the result. Obtaining 70 out of 100 when 50 would be expected is very surprising ($p < .001$). This $p$-value is well below the conventional cut-off described next, so the null hypothesis is rejected.

A result is usually called **statistically significant** when $p$ falls at or below a cut-off, most often .05 (Greenland et al., 2016, p. 339). That threshold is a convention, not a law. Two misreadings of the $p$-value are especially common.

The first is **mistaking significance for importance**. The American Statistical Association's statement on $p$-values sets out, as one of its six principles, that neither a $p$-value nor statistical significance measures the size of an effect or the importance of a result (Wasserstein & Lazar, 2016). For

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

this reason an effect size should be reported alongside every *p*-value.

The second is **reading a non-significant result as proof of no effect**. Greenland et al. (2016) list this reading among the common fallacies, observing that even when a *p*-value is large, many hypotheses other than the null remain highly consistent with the data, "so that a definitive conclusion of 'no association' cannot be deduced from a P value, no matter how large" (p. 341). A study can simply be too small to detect a real effect, which is why statistical power, the pre-study probability that a test will give a significant result if the true effect is of a given size, should be checked before data are collected (Greenland et al., 2016, p. 345).

A study therefore usually reports both a test and an interval, because they answer different questions. The test indicates whether there is evidence of an effect; the confidence interval and the effect size indicate how large that effect plausibly is. A results section that reports only *p*-values has answered the less informative of the two questions.

## 5. Common Inferential Tests

There are many inferential tests, and the appropriate one depends on the research question and the data. The five described below are widely used and common in student research. Broadly, the *t*-test and analysis of variance compare group means, the chi-square test examines counts in categories, correlation examines the relationship between two numerical variables, and regression predicts one variable from one or more others. The list is not complete, and each test carries assumptions that the data must meet before its result can be trusted.

### 5.1 *t*-tests

A *t*-test **compares two means**, most often those of two groups, to assess whether the difference between them is larger than sampling variation alone would be likely to produce if the groups did not differ in the population. This matters because two means can differ by a few points simply because of who happened to fall into each group. The *t*-test weighs the gap between the means against how much the scores vary within each group and how many scores each group contains, so the same difference produces a larger test statistic when the scores are closely grouped and the groups are large.

A *t*-test could be used, for example, to assess whether the exam scores of two mathematics classes taught by different teachers differ significantly. A significant result indicates that a gap this large would be unlikely if the two classes did not differ in the population. Whether the teacher caused the difference is a separate question that depends on the research design: the classes might also differ in prior ability, timetable or anything else the study did not control. Figure 3 shows the example with simulated scores. The number in parentheses after *t* is the degrees of freedom; for an independent-samples *t*-test that assumes equal variances, it is the total number of participants minus two (Kim, 2015), 48 for two classes of 25.

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

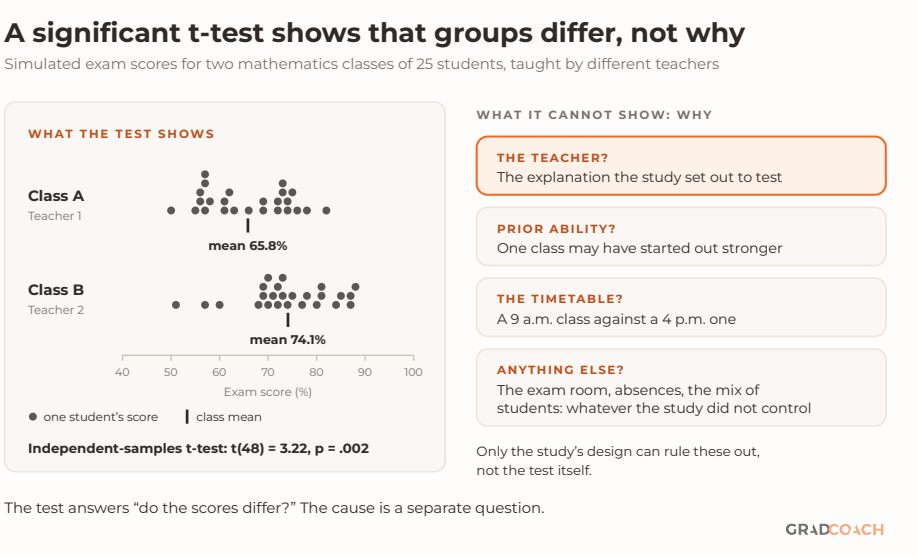



**Figure 3.** A significant *t*-test shows that the classes differ; only the design can show why.

Three forms of the t-test are in common use. The **one-sample t-test** compares one group's mean with a fixed value, such as a national average. The **independent-samples t-test** compares two separate groups, such as the two classes. The **paired-samples t-test** compares the same people measured twice, such as before and after a training course. Kim (2015) divides t-tests into independent and paired types and classifies the paired test as a form of the one-sample test.

The following extract, from a clinical psychology doctorate at the University of Lincoln, gives a clear paired example. Its 418 participants were UK professionals who are exposed to accounts of other people's trauma at work, among them doctors, paramedics, psychologists, nurses and counselors. Of these, 318 completed a measure of vicarious trauma (VT), the Trent Measure of Vicarious Trauma (TMVT), twice, about two weeks apart (T1 and T2 in the extract, where M is the mean and SD the standard deviation):

> A paired-samples t-test further examined whether group average VT levels changed over time. Mean TMVT scores did not significantly differ between T1 (M=36.64, SD=14.44) and T2 (M=35.77, SD=15.68), t(317)=0.78, p =.437. The mean difference (0.87) was small, with a negligible effect size (Cohen's d=.04).
>
> Tunstall (2026, pp. 82–83)

The result is not significant ($p$ = .437), and the passage goes further. It reports the **size of the difference** (0.87 points) and an effect size, Cohen's $d$, which expresses the difference in standard deviations (0.04, far below the 0.2 conventionally treated as a small effect; Sullivan & Feinn, 2012). Together, the mean difference and the effect size indicate that the group average changed little over the two weeks. A 95% confidence interval for the mean difference (T1 minus T2) can be derived from the reported t and mean difference. It runs from about −1.3 to 3.1 points. Since the scale's standard deviation is about 15, the interval excludes changes larger than about a fifth of the scale's standard deviation in either direction. A non-significant result is best reported in the same way, with the statistic, the $p$-value and the effect size, rather than with a simple statement that no difference was found.

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

### 5.2 Analysis of variance

Whereas a t-test compares the means of two groups, an **analysis of variance (ANOVA)** compares the means of three or more groups at once. An ANOVA also assesses whether the differences between the means are larger than sampling variation alone would be likely to produce if the groups did not differ in the population. An ANOVA could be used, for example, to compare the average standardized test scores of students at three types of school (public, private and home schooling), or the average sales of a product across several stores.

Table 2 shows the kind of table that statistical software such as SPSS produces for a one-way ANOVA, here for 30 people divided into three groups (the data are illustrative).

**Table 2.** One-way ANOVA for three groups of 10 (illustrative data, SPSS layout).

| Source | Sum of squares | *df* | Mean square | *F* | Sig. |
|---|---|---|---|---|---|
| Between groups | 339.800 | 2 | 169.900 | 4.545 | .020 |
| Within groups | 1009.400 | 27 | 37.385 | | |
| Total | 1349.200 | 29 | | | |

Each mean square is a sum of squares divided by its degrees of freedom (*df*), and the F value is the ratio of the two mean squares (169.900 divided by 37.385). It compares how much the group means vary with how much individuals vary within their groups (Kim, 2014). The "Sig." column is the *p*-value. In APA style the result is reported as $F(2, 27) = 4.54$, $p = .020$. The exact ratio is 4.5446: the software displays it to three decimals (4.545), and the APA-style report to two. Because .020 is below .05, there is **evidence that at least one group mean differs** from at least one other.

An effect size should be added. **Eta squared** ($\eta^2$) is the between-groups sum of squares divided by the total sum of squares (Lakens, 2013): 339.8 / 1349.2 = .25, so about a quarter of the variation in scores lies between the groups.

What the ANOVA does not show is **which groups differ**. A significant F is therefore followed by post hoc tests that compare the groups in pairs (Kim, 2014). The example above is a **one-way ANOVA**, which has a single grouping variable. Other designs call for other forms, such as the repeated-measures ANOVA when the same people are measured three or more times, so the form should be matched to the study before it is run.

### 5.3 Chi-square tests

The **chi-square test works with categories** rather than means. In its most common form, the chi-square test of independence, it assesses whether two categorical variables are related, such as sex and handedness. A second form, the **goodness-of-fit test**, assesses whether the counts for one variable match an expected pattern. That is the test behind the pizza survey, which compares the observed 70 with the 50 expected under the null hypothesis that pizza is the favorite of exactly half the city.

A chi-square test of independence starts from a contingency table (a cross-tabulation) of counts. Figure 4 shows one for sex and handedness in 100 people.

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

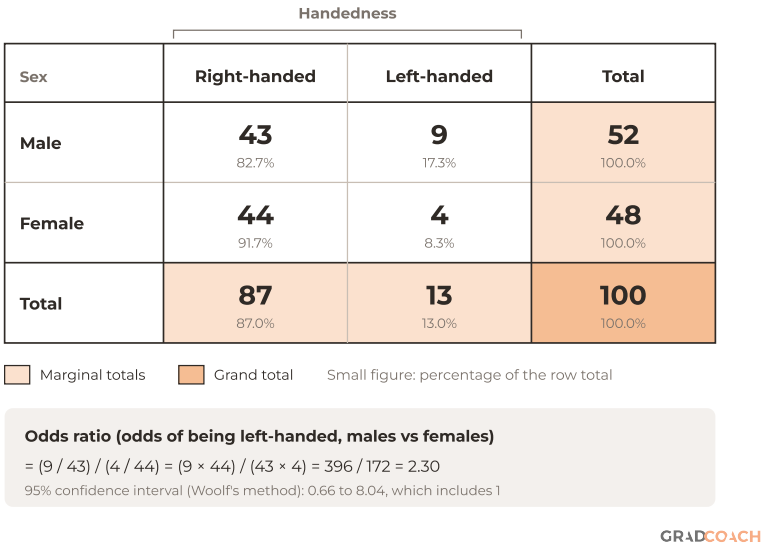



**Figure 4.** A 2 × 2 contingency table with row percentages and an odds ratio. From Jansen (2026b), CC BY 4.0.

The row percentages allow the two groups to be compared fairly, even though they differ in size. Left-handedness is about twice as common among the men in this sample (17.3% against 8.3%), but with only 13 left-handed people in the sample, a gap of this size could easily arise by chance even if handedness did not differ by sex. A chi-square test of independence on these counts agrees: $\chi^2(1, N = 100) = 1.78$, $p = .18$ (Pearson's version, without a continuity correction), so the difference is not statistically significant. Every expected count (the count a cell would hold if handedness did not differ by sex) is above 5, which meets the test's usual requirement (McHugh, 2013).

The figure examines the same gap a second way, through the **odds ratio**. This is the odds of being left-handed among the men (left-handed men divided by right-handed men) divided by the odds of being left-handed among the women (Szumilas, 2010). Here the odds ratio is 2.30, with a 95% confidence interval of 0.66 to 8.04. An odds ratio of 1 means equal odds, and because the interval includes 1, the two approaches agree: the sample gives **no clear evidence of a difference**.

The following extract is from a Doctor of Nursing Practice project at the University of Massachusetts Amherst, which tested whether the share of patient records showing a health care proxy (HCP), an advance directive that names a person to make medical decisions for a patient who cannot communicate their wishes, rose after a nurse-led intervention on a hospital unit. Its 687 records with usable proxy data form a 2 × 2 table (before or after the intervention, proxy present or absent). Each record either did or did not show a proxy (nominal data), so the analysis is a test of independence:

> To determine if the increase in HCP rates in the post-intervention group was significant, Chi-square analysis was used to compare expected frequencies. The observed post-intervention HCP completion rate was significantly greater than expected, $\chi^2(1, N = 687) = 4.127$, $p = 0.042$.
>
> Baldassari (2026, p. 18)

The project's own table gives the counts behind the result: 69.0% of the 403 pre-intervention records with usable proxy data showed a proxy, against 76.1% of the 284 post-intervention records (Baldassari, 2026, p. 19). Because the two periods involve different patients, a before-and-after comparison of this kind cannot, on its own, show that the intervention caused the rise.

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

The result follows a standard order: **the test, its degrees of freedom and sample size in parentheses, the statistic and the *p*-value**, after a sentence stating what was tested. Two additions strengthen such a report. First, the percentages in the text let the reader see the size of the change, not only that it is significant. Second, because the chi-square statistic indicates significance rather than strength, it should be followed by a measure of the strength of the association (McHugh, 2013), such as phi (ϕ). Here phi is about .08, a very weak association that falls below the .10 that Cohen (1992, p. 157) gave as a small effect for a chi-square test (for a 2 × 2 table, Cohen's index $w$ equals phi).

### 5.4 Correlation

Correlation analysis examines the **relationship between two numerical variables** to assess whether they move together. The result is a correlation coefficient, usually Pearson's $r$, which ranges from −1 to +1 (Schober et al., 2018). A value close to +1 means that the two variables tend to rise together, a value close to −1 means that one tends to fall as the other rises, and a value close to zero means that there is no clear straight-line relationship.

For example, the more hours students spend studying, the higher their exam scores may tend to be (a positive correlation), and the more hours they spend in paid work, the lower their scores may tend to be (a negative correlation). Figure 5 shows both patterns, and the absence of a relationship, in simulated data.

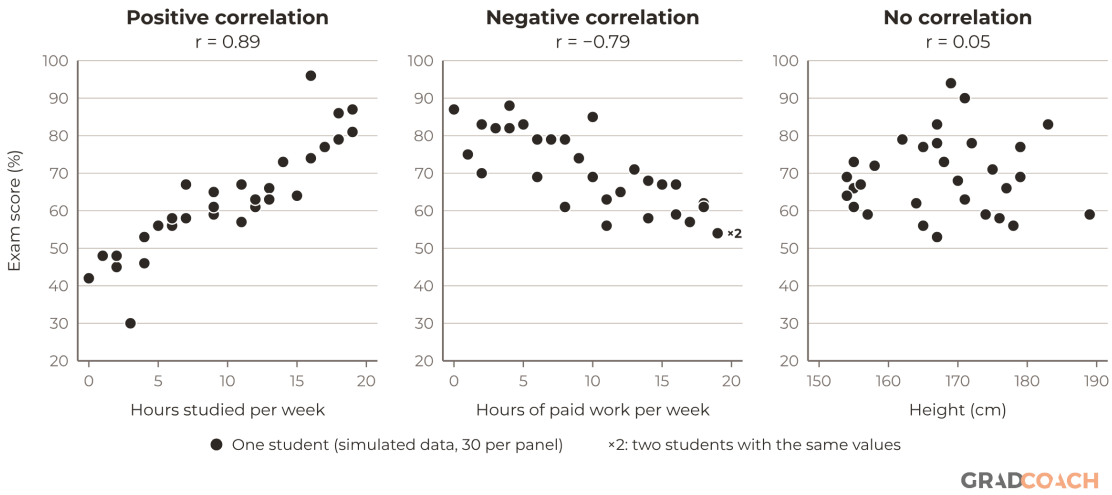



**Figure 5.** Positive, negative and no correlation (simulated data). From Jansen (2026d), CC BY 4.0.

The closer the points cluster around a straight line, the further $r$ moves from zero. One caution applies to the right-hand panel: a coefficient near zero indicates only the absence of a straight-line (linear) relationship (Schober et al., 2018); a strong curved relationship can still produce a coefficient near zero.

On its own, $r$ describes only the sample. It becomes inferential when a test assesses whether the relationship holds in the population, which is where the accompanying $p$-value comes from. Both the correlation and its $p$-value appear in the following extract from a PhD dissertation from

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

Purdue University's Department of Agricultural Sciences Education and Communication. The study surveyed undergraduates about their teachers' practices and about their confidence in their own academic ability (academic self-efficacy):

> As shown in Table 11, there was a moderate and significant positive relationship between students' perceptions of diverse teaching practices and students' academic self-efficacy ($r = .47$, $n = 921$, $p < 0.001$). Coefficient of determination $R^2 = .22$ indicated that 22% of the variance in student' academic self-efficacy was explained by students' perceptions of diverse teaching practices.
>
> Huang (2019, p. 95)

The passage names the strength of the relationship in words ("moderate"), gives $r$, the sample size and the $p$-value, and then squares $r$ ($.47 \times .47 \approx .22$) to state **how much of the variation the relationship accounts for**. That last step is important: with 921 students, a correlation as weak as .065 would already reach $p < .05$, so the $p$-value alone says little about how strong the relationship is. The word "explained" needs care, however: it is statistical shorthand for shared variance, not a claim that one variable causes the other.

Correlation can show how two variables are related, but **it does not show that one causes the other**. Figure 6 sets out three explanations that fit the same coefficient of .47.

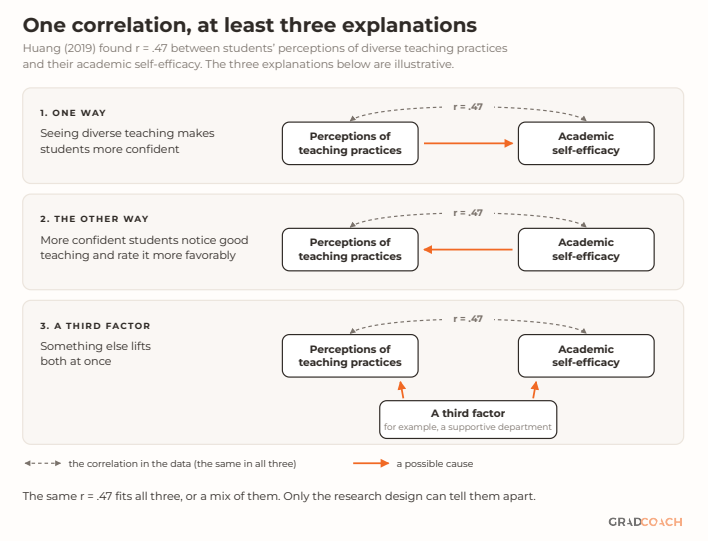



**Figure 6.** One correlation fits several causal explanations; the data cannot choose between them.

## 5.5 Regression

Regression goes a step beyond correlation: it **predicts the value of one variable** (the dependent variable) from the values of one or more others (the independent variables). Regression analysis could be used, for example, to predict house prices from the number of bedrooms, the location and the age of the house. The analysis produces an equation into which those factors can be entered to estimate a house's price. Similarly, a person's weight could be predicted from their height, age and daily calorie intake.

These examples describe **multiple regression**, because each has more than one independent variable. There are many other types, including simple linear regression, with a single predictor, and logistic regression, for an outcome with two categories such as pass or fail. The kind of outcome

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

variable largely determines which type is appropriate.

Before a regression is trusted, its **residuals** should be checked. A residual is the gap between an observed value and the value the model predicts. Residuals are plotted against the fitted values, the model's predictions (Kim, 2019). A straight-line model that fits the data leaves a random, even band of residuals around zero. A curved pattern indicates that the relationship is not a straight line, and a funnel shape, in which the spread of the residuals grows with the fitted value, indicates unequal variance, or *heteroscedasticity* (Kim, 2019). Figure 7 shows the three patterns in simulated data.

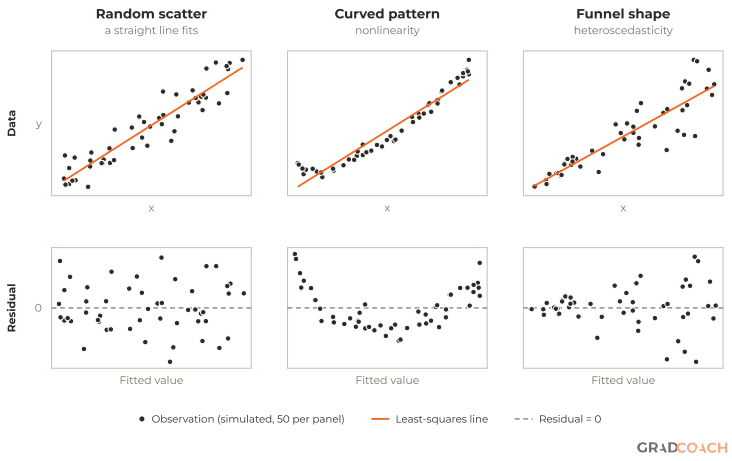



**Figure 7.** Residual plots: random scatter, a curve and a funnel. From Jansen (2026c), CC BY 4.0.

A curved pattern in the residuals means that the straight-line model has missed the shape of the relationship. A funnel violates the equal-variance assumption (Kim, 2019); the standard errors, and the significance tests and confidence intervals built on them, may then be misleading.

As with correlation, **regression alone does not establish causation**. Factors that were not included in the model could be responsible for the results, and a causal claim requires a design that deals with them, such as a randomized experiment, in which random assignment balances such factors between groups on average.

## 6. Conclusion

Inferential statistics use a sample to draw conclusions about a population, and they are only as trustworthy as the sample: they work best when it is random and representative. Estimation gives a best estimate of a population value with a confidence interval that shows its precision; hypothesis testing gives a test statistic and a *p*-value that indicate how surprising the pattern would be if the null hypothesis and the test's assumptions were correct. Of the five tests described here, the t-test and ANOVA compare group means, the chi-square test works with categories, and correlation and regression examine relationships.

A significant result is not proof of a large effect or of a cause, and a non-significant one is not proof of no effect. Each result is therefore best reported in full: the test, its degrees of freedom, the statistic, the *p*-value and an effect size, then a plain statement of what it shows. A causal claim rests on the design, not the statistic.

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

## References

Baldassari, J. (2026). *Improving health care proxy procurement rates for hospitalized adults: A nurse led intervention* [Doctor of Nursing Practice project, University of Massachusetts Amherst]. https://doi.org/10.7275/10651

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

Huang, Y. (2019). *Undergraduate students' perceptions of culturally responsive teaching and their sense of belonging and academic self-efficacy in higher education* [Doctoral dissertation, Purdue University]. https://doi.org/10.25394/PGS.9116885

Illowsky, B., & Dean, S. (2023). *Introductory statistics* (2nd ed.). OpenStax. https://openstax.org/details/books/introductory-statistics-2e

Jansen, D. (2026a). *Central limit theorem with an exponential population* [Figure]. Grad Coach. https://gradcoach.com/open-educational-resources/central-limit-theorem-example/

Jansen, D. (2026b). *Contingency table 2x2 example with marginal totals and odds ratio* [Figure]. Grad Coach. https://gradcoach.com/open-educational-resources/contingency-table-example/

Jansen, D. (2026c). *Residual plots showing random scatter, curvature and heteroscedasticity* [Figure]. Grad Coach. https://gradcoach.com/open-educational-resources/residual-plot-examples/

Jansen, D. (2026d). *Scatter plots showing positive, negative and no correlation* [Figure]. Grad Coach. https://gradcoach.com/open-educational-resources/scatter-plot-examples/

Kim, H.-Y. (2014). Analysis of variance (ANOVA) comparing means of more than two groups. *Restorative Dentistry & Endodontics*, 39(1), 74–77. https://doi.org/10.5395/rde.2014.39.1.74

Kim, H.-Y. (2019). Statistical notes for clinical researchers: Simple linear regression 3 – residual analysis. *Restorative Dentistry & Endodontics*, 44(1), Article e11. https://doi.org/10.5395/rde.2019.44.e11

Kim, T. K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6), 540–546. https://doi.org/10.4097/kjae.2015.68.6.540

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

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia Medica*, 23(2), 143–149. https://doi.org/10.11613/BM.2013.018

Purdue Online Writing Lab. (n.d.). *Basic inferential statistics: Theory and application*. Purdue University. Retrieved November 12, 2026, from https://owl.purdue.edu/owl/research_and_citation/using_research/writing_with_statistics/basic_inferential_statistics.html

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5), 1763–1768. https://doi.org/10.1213/ANE.0000000000002864

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3), 227–229. https://pmc.ncbi.nlm.nih.gov/articles/PMC2938757/

Tunstall, C. M. (2026). *Confirmatory factor analysis and further validation of the Trent Measure of Vicarious Trauma (TMVT)* [Doctoral thesis, University of Lincoln]. https://doi.org/10.24385/lincoln.33113120

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA 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