SAINT MARY'S SCHOOL • MEDFORD, OREGON

AP Statistics

Student Reference

Definitions, explanations, worked examples, and connections for the full course. Use this book to recover the reasoning behind a method, then practice in the companion workbook.



01 02 03 04 05 06 07 08 09 10

Materials customized by **Jason Javarone** for use in the St. Mary's School AP Statistics course.

Source-based classroom edition | November 11, 2026

Ten SMS units follow the five-unit AP Statistics Course and Exam Description, effective Fall 2026.

About this edition

For each statistical conclusion, name the group and the quantity, give numerical evidence with units when appropriate, and explain what the evidence supports. Read your answer with the question covered. Someone who has never seen the prompt should still understand your claim.

Match the lesson ID across books. For example, S01.01 in the Student Reference pairs with S01.01 in the Workbook. The homework page gives the next reading number.

The primary instructional source is Introductory Statistics 2e by Barbara Illowsky and Susan Dean, OpenStax, Rice University (supplied 2026 printing of the second edition). The explanations and credited worked examples adapt this textbook to the AP course. Source notes identify the original sections and problems and any changes to context, assumptions, or calculations. Local applications and class-data investigations provide additional practice.

This adaptation is distributed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0), consistent with the supplied OpenStax edition. The original OpenStax textbook is available in full for free at OpenStax.org: https://openstax.org/details/books/introductory-statistics-2e

AP® and Advanced Placement® are trademarks of the College Board, which is not affiliated with and does not endorse this package. College Board CED content is used as an alignment reference; official exam questions are not reproduced. Worked data are illustrative unless a source is named or students collect the observations themselves.

Local artwork was created for this SMS edition. Chapel architecture follows the supplied Ron Grimes/Architects reference; the historical academy follows a supplied archival photograph credited to the Southern Oregon Historical Society. Two approved Table Rock landscapes and the supplied SMS logo are reused unchanged. Captions distinguish local settings from observed data. Original reference art and the school logo retain their original ownership.

Print US Letter, actual size, long-edge duplex. The inner margin is 1 inch; the outer margin is 0.75 inch. Color assists navigation, but labels carry the meaning. Fonts are embedded Caladea and Carlito. Keep teacher solutions separate from student copies.

Contents

About this edition 2

Visual guide and navigation 5

Unit 01 | Analyzing Data 6

S01.01 | Questions, Variables, and Categorical Data . . . 7
S01.02 | Two Categorical Variables and Association . . . 10
S01.03 | Seeing a Quantitative Distribution . . . 14
S01.04 | Center, Variability, and Resistance . . . 17
S01.05 | Quartiles, Outliers, and Group Comparisons . . . 21
S01.06 | Relative Standing and Changing Units . . . 24

Unit 02 | Collecting Data 28

S02.01 | Choosing a Random Sample . . . 29
S02.02 | Bias, Observational Studies, and Ethical Questions . . . 32
S02.03 | Designing a Completely Randomized Experiment . . . 37
S02.04 | Blocking, Matched Pairs, and Scope of Inference . . . 41

Unit 03 | Probability 46

S03.01 | Chance models and reproducible simulation . . . 47
S03.02 | Events, complements, and overlapping categories . . . 51
S03.03 | Conditional probability and independence . . . 55
S03.04 | Multi-stage chance processes and probability trees . . . 58

Unit 04 | Random Variables 63

S04.01 | Random variables, distributions, and expected value . . . 64
S04.02 | The binomial model . . . 68
S04.03 | Continuous models and the normal distribution . . . 71
S04.04 | Normal probabilities, percentiles, and cutoffs . . . 75

Unit 05 | Sampling Distributions 80

S05.01 | Sampling distributions and randomization distributions . . . 81
S05.02 | Sampling distributions of proportions and their differences . . . 84
S05.03 | Sample means, differences in means, and the CLT . . . 89

Unit 06 | Confidence Intervals for Proportions 94

S06.01 | What a confidence interval says . . . 95

S06.02 | Build and justify a one-proportion interval . . . 99
S06.03 | Plan precision before collecting data . . . 103
S06.04 | Estimate a difference in population proportions . . . 106

Unit 07 | Hypothesis Tests for Proportions 111
S07.01 | Hypotheses, null models, and p-values . . . 112
S07.02 | Carry out a one-proportion z test . . . 116
S07.03 | Errors, power, and consequences . . . 120
S07.04 | Test a difference in population proportions . . . 124
S07.05 | Compare categorical distributions with chi-square . . . 128
S07.06 | Investigate association with chi-square . . . 133

Unit 08 | Confidence Intervals for Means 138
S08.01 | Estimating a population mean . . . 139
S08.02 | Confidence, precision, and defensible claims . . . 143
S08.03 | Paired observations: analyze the differences . . . 147
S08.04 | Estimating a difference between independent means . . . 151

Unit 09 | Hypothesis Tests for Means 156
S09.01 | Testing a claim about one population mean . . . 157
S09.02 | Testing a population mean difference from paired data . . . 161
S09.03 | Testing a difference between independent means . . . 165
S09.04 | Choosing inference and linking tests with intervals . . . 169

Unit 10 | Two-Variable Data and Regression 174
S10.01 | Seeing relationships in paired quantitative data . . . 175
S10.02 | Correlation: strength of a linear association . . . 179
S10.03 | Linear predictions and their limits . . . 183
S10.04 | Residuals and checking a linear model . . . 187
S10.05 | Least squares and explained variation . . . 191

Visual reference 196

Choosing an inference method 199

Responding to AP questions 200

Using technology and writing digitally 201

Notation and vocabulary 202

Visual guide and navigation

Unit numbers, titles, colors, and page-edge tabs identify your place in the course. Lesson IDs match across the reference, workbook, and teacher guide. Read every graph with its labels, units, and caption; color is a second cue, not the only cue.

01 02 03 04 05 06 07 08 09 10

KEY IDEA

A definition or principle to retain and use in a new situation.

WATCH FOR

A common error, a condition that needs checking, or a conclusion that the evidence does not support.

CONCEPT → WORKED EXAMPLE → YOUR INVESTIGATION

Instructional graphics make relationships visible: compare groups, follow a probability path, locate a statistic in a distribution, or connect a calculation to its meaning. Workbook organizers leave the reasoning for you to complete. Occasional context illustrations introduce a setting; they are not measured data.

A good statistical graphic helps answer a question. Identify what is shown, describe the pattern with evidence, and state what the display cannot establish.

How to use the reference

Before class, complete the assigned preview reading and its two checks. After the entry ticket, follow the explanation and worked example in this book. Then turn to the workbook, try the first step yourself, and use the class discussion to check your reasoning. On a lesson marked Opening prediction, start with the brief workbook question. Return to the reference when you need to check a definition or example.

KEY IDEA

A statistical answer connects a question, the data-collection process, an analysis, and a contextual conclusion. The strength of the conclusion depends on every link.

Unit 01 | Analyzing Data

Statistics begins with a question, a defined population, and measurements that can answer the question. This unit develops the habit of examining a distribution before selecting a numerical summary. You will distinguish categorical and quantitative variables, compare groups with appropriate denominators, and connect every calculation to the original context. Displays and summaries can reveal patterns, but they can also conceal important features. You will learn to identify those limits and to explain what the data support. The local examples use illustrative data; they are not claims about actual Saint Mary's students or Southern Oregon residents.

CED topics: 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.1, 2.2

IN THIS UNIT

Lesson	Focus
S01.01	Questions, Variables, and Categorical Data
S01.02	Two Categorical Variables and Association
S01.03	Seeing a Quantitative Distribution
S01.04	Center, Variability, and Resistance
S01.05	Quartiles, Outliers, and Group Comparisons
S01.06	Relative Standing and Changing Units

Two rows from an illustrative arrival log

Student ID	Main travel mode	Travel time (min)
S01	Bus	18
S02	Walk	12

Each row records one student's arrival on the selected morning.

LOCAL QUESTION

At SMS, what would you record to describe how students arrive at school? Define the population, observational unit, and variables before collecting data.

S01.01 | Questions, Variables, and Categorical Data

CED 1.1, 1.2, 1.3, 1.4

- Write an investigative question with a defined population and measurable variable.
- Distinguish observational unit, population, sample, parameter, and statistic.
- Classify variables and construct categorical frequency and relative-frequency displays.

CONCEPT AND INTERPRETATION

Start with a question

How common is bus travel to school? To make that question precise, ask, “What proportion of our current students usually arrive by bus?” The population is the group we want to learn about: all students enrolled here now. The variable is the characteristic recorded for each student: usual arrival method. Students can give different answers, and data can help estimate the proportion. “Is transportation good?” leaves the group, the measurement, and even the meaning of “good” unspecified. Set the question before examining the results. If a pattern suggests a new question, call it exploratory and look for fresh evidence.

Identify the units and the target

An observational unit is the person or object on which measurements are recorded. The population includes all observational units relevant to the question; its size is N. The sample is the subset actually observed; its size is n. A census attempts to obtain information from every population member. One student’s recorded arrival method is a datum. Together the records form the data set. A parameter summarizes the population, and a statistic summarizes the sample. A sample percentage is known once the data are collected, but the population percentage may remain unknown. A large sample is not automatically representative. The selection process matters, a point developed in Unit 2. The distinction between a parameter and a statistic depends on which group is summarized, not on whether a number is large or known. In the uniform-spending example, changing the sample size would change which families contribute to the statistic but would not change the population parameter being estimated.

Classify the variable by meaning

Each student is an observational unit. Usual arrival method is a categorical variable because its values are labels such as bus, car, walk, and bicycle. A quantitative variable measures or counts a numerical amount for which arithmetic has a meaningful interpretation. A numerical code such as a jersey number is categorical, because averaging the codes does not measure a useful amount. A discrete quantitative variable has countable possible values, such as number of absences. A continuous quantitative variable can take any value in an interval, such as travel time. Rounding measured time to whole minutes does not change its underlying continuous nature. Photographs, text, and sound can also supply data when a consistent coding rule is defined.

01

Make the denominator visible

A frequency table records category counts. Divide each count by the total number of observations to obtain relative frequency. Multiply by 100 to express a percentage. When each observation belongs to exactly one of the listed categories, the counts add to the number of observations and the proportions add to 1, apart from rounding. Missing responses must be reported rather than quietly removed. If respondents may select several choices, percentages can sum above 100%; that is a different design and should be labeled clearly. State whether a percentage uses all selected people, all respondents, or only a particular subgroup.

Choose and critique a display

A bar chart displays categorical counts or proportions with separate bars and labeled axes. Begin a bar-length scale at zero so the lengths preserve the comparison. A pie chart represents the whole with sectors whose angles are the category proportions times 360 degrees. Pie charts require nonoverlapping categories that account for the whole. Avoid three-dimensional decoration, inconsistent widths, and enlarged pictures whose areas exaggerate differences. A useful written conclusion identifies the category, gives supporting numbers, and limits the claim to the observed group unless the sampling design supports broader inference.

KEY IDEA

Name the population, variable, and denominator before interpreting a number.

WATCH FOR

A number used as a label is not automatically a quantitative variable.

OPENSTAX WORKED EXAMPLE

School-uniform spending: population and sample

A school wants to estimate the mean amount its families spend on school uniforms in one year. It randomly surveys 100 families with children enrolled at the school. Three reported amounts are $65, $75, and $95. Identify the population, sample, variable, population parameter, sample statistic, and data. Explain whether the three amounts are enough to calculate the sample mean.

1. The observational unit is one family. The population consists of all families with children enrolled at the school during the year of interest. The sample consists of the 100 families selected and surveyed.

2. The variable is the amount one family spends on school uniforms during that year, measured in dollars. It is quantitative. The population parameter is the mean annual uniform expenditure of all enrolled families.

3. The sample statistic is the mean annual expenditure reported by the 100 sampled families. The values $65, $75, and $95 are three individual observations, not three statistics.

4. Only three of the 100 expenditures are given. Their mean would describe those three observations; it is not the mean of the full sample. All 100 amounts are needed to calculate the sample mean.

Conclusion. The survey estimates a population mean using the mean for 100 randomly sampled families. The three listed expenditures illustrate the variable but do not determine the sample mean.

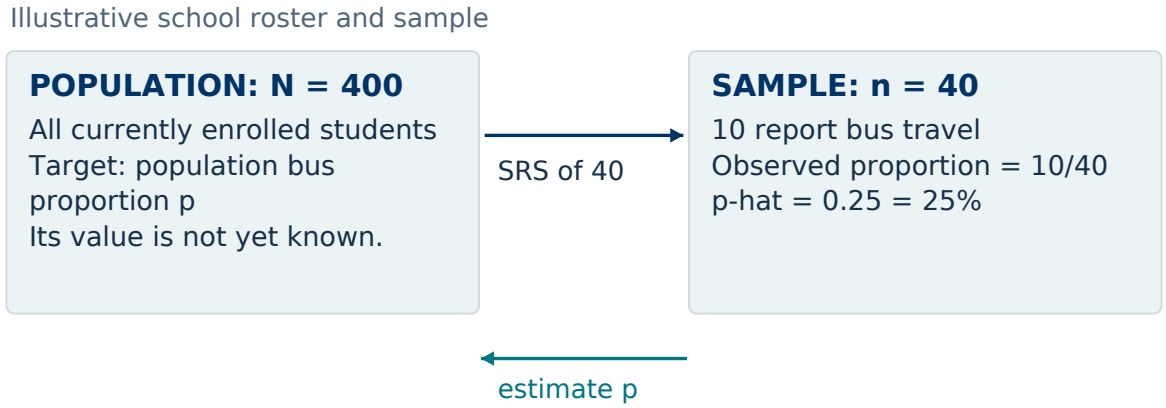
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.1, Try It 1.1, printed p. 8. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

How did the sampled students arrive?

An illustrative simple random sample of 40 students from a school roster of 400 reports usual arrival method: car 22, bus 10, walk 6, bicycle 2. Identify the observational unit, population, sample, and variable. Calculate the relative frequency of each arrival method and interpret the sample bus proportion as an estimate of the population proportion.

From a population question to a sample statistic



One student is a unit; usual arrival method is the categorical variable.

The sample percentage estimates a population proportion; it does not reveal the population value exactly.

The same distribution as counts, percentages, and angles

Arrival method	Count	Relative frequency	Pie-chart angle
Car	22	22/40 = 55%	0.55 × 360° = 198°
Bus	10	10/40 = 25%	90°
Walk	6	6/40 = 15%	54°
Bicycle	2	2/40 = 5%	18°
Total	40	100%	360°

Illustrative worked data. Every relative frequency uses all 40 sampled students.

1. The observational unit is one student. The population is the 400 currently enrolled students; the sample is the 40 selected students. The variable is usual arrival method, which is categorical.

2. The parameter of interest could be the proportion of all 400 students who usually arrive by bus. The corresponding sample statistic is 10/40 = 0.25.

3. The relative frequencies are car 55%, bus 25%, walk 15%, and bicycle 5%. Their sum is 100%, so the categories account for all 40 responses.

4. A relative-frequency bar chart uses those four categories and a vertical axis labeled Percent of sampled students, beginning at zero. A pie chart would use angles 198, 90, 54, and 18 degrees.

5. Car was the most common method in the sample: 22 of 40 students, or 55%, reported it. For the bus question, 10 of 40 students reported riding the bus. The sample estimate for the whole roster is 25%; the actual percentage among all 400 students is unknown.

Conclusion. Ten of the 40 randomly sampled students usually rode the bus, so the estimated bus share for the 400-student roster is 25%. The exact roster percentage is still unknown.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.1, Try It 1.1, printed p. 8. Adapted for this course; source data retained unless otherwise stated.

The school setting is retained without claiming these are actual SMS expenditures. Unlike the inconsistent sample description in the key to Example 1.1, the sample here is exactly the 100 surveyed families.

S01.02 | Two Categorical Variables and Association

CED 2.1, 2.2

- Compute joint, marginal, and conditional relative frequencies.
- Compare groups using appropriate conditional distributions.
- Interpret side-by-side bars, segmented bars, and mosaic plots.

CONCEPT AND INTERPRETATION

One observation, two classifications

A two-way table records the joint distribution of two categorical variables measured on the same observational units. Each interior cell counts units with both the row and column characteristics. Row and column totals are called marginal totals because they appear along the margins. Check that the table includes the same units in every calculation and that each unit is counted once. A table of counts is not yet a table of conditional percentages. Before dividing, translate the question into ordinary language so the reference group is explicit.

Three kinds of relative frequency

A joint relative frequency divides one interior cell by the grand total. It describes the fraction of all observations with both characteristics. A marginal relative frequency divides a row or column total by the grand total and describes one variable without restricting the other. A conditional relative frequency first restricts attention to one category of one variable, then divides the relevant cell count by that restricted total. The wording "among students who work" fixes the denominator as the working-student total. Reversing the condition usually produces a different answer because the denominator changes.

Compare proportions when group sizes differ

A raw count can be larger merely because a group is larger. To compare how a response is distributed across groups, calculate the response percentages within each group using a consistent conditioning direction. The percentages within one complete group should sum to 100%. Two categorical variables are associated in the observed data when the conditional distribution of one differs across categories of the other. Give numerical evidence, such as 60% compared with 35%, and name the groups and response. Similar conditional distributions suggest little or no association in these data. For example, the source hiking table compares 18/45 with 16/55, not 18 with 16. To justify association, write a comparative sentence with both conditional percentages. A sentence reporting only the grand total or a joint percentage does not establish how the distributions differ.

Read categorical comparison graphs

Side-by-side bar charts place the categories of one variable next to one another within each category of the other. For unequal group sizes, proportions are generally more informative than counts. Segmented bar charts give each group a bar of total height 100%, divided into the response categories; compare corresponding segment heights. A mosaic plot also shows group sizes: the widths represent marginal proportions of the grouping variable, while segment heights represent conditional proportions. Therefore rectangle areas represent joint proportions. A narrow group can have a large conditional percentage yet a small share of all observations.

Interpret without overstating

An observed association is a descriptive finding, not automatically evidence of causation or a population relationship. Sampling variation can produce differences, and other variables may help explain the pattern. Later units provide formal tests for population association. For now, distinguish what was observed from why it occurred. Aggregating across a third variable can also change or reverse an association because group composition differs. Inspect subgroup information when it is relevant; do not assume either the combined table or the subgroup table answers every investigative question. Choose the comparison that matches the question and the study design.

KEY IDEA

A conditional percentage answers a question about a restricted group; name that group before dividing.

01

WATCH FOR

P(A given B) and P(B given A) use different denominators.

OPENSTAX WORKED EXAMPLE

Comparing hiking preferences with unequal group sizes

An OpenStax exercise reports a random sample of 100 hikers. Among 45 women, 18 prefer the coastline, 16 prefer lakes and streams, and 11 prefer mountain peaks. Among 55 men, the corresponding counts are 16, 25, and 14. Calculate the conditional percentage preferring the coastline in each group. Is there evidence of an association in this sample between the recorded sex category and hiking preference? Justify your answer with the conditional percentages.

1. The two variables are recorded sex category and preferred hiking area. Each hiker contributes one response to one cell. The coastline total is 18+16=34, and the sample total is 45+55=100.

2. Among the women, 18/45=0.40, or 40%, prefer the coastline. Among the men, 16/55≈0.2909, or 29.1%, prefer the coastline. These denominators compare preference within each group.

3. The coastline preference differs by about 40.0−29.1=10.9 percentage points. Thus at least one category of the conditional preference distribution differs between the groups, indicating association in the sample.

4. The joint percentage who are women and prefer the coastline is 18/100=18%. The overall coastline percentage is 34/100=34%. Neither is the conditional 40% among women.

Conclusion. In this sample of hikers, coastline preference is more common among women (40%) than among men (about 29%). This is a descriptive association, not evidence that one characteristic causes a preference.

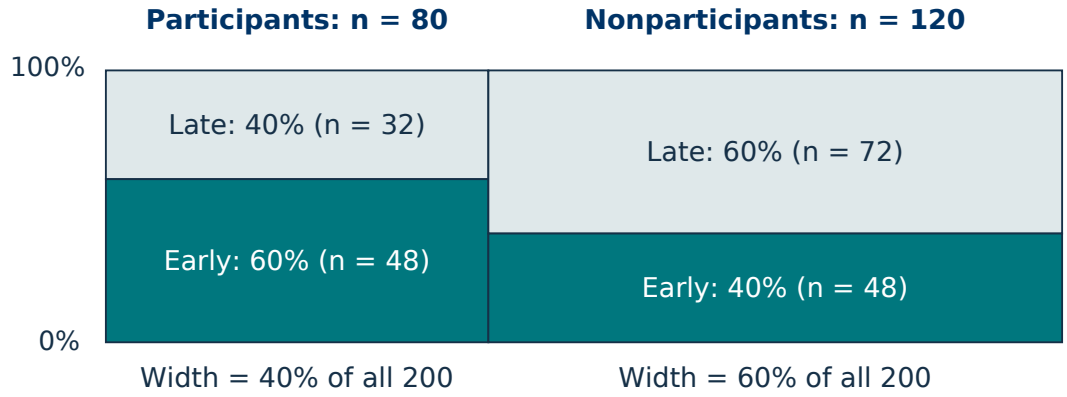
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.4, Example 3.21, Tables 3.5-3.6, printed p. 185-186. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Do the two groups choose differently?

Illustrative survey counts: among 80 students who participate in a school sport, 48 prefer an early activity and 32 a late activity. Among 120 nonparticipants, 48 prefer early and 72 late. Compare preference across participation groups.

A mosaic plot shows two denominators at once



Early rectangle areas: 0.40 × 0.60 = 0.24 and 0.60 × 0.40 = 0.24.

Illustrative worked data. Widths show group shares; heights show within-group preferences; areas show joint proportions.

Complete the margins before choosing a denominator

Participation	Early	Late	Row total
Participant	48	32	80
Nonparticipant	48	72	120
Column total	96	104	200

Illustrative worked data: joint = 48/200; early among participants = 48/80; participants among early = 48/96.

1. There are 200 students. Early preference totals 96; late preference totals 104. The equal early counts, 48 and 48, do not establish equal preference rates because group sizes differ.

2. Among sport participants, the early percentage is 48/80 = 60%; among nonparticipants it is 48/120 = 40%. The observed difference is 20 percentage points.

3. The joint proportion who both participate and prefer early is 48/200 = 24%. The marginal proportion who prefer early is 96/200 = 48%. Neither answers the within-group comparison.

4. A segmented bar chart has heights 60% early and 40% late for participants, compared with 40% early and 60% late for nonparticipants.

5. For a mosaic plot, group widths are 80/200 = 40% and 120/200 = 60%. The first early rectangle has area 0.40 times 0.60 = 0.24, the joint proportion.

Conclusion. Participation and activity preference are associated in this illustrative sample. The observational comparison does not show that sports participation causes a preference.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.4, Example 3.21, Tables 3.5-3.6, printed p. 185-186. Adapted for this course; source data retained unless otherwise stated.

The complete source table and its recorded categories are retained. The original independence question is recast as a descriptive association question because formal probability notation is taught in Unit 3.

S01.03 | Seeing a Quantitative Distribution

CED 1.5, 1.6

- Construct dotplots, stemplots, and equal-width histograms.
- Describe shape, center, variability, gaps, clusters, and unusual values in context.
- Explain how bin choices affect a histogram.

CONCEPT AND INTERPRETATION

A distribution answers more than one question

The distribution of a quantitative variable describes what values occur and how often they occur. A display should preserve numerical order and use a scale with units. Before summarizing, identify the observational unit and the measured variable. Travel times in minutes are not interchangeable with numbers of travelers. A complete description considers shape, center, variability, and unusual features. The purpose is to explain the pattern in context, not to recite vocabulary. A statement such as "the times cluster around 12 minutes and range from 6 to 32 minutes" tells the reader more than "the distribution has spread."

Dotplots and stemplots preserve observations

In a dotplot, each observation contributes one dot at its numerical value. Stack dots for repeated values; do not shift them horizontally just to make the display attractive. Dotplots work well for small or moderate data sets. A stem-and-leaf plot splits each number into a stem and a leaf, orders both, and includes a key showing the place value and units. For example, a key 1|4=14 minutes prevents ambiguity. Empty stems preserve gaps. Splitting stems can reveal patterns in crowded data, but the splitting rule must remain consistent. The source supermarket example illustrates why a key and empty stems matter. Omitting stems 7 through 11 would visually compress the gap before 12.3 kilometers. With 21 observations, a valid stemplot must contain exactly 21 leaves.

Histograms group nearby values

A histogram divides the numerical axis into nonoverlapping intervals called bins. In this course use equal-width bins unless a density scale is explicitly provided. State which endpoints belong to each bin so boundary observations are not counted twice. A bin labeled 10 to less than 15 contains 10 but not 15. Bar heights show counts or relative frequencies and adjacent bins touch because the numerical intervals are connected. Larger bins simplify the picture but can hide clusters; smaller bins expose detail but may make random irregularities look important. Check several sensible bin widths before making a strong claim about shape.

Describe shape with the tails and peaks

A roughly symmetric distribution has halves that approximately mirror each other. A right-skewed distribution has a longer tail toward larger values; a left-skewed distribution has a longer tail toward smaller values. Name skew by the tail, not by the side with most observations. Unimodal means one main peak, bimodal two main peaks, and approximately uniform means no prominent peak with broadly similar frequencies. Small samples can look irregular, so avoid forcing a precise shape label. A gap is an interval with no observed values between populated regions; clusters are concentrations separated by gaps.

Support claims and recognize display limits

An unusually distant value is a potential outlier, which deserves checking rather than automatic deletion. It may be a recording error, a legitimate rare outcome, or a clue that groups were mixed. A graph can suggest unusualness before a numerical rule is chosen. State an approximate center and a meaningful spread with units, and identify unusual values directly. A histogram usually does not recover exact observations within bins, so it cannot generally provide an exact median or mean. A distribution shows the pattern of values; it does not show the original time order unless time itself is on the axis.

KEY IDEA

Choose a display that preserves the information needed to answer the question.

WATCH FOR

A histogram describes quantitative intervals; a bar chart compares separate categories.

OPENSTAX WORKED EXAMPLE

A stemplot of distances to supermarkets

The following distances, in kilometers, are from a home to local supermarkets: 1.1, 1.5, 2.3, 2.5, 2.7, 3.2, 3.3, 3.3, 3.5, 3.8, 4.0, 4.2, 4.5, 4.5, 4.7, 4.8, 5.5, 5.6, 6.5, 6.7, 12.3. Describe how to construct a stemplot and use it to describe the distribution.

1. Use the whole-number part as the stem and the tenths digit as the leaf. Include the key 3|2=3.2 kilometers. Keep a separate leaf for every observation, including both 3.3 values and both 4.5 values.

2. The occupied rows are 1|1 5; 2|3 5 7; 3|2 3 3 5 8; 4|0 2 5 5 7 8; 5|5 6; 6|5 7; and 12|3. Include empty stems 7, 8, 9, 10, and 11 to preserve the large gap.

3. There are 21 observations, so the middle observation is the 11th ordered value, 4.0 kilometers. Distances range from 1.1 to 12.3 kilometers, a range of 11.2 kilometers.

4. Most distances are between about 2 and 6 kilometers. The 12.3-kilometer supermarket is separated from the next largest distance, 6.7 kilometers. The long tail toward larger values makes the distribution right-skewed.

01

Conclusion. The distances are concentrated around 3-5 kilometers, with median 4.0 kilometers. A single distance of 12.3 kilometers is unusually large relative to the rest and creates a long right tail.

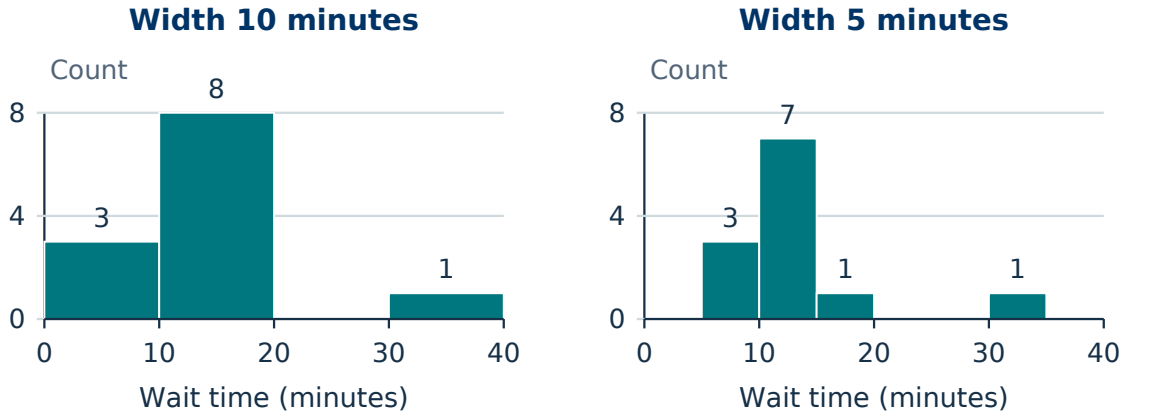
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.1, Example 2.2, printed p. 67-68. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

An unusually long wait

Illustrative wait times in minutes are 6, 8, 9, 10, 10, 11, 12, 12, 13, 14, 16, 32. Construct a useful display and describe the distribution.

Bin width changes the picture, not the observations



Same 12 illustrative waits. Bins include the left endpoint, not the right.

The 32-minute wait remains isolated. Both panels use the same horizontal and vertical scales, so the difference comes from bin width.

A stemplot retains each observed wait

Stem (tens)	Ordered leaves (ones)
0	6 8 9
1	0 0 1 2 2 3 4 6
2	(no observations)
3	2

Illustrative worked data. Key: 1 | 4 means 14 minutes. Repeated leaves represent repeated observations.

1. The observational unit is one service visit and the variable is waiting time in minutes. Sort the values before constructing a stemplot or finding a middle value.

2. A dotplot contains 12 dots, including two at 10 and two at 12. A stemplot with tens as stems has stem 0 with leaves 6, 8, 9; stem 1 with 0, 0, 1, 2, 2, 3, 4, 6; stem 2 empty; stem 3 with 2.

3. For bins [0, 10), [10, 20), [20, 30), and [30, 40), the counts are 3, 8, 0, and 1. The empty bin from 20 to just under 30 minutes shows part of the gap.

4. The distribution has a long right tail, with most waits between 8 and 16 minutes and one isolated value at 32 minutes. Its median is $(11+12)/2=11.5$ minutes.

5. The range is 32-6=26 minutes. The 32-minute visit should be checked against the original record; a graph alone does not establish that it is an error.

Conclusion. The waits are right-skewed, centered near 11.5 minutes, and include an unusually long 32-minute wait separated from the rest.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.1, Example 2.2, printed p. 67-68. Adapted for this course; source data retained unless otherwise stated.

Explanations expand the source reasoning and use AP terminology. The local application uses a separate illustrative data set.

S01.04 | Center, Variability, and Resistance

CED 1.7, 1.8

- Calculate and interpret mean, median, range, variance, and sample standard deviation.
- Select resistant summaries when skewness or extreme values matter.
- Interpret standard deviation in the units of the variable.

CONCEPT AND INTERPRETATION

Center describes location

For observations $x_1, \ldots, x_n$, the sample mean is $\bar{x} = \sum x_i/n$. It is the arithmetic balance point of the data. The median is the middle ordered value, or the mean of the two middle ordered values when n is even. Both measure center, but they answer different questions. The mean uses every numerical value and is useful when totals matter. The median separates the ordered observations into halves and describes a typical location without allowing a single extreme value to dominate. Neither alone describes variability. To find the median of an even-sized data set, identify the two middle positions before averaging their values. For example, the median of 4, 6, 6, 12 is $(6+6)/2=6$, whereas its mean is $(4+6+6+12)/4=7$. These calculations answer different questions even though both use division.

Spread describes how much values differ

The range is maximum minus minimum. It is simple but depends entirely on the two extremes. Standard deviation uses deviations from the mean to summarize the typical distance of observations from that mean. It is not literally the mean absolute distance. Squaring deviations prevents positive and negative deviations from canceling; the square root restores the original units. Variance is the squared standard deviation and has squared units. A standard deviation of 3 minutes concerns the spread of individual times around their mean, not uncertainty in the mean itself.

01

Calculate sample standard deviation carefully

The sample standard deviation is $s = \sqrt{\sum(x_i - \bar{x})^2/(n-1)}$. First find the mean, then each deviation, then square and add the deviations, divide by n-1, and take the square root. The n-1 denominator accounts for using the sample mean in estimating population variability: after the mean is fixed, the deviations sum to zero. A calculator may also report a population standard deviation using n in the denominator. Select the sample result when the observations are a sample. Retain extra digits during calculations and round the final answer with units.

Resistance matters when extremes move

A resistant statistic changes relatively little when a small number of extreme values move farther away. The median and the interquartile range, the spread of the middle half of the ordered data, are resistant; the mean, range, and standard deviation are not. Resistance does not mean a statistic never changes or that extreme observations should be discarded. It means the summary is less sensitive to their magnitudes. For a skewed distribution or one with unusual extreme values, median and IQR often summarize a typical observation more clearly. Mean and standard deviation remain useful when the investigative question concerns an average or total.

Use summaries with a graph

For approximately symmetric, unimodal data without extreme values, the mean and median are often close. In a right-skewed distribution the mean is usually larger than the median, and the reverse usually occurs with left skew. These are useful tendencies, not a method for reconstructing the whole distribution from two numbers. Different shapes can share the same mean and median. Compare groups directly using the same variable and units; report both center and spread. A smaller mean wait is desirable only if that matches the question, and a similar mean can conceal very different consistency.

KEY IDEA

A center locates a distribution; a spread describes its variability. Report both.

WATCH FOR

Sample standard deviation uses n-1; a calculator may display a different population standard deviation beside it.

OPENSTAX WORKED EXAMPLE

When one income changes the mean

In the OpenStax scenario, a town has 50 people. One person earns \$5,000,000 per year and the other 49 each earn \$30,000 per year. Calculate the mean and median annual income. Which measure better describes the income of a typical resident? Justify your choice.

1. The combined annual income is \$5,000,000+49(\$30,000)=\$6,470,000. The mean is \$6,470,000/50=\$129,400 per person per year.

2. With 50 ordered values, the median is the mean of the 25th and 26th values. Both are $30,000, so the median annual income is $30,000.

3. Forty-nine of the 50 residents earn $30,000. The median therefore describes the income of a typical resident more closely. The single very large income raises the mean far above the incomes of almost everyone.

4. The mean still answers a useful question about total income per resident. Multiplying the mean by 50 recovers the total; multiplying the median by 50 does not. The appropriate summary depends on the question.

Conclusion. The median annual income of $30,000 represents the typical resident in this scenario. The mean of $129,400 is strongly affected by the one $5,000,000 income.

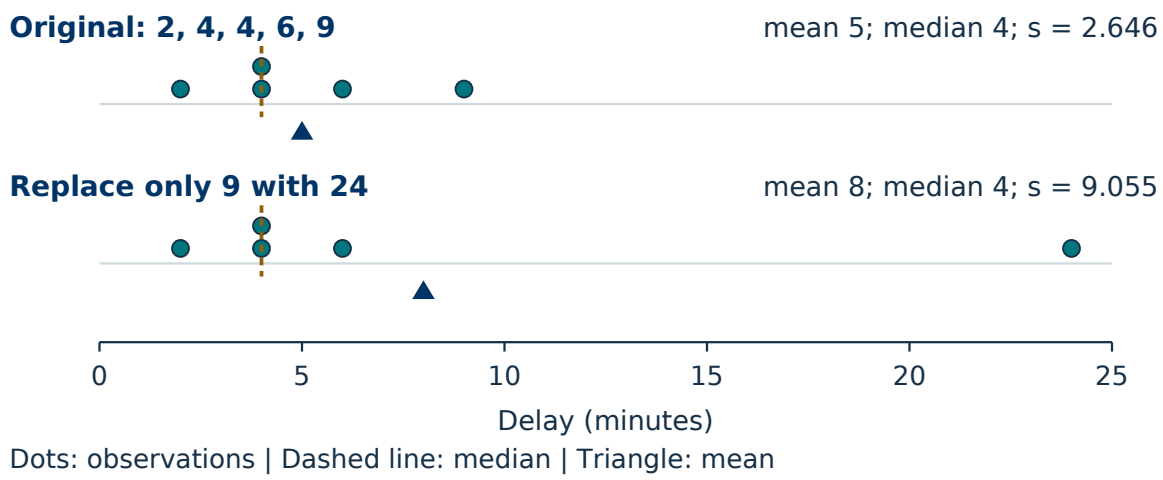
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.5, Example 2.27, printed p. 100. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

How variable are the delays?

Five illustrative delivery delays are 2, 4, 4, 6, and 9 minutes. Compute the mean and sample standard deviation, then consider replacing 9 with 24.

One extreme value moves the mean and spread



Illustrative worked data. The median remains 4 minutes; moving the largest value increases the mean and sample standard deviation.

01

Build sample standard deviation from deviations

Delay x (min)	x − mean (min)	Squared deviation (min²)
2	−3	9
4	−1	1
4	−1	1
6	1	1
9	4	16
Total	0	28

Mean = 5 minutes. Sample variance = 28/(5 − 1) = 7 min²; sample SD = √7 ≈ 2.646 minutes.

1. The sum is 25, so the mean is 25/5=5 minutes. The median is 4 minutes after ordering.

2. Deviations from 5 are -3,-1,-1, 1, and 4 minutes. Their squares are 9, 1, 1, 1, and 16, summing to 28 square minutes.

3. Sample variance is 28/(5-1)=7 square minutes. The sample standard deviation is the square root of 7, approximately 2.646 minutes. Delays typically differ from their mean by about 2.65 minutes.

4. The range is 9-2=7 minutes. The mean exceeds the median, consistent with the higher values pulling the mean upward, although five observations are too few for a confident shape label from summaries alone.

5. Replacing 9 with 24 changes the mean to 8 and the sample standard deviation to about 9.055 minutes, while the median remains 4. The large observation strongly affects the nonresistant summaries.

Conclusion. Replacing the 9-minute delay with a 24-minute delay raises the mean from 5 to 8 minutes and the sample SD from about 2.65 to 9.06 minutes. The median remains 4 minutes.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.5, Example 2.27, printed p. 100. Adapted for this course; source data retained unless otherwise stated.

The hypothetical source scenario is retained. The calculation of the mean and the distinction between a typical individual and a per-person total are developed explicitly.

S01.05 | Quartiles, Outliers, and Group Comparisons

CED 1.7, 1.8, 1.9

- Construct a five-number summary and modified boxplot.
- Apply and distinguish the 1.5 IQR and two-standard-deviation outlier criteria.
- Compare quantitative distributions with context and numerical evidence.

CONCEPT AND INTERPRETATION

Find quartiles with a declared convention

The five-number summary consists of minimum, first quartile Q_1, median, third quartile Q_3, and maximum. Q_1 and Q_3 mark the lower and upper boundaries of the middle half of the ordered data. For hand calculations in this package, find the median and then find the median of each half; when n is odd, omit the overall median from both halves. Other software uses interpolation conventions, which can produce slightly different quartiles in small samples. State the convention when an exact result depends on it and use supplied quartiles when the problem provides them.

Measure the spread of the middle half

The interquartile range is $IQR=Q_3-Q_1$. It describes the width of the interval containing the middle approximately 50% of observations. Its endpoints are quartiles, values determined by position in the ordered data. Moving the largest observation farther out need not change either quartile. An IQR of 4 minutes does not mean every observation is within 4 minutes of the median. The range measures the full distance between the two extremes, while the IQR focuses on the middle. Comparing both can reveal whether two groups differ mainly in their typical variation or in unusually extreme outcomes.

Use an explicit rule for unusual values

The 1.5 IQR rule flags values below Q_1-1.5 IQR or above Q_3+1.5 IQR. A value exactly on a fence is not beyond it. These fences are thresholds, not necessarily observed values. The revised CED also describes a rule flagging values more than two standard deviations above or below the mean. The rules are not equivalent: mean and standard deviation are sensitive to extremes, whereas quartiles and IQR are resistant. Name the rule used, compare actual values to its cutoffs, and do not automatically delete flagged observations.

Read and construct a modified boxplot

A box extends from Q_1 to Q_3 and includes a line at the median. In a modified boxplot, whiskers end at the most extreme observed values that are not flagged by the chosen outlier rule, generally 1.5 IQR; plot flagged values separately. Do not extend whiskers to the fences unless a fence happens to equal the endpoint observation. Label the numerical axis and units. Boxplots efficiently compare centers, middle spreads, and extremes across groups on a common scale, but they do not reveal all gaps, clusters, or modes within the quartile intervals. If the two quartiles and the median coincide or several observations are tied, do not assume every closed interval between quartiles contains exactly 25% of the observations. The ranks define the quartiles, while repeated numerical values can fall on shared boundaries. Describe the middle approximately 50% and inspect the original data when an exact count is required.

01

Two routes on one back-to-back stemplot

North route	Stem	South route
6 4 2	1	5 8
9 3 3 0	2	1 4 6
	3	0 2

Key: stem 2 with leaf 3 means a travel time of 23 minutes on either route.

Each leaf is one illustrative trip. Read outward from the common stem toward larger leaves. Each side has seven observations.

Compare rather than list

Compare the groups on a common scale and support the comparison with values and units. In a back-to-back stemplot, both groups use the same stems. Each group has its own leaves, with the smallest leaf nearest the stem. Count repeated leaves as separate observations. Boxplots make medians and quartiles quick to compare, while a stemplot preserves the observed values, gaps, and clusters. Report what the displays support about the groups and keep any broader claim within the study design.

KEY IDEA

An outlier rule flags a value for investigation; it does not prove the value is an error.

WATCH FOR

Boxplot whiskers end at observed nonoutlier values, not at calculated fences.

OPENSTAX WORKED EXAMPLE

Which house price is a potential outlier?

OpenStax gives these 13 house prices, in dollars: 389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000. Use the median-of-halves convention, excluding the overall median from both halves, to calculate the interquartile range and identify any potential outliers using the 1.5 IQR rule.

1. Order the values: 114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000. The seventh value, $488,800, is the median.

2. Exclude the median, leaving six values in each half. The lower quartile is $(230{,}500+387{,}000)/2=\$308{,}750$. The upper quartile is $(639{,}000+659{,}000)/2=\$649{,}000$.

3. The IQR is $\$649{,}000-\$308{,}750=\$340{,}250$. The lower fence is $\$308{,}750-1.5(\$340{,}250)=-\$201{,}625$. The upper fence is $\$649{,}000+1.5(\$340{,}250)=\$1{,}159{,}375$.

4. Only \$5,500,000 exceeds the upper fence. No value is below the lower fence. In a modified boxplot, the upper whisker ends at the largest nonoutlier, \$1,095,000, not at the upper fence.

Conclusion. The \$5,500,000 house is the only potential outlier under the 1.5 IQR rule. The numerical rule flags it for investigation; it does not establish that its recorded price is incorrect.

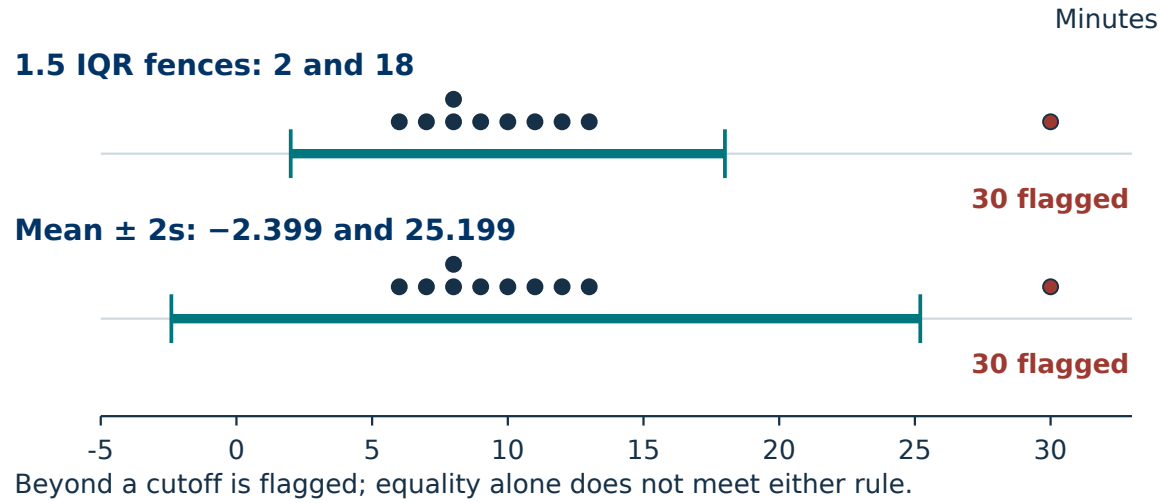
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.3, Example 2.13, printed p. 87. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

A modified boxplot for trip times

Illustrative trip times are 6, 7, 8, 8, 9, 10, 11, 12, 13, 30 minutes. Use the median-of-halves convention to construct a modified boxplot and compare two outlier criteria.

Two rules, two cutoffs, one flagged trip



Illustrative worked data. Q1 = 8, Q3 = 12, IQR = 4; mean = 11.4 and sample SD ≈ 6.899. The rules happen to agree here.

1. The median is (9+10)/2=9.5 minutes. The lower half 6, 7, 8, 8, 9 has median 8; the upper half 10, 11, 12, 13, 30 has median 12. Thus Q_1=8 and Q_3=12.

2. IQR=12-8=4 minutes. The fences are 8-1.5(4)=2 and 12+1.5(4)=18 minutes. Only 30 lies beyond a fence.

3. The modified box runs 8 to 12 with median 9.5. Whiskers reach 6 and 13, the smallest and largest observed nonoutliers. Plot 30 separately.

4. The sample mean is 11.4 minutes and sample standard deviation is approximately 6.899 minutes. The two-SD cutoffs are approximately -2.399 and 25.199 minutes. This rule also flags 30 in this particular data set.

5. The rules happened to agree here; that agreement is not guaranteed. The mean and SD shift when the largest observation changes, whereas the median and IQR can remain stable.

Conclusion. Typical trip times are much shorter than 30 minutes. The value should be checked and discussed, not discarded merely because a rule flags it.

01

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.3, Example 2.13, printed p. 87. Adapted for this course; source data retained unless otherwise stated.

Explanations expand the source reasoning and use AP terminology. The local application uses a separate illustrative data set.

S01.06 | Relative Standing and Changing Units

CED 1.7, 1.9

- Interpret percentile rank using the CED at-or-below convention.
- Compute and compare standardized scores.
- Predict effects of linear transformations on center and spread.

CONCEPT AND INTERPRETATION

Percentiles describe position in order

A percentile describes relative position within an ordered distribution. The revised CED uses the convention that the pth percentile has approximately p percent of observations less than or equal to it. For an observed value in a finite data set, this package calculates its empirical percentile rank as 100 times the number at or below the value divided by n. Ties can cause jumps, so not every percentile is attained exactly. Software may interpolate percentile cutoffs differently. Always distinguish a percentile rank, which is a percentage, from the data value at that rank, which retains the variable's units.

Standardize a distance from the mean

For a population mean μ and positive standard deviation σ, the standardized score is $z = (x - \mu)/\sigma$. It tells how many standard deviations x lies above or below the mean. Positive values are above; negative values are below. If only sample summaries are available, use $\bar{x}$ and s and identify that choice. A z-score has no units because the numerator and denominator have the same units. Standardizing requires positive spread; if every value is identical, the denominator is zero and the usual z-score is undefined.

Compare standing, not raw magnitudes

A raw score can be larger while its standardized position is lower. Compare z-scores to compare distances relative to the distribution's spread. Whether a larger z-score is better depends on context: high points may be desirable, while a high race time is not. A z-score is not itself a percentile. Converting z to a normal-model percentage requires an appropriate normal model, developed in Unit 4. Distributions with different shapes can assign different percentile ranks to the same z-score. State precisely whether you are comparing standard-deviation distances or percentiles. The swimmers in the source example show why the direction matters. A z-score of −2 is a shorter time relative to the team mean than −1.25. Write the contextual comparison before attaching a value judgment such as better.

Translate and rescale data

A linear transformation changes every value using $y = a + bx$. The mean becomes $a + b\bar{x}$ and standard deviation becomes $|b|s$. Adding a shifts all values equally without changing distances between them. Multiplying by b rescales every distance by its absolute value. Range and IQR also multiply by $|b|$; variance multiplies by b^2. For b positive, order and shape are preserved. For b negative, order reverses and the shape is reflected. Recompute or reverse quartile roles when b is negative rather than blindly calling a+bQ 1 the new lower quartile.

Use transformations to check reasoning

Converting measurements from centimeters to millimeters multiplies center and spread by 10. Converting Celsius to Fahrenheit multiplies by 1.8 and adds 32, so the mean changes by both operations but standard deviation only by the factor 1.8. Standardizing an entire population produces mean 0 and standard deviation 1; standardizing a sample by its own sample mean and s produces sample mean 0 and sample SD 1. These transformations do not make an irregular distribution normal. They change the scale while retaining its pattern, except for reflection when the multiplier is negative.

KEY IDEA

A percentile is a data value. Its percentile rank is the percentage of observations at or below it. A z-score measures distance from the mean in standard deviations.

WATCH FOR

Shifting every observation changes center but leaves spread unchanged.

OPENSTAX WORKED EXAMPLE

Comparing two swimmers with their own teams

Angie swims the 50-meter freestyle in 26.2 seconds. Her team has mean time 27.2 seconds and standard deviation 0.8 second. Beth swims it in 27.3 seconds. Her team has mean time 30.1 seconds and standard deviation 1.4 seconds. Which swimmer is faster relative to her own team, using standardized scores? Show your calculations and interpret the signs.

1. For Angie, z=(26.2−27.2)/0.8=−1.25. Her time is 1.25 standard deviations below her team mean.

2. For Beth, z=(27.3−30.1)/1.4=−2.00. Her time is 2 standard deviations below her team mean.

3. A smaller race time represents a faster performance. Beth is farther below her team mean and therefore has the faster relative performance by this standardized comparison, although Angie has the faster raw time.

4. No normal distribution was stated. These z-scores describe standard-deviation distances; they do not by themselves give the swimmers' percentiles.

Conclusion. Beth performed better relative to her team by the z-score comparison: 2 standard deviations faster than her team mean, compared with Angie's 1.25 standard deviations.

01

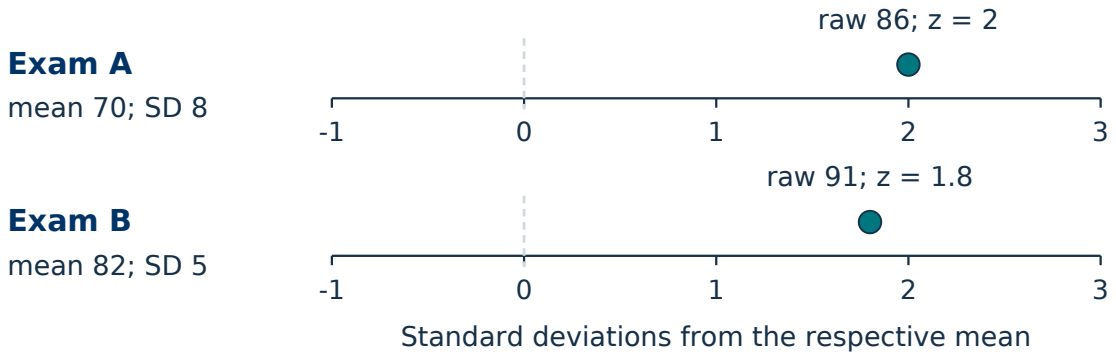
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.7, Try It 2.35, Table 2.36, printed p. 117. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Which performance stands farther above its group?

Illustrative exam A has population mean 70 and SD 8; exam B has mean 82 and SD 5. A student earns 86 on A and 91 on B. Compare standardized standing. Then exam A scores are transformed by y=10+0.9 x.

A common scale for different exams



A stands farther above its mean: z = 2 > 1.8. No normal model is assumed.

Illustrative worked scores. A higher raw score need not have the higher standardized position.

Track every quantity through y = 10 + 0.9x

Quantity	Original exam A	Transformed exam A
Population mean	70	10 + 0.9(70) = 73
Population SD	8	0.9(8) = 7.2
Student score	86	10 + 0.9(86) = 87.4
Standardized score	(86 − 70)/8 = 2	(87.4 − 73)/7.2 = 2

A positive linear transformation preserves z-scores. The additive constant changes location, not spread.

1. For exam A, z=(86-70)/8=2. The score lies 2 standard deviations above its exam mean.

2. For exam B, z=(91-82)/5=1.8. The raw score is larger on B, but the standardized distance above the mean is smaller.

3. The student stands farther above the mean on A by this standard-deviation comparison. Without distribution shapes, we cannot conclude exact percentile ranks or assume they follow a normal model.

4. For transformed exam A, the mean is 10+0.9(70)=73 and SD is 0.9(8)=7.2. The student's transformed score is 10+0.9(86)=87.4.

5. The new standardized score is (87.4-73)/7.2=2, unchanged because a positive linear transformation preserves relative standardized position.

Conclusion. The student stood 2 SDs above the mean on exam A and 1.8 SDs above the mean on exam B. Exam A shows the stronger relative performance even though its raw score was lower. The z-scores alone do not give percentile ranks.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 2.7, Try It 2.35, Table 2.36, printed p. 117. Adapted for this course; source data retained unless otherwise stated.

Explanations expand the source reasoning and use AP terminology. The local application uses a separate illustrative data set.

Unit synthesis

1. An illustrative survey of 200 students records grade, arrival mode, and travel time. Identify a categorical and a quantitative variable and write an investigative question with a defined population.

2. Among 80 students in group A, 48 prefer option 1; among 120 in group B, 48 prefer option 1. Compare groups and compute the joint proportion in A and preferring 1.

3. Use data 2, 4, 4, 6, 9. Find mean, median, and sample SD; interpret the SD if the data are minutes.

4. Use data 6, 7, 8, 8, 9, 10, 11, 12, 13, 30. Find IQR, fences, and whisker endpoints under the stated convention.

5. Population summaries are mean 50 and SD 10. Find z for 65. Then transform all values by y=100-2 x and determine transformed mean, SD, and z for the transformed 65.

6. A report claims that the school's commuting habits cause a difference in test scores because survey groups have different means. Identify two distinct limits.

Unit 02 | Collecting Data

Good analysis cannot repair every weakness in data collection. This unit connects the investigative question to the population, sampling plan, measurements, and conclusions. You will describe reproducible random selection and random assignment, distinguish bias from ordinary sampling variation, and choose between completely randomized, block, and matched-pairs experiments. The central distinction is practical: random sampling supports population generalization, while random assignment supports causal interpretation when the experiment is well designed. Both require careful implementation. Activities use invented populations or low-risk classroom tasks, with voluntary participation and an equivalent alternative when actual data are collected.

CED topics: 1.10, 1.11, 1.12, 1.13

IN THIS UNIT

Lesson	Focus
S02.01	Choosing a Random Sample
S02.02	Bias, Observational Studies, and Ethical Questions
S02.03	Designing a Completely Randomized Experiment
S02.04	Blocking, Matched Pairs, and Scope of Inference



A sample from the orchard

LOCAL QUESTION

How could an orchard sample avoid favoring fruit that is easiest to reach? Compare random selection across trees with convenience selection beside the path.

S02.01 | Choosing a Random Sample

CED 1.10, 1.11

- Describe reproducible simple random, stratified, cluster, and systematic sampling procedures.
- Choose a sampling method that fits a population and investigative question.
- Distinguish selection variability from systematic bias.

CONCEPT AND INTERPRETATION

Define the population and sampling frame

The target population consists of the units about which the question is asked. A sampling frame is the list or mechanism used to reach them. If the population is current students, a current complete roster is more suitable than last year's club list. A census records information from every population member; it can still suffer measurement or nonresponse problems. A sample studies only a subset. Random sampling uses a chance mechanism rather than convenience or personal choice. The frame, selection method, and response process together determine whether a sample can support a useful population estimate.

A simple random sample selects sets fairly

A simple random sample of size n gives every possible subset of n distinct population members the same chance of selection. Equal selection chances for individuals alone are not sufficient for an SRS; some designs allow only restricted combinations. To describe an SRS, retrieve a complete frame, label units uniquely, and use a random generator to select n distinct labels. If the generator samples with replacement, ignore duplicate labels and continue until n distinct units have been selected. State how invalid labels are handled. Saying "pick randomly" does not describe a procedure another person could reproduce.

Stratify to represent meaningful groups

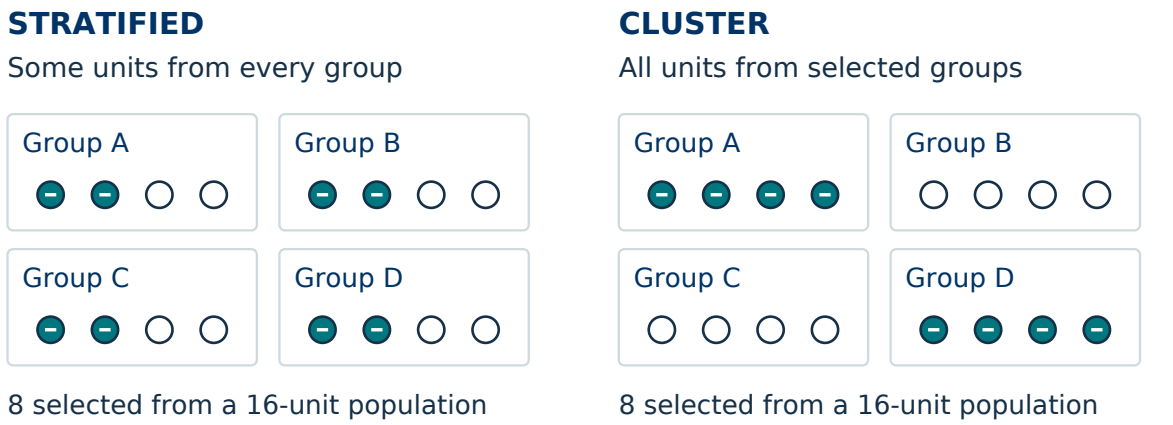
A stratified random sample partitions the population into nonoverlapping strata, then selects an independent random sample from every stratum. Strata should relate to the variable of interest when possible, so units within a stratum are relatively similar. This can reduce variability of estimates and ensure representation. Proportional allocation selects each stratum in proportion to its size. Disproportionate allocation can deliberately obtain more information about a small stratum, but an unweighted overall percentage may then misrepresent the population. Sampling within every stratum distinguishes this method from selecting a few clusters. The source tuition example uses the same four year groups in two different ways. Sampling within all four creates strata; selecting two whole year groups creates clusters. Drawing four boxes is therefore not enough to identify the method. Mark which groups are selected and which students within them can enter the sample.

02

Clusters and systematic selection address logistics

A cluster sample randomly chooses some natural groups and studies every unit in the selected groups. For example, select several homerooms and survey every student in those rooms. Clusters ideally resemble small versions of the whole population; strong within-cluster similarity can increase sampling variability. A systematic sample chooses a random starting position and then every kth unit. For a 240-unit ordered frame and a 24-unit sample, choose a random start 1 through 10, then select every tenth unit. Periodicity in the frame that matches the interval can produce unrepresentative samples. Describe the start, interval, and stopping rule explicitly.

The groups enter the two sampling designs differently



Schematic: each group has 4 units. Filled circles with white bars are selected.

This schematic shows one possible random-selection outcome. Strata sample every group; cluster sampling selects whole groups.

Justify the method, not just the label

No method is universally best. An SRS is straightforward with a complete accessible frame. Stratification helps when meaningful subgroups differ and subgroup representation matters. Clustering can reduce travel or administrative cost when individual sampling is difficult. Systematic selection can be convenient on a list or production line if its ordering is understood. Random methods still yield different samples and statistics by chance; this sampling variability is not itself evidence of bias. Larger samples usually reduce variability under a sound design, but they do not automatically correct a defective frame or a response process that systematically excludes some people.

KEY IDEA

An SRS gives every possible sample of the specified size the same chance.

WATCH FOR

Stratified sampling takes some units from every group; cluster sampling takes all units from selected groups.

OPENSTAX WORKED EXAMPLE

Identify a sampling method from its selection procedure

A college wants to estimate the mean tuition paid by its undergraduate students in the fall semester. Consider three plans from OpenStax. A: divide the complete student list into first-year, sophomore, junior, and senior groups, and randomly select 25 distinct students from each group. B: randomly select two of the four year groups and survey every student in the selected groups. C: survey the first 100 undergraduates encountered outside the library on Wednesday. Identify each sampling method and explain the defining feature.

1. Plan A is a stratified random sample. Every year group contributes a random sample. The year groups are the strata. Random selection must occur within each stratum; merely taking 25 available students would not meet that description.

2. Plan B is a cluster sample. Only two year groups are selected, but every student in each selected group is included. The year groups are used as clusters, not strata.

3. Plan C is a convenience sample. Selection depends on being outside the library at that time and being among the first 100 encountered. Students unavailable there may have different tuition patterns.

4. The same grouping variable can define strata or clusters. The distinction is the selection procedure: some students from every stratum versus all students from selected clusters.

Conclusion. Plans A, B, and C are stratified random, cluster, and convenience samples respectively. The method is identified by how students are selected, not by the mere presence of groups.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.2, Example 1.11, parts a, d, and e, printed p. 21. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Sampling a 240-student roster

An invented school has 240 students: 120 ninth graders, 60 tenth graders, 40 eleventh graders, and 20 twelfth graders. The goal is to estimate the proportion who prefer a new lunch option using 24 responses. Compare an SRS, a proportional stratified sample, and a systematic sample.

Three ways to select 24 from the illustrative 240-student roster

Design	Selection procedure	What to check
SRS	Choose 24 distinct labels uniformly from 001-240.	Complete frame; duplicates excluded.
Stratified by grade	Select 10% within each grade at random: 12, 6, 4, and 2 students.	All four grades contribute; selection is random within each.
Systematic	Choose r uniformly from 1-10; select r, r+10, …, r+230.	Exactly 24 positions; inspect periodic patterns in roster order.

The sample size is 24 in each plan. The mechanism, rather than the word random alone, defines the design.

1. For an SRS, label the complete roster 001 through 240. Use a random generator to choose 24 distinct integers in that range, then survey those students. If a selected student does not respond, record nonresponse rather than quietly substituting a convenient friend.

2. For proportional stratification, sample 24/240 = 10% of each grade. Select 12, 6, 4, and 2 students respectively. Number each grade's complete roster and independently draw that grade's required number of distinct labels at random.

3. These selections total 24 and preserve the four population grade shares. Choosing six students per grade would give each grade 25% of the sample, although ninth graders make up 50% of the school. Differences in lunch preference by grade can make proportional stratification more precise.

4. For systematic sampling, choose one random integer r from 1 through 10 and select roster positions r, r+10,..., r+230. This yields 24 students.

5. Inspect the ordering: if each block of 10 contains a recurring category in the same position, the systematic sample might repeatedly select that category. Choose a different method or randomize the list first.

Conclusion. If the lunch survey needs each grade represented in its population share, randomly select 12 ninth graders, 6 tenth graders, 4 eleventh graders, and 2 twelfth graders. An SRS of 24 allows those counts to vary.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.2, Example 1.11, parts a, d, and e, printed p. 21. Adapted for this course; source data retained unless otherwise stated.

Random selection within strata is made explicit because the original part a does not state it. The original source setting and group counts are retained.

S02.02 | Bias, Observational Studies, and Ethical Questions

CED 1.10, 1.12

- Identify undercoverage, nonresponse, response, and voluntary-response bias.
- Distinguish surveys, prospective studies, retrospective studies, and experiments.
- Write a measurable investigative question and evaluate ethical data collection.

CONCEPT AND INTERPRETATION

Make the question guide the design

An investigative question should specify variables, population, and the kind of conclusion sought. An estimation question asks for a plausible value or range for a population quantity, such as the mean one-way commute time of current students. A testing question asks whether evidence supports a specified direction, such as whether more than half prefer an option. A relationship question names both variables. A causal question requires an appropriate experiment, not just different survey averages. Set the question before collecting and analyzing data; changing the target after seeing results can make chance patterns look more convincing than they are.

Observe without imposing treatment

An observational study records variables without assigning treatments. A survey obtains information from people using a standard set of questions. A prospective study selects units now and follows them forward; a retrospective study selects units now and gathers information about their past. Neither direction in time makes the study an experiment. An experiment imposes treatments on experimental units. In an observational association, a confounding variable offers an alternative explanation because it is associated with both the explanatory and response variables. Naming any unrelated variable does not establish confounding; explain both relationships.

Separate selection failures

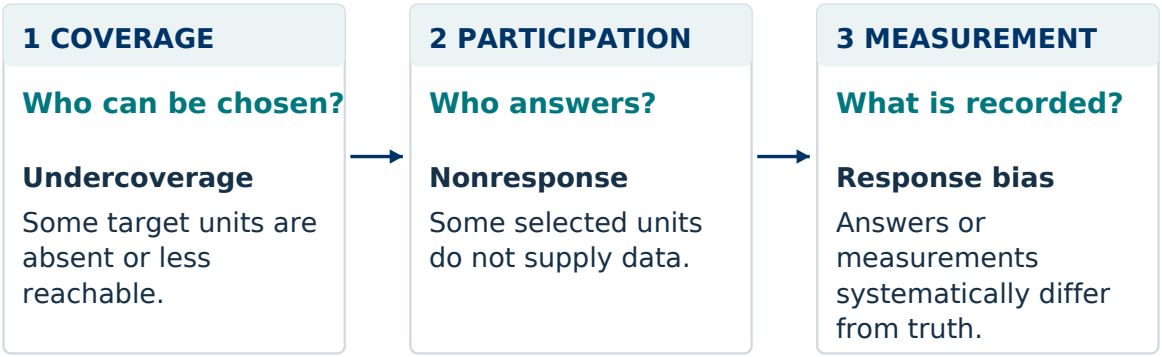
Undercoverage occurs when some population members are missing from the frame or have too little chance of selection. Nonresponse arises when selected individuals do not supply the requested data. Voluntary-response samples arise when people decide to enter an open invitation; strong opinions or special interests may be overrepresented. A convenience sample uses readily available units without a probability selection scheme. These problems can make a statistic systematically too high or too low, but the direction requires contextual evidence. A large sample with a biased selection mechanism can estimate the wrong quantity very precisely. A useful explanation of bias connects the method to the response being measured. For example, a library poll restricted to frequent library visitors could overestimate support for longer hours if those visitors are more likely to favor the change. Naming undercoverage alone does not explain that mechanism; without a defensible relationship, report that the direction is unknown.

Separate measurement failures

Response bias occurs when recorded responses systematically differ from the underlying truth. Leading wording, confusing time frames, social pressure, interviewer effects, and inaccurate recall can contribute. A respondent who gives an inaccurate answer is different from a selected person who never responds. Improve measurement with neutral wording, a specific reference period, clear response options, and privacy protections. Pilot the instrument to discover misunderstandings. Do not assume all errors cancel: consistent pressure in one direction can remain even in a huge data set. Report missingness and explain whether the analysis excludes missing responses.

02

Find the step that can distort the data



An open invitation adds self-selection: volunteers decide whether to enter.

A larger sample does not automatically repair any of these mechanisms.

These mechanisms can coexist. Explain how the process could change the statistic before asserting a direction of bias.

Respect participants and scope

Collect only information needed for the question, explain its purpose, protect identifiers, and allow meaningful choice about participation. Classroom investigations should avoid unnecessary sensitive information and provide an equivalent alternative. Consent alone does not make any design acceptable: consider risk, privacy, and whether assigning the proposed treatment is ethical. Random sampling supports generalization to the sampled population when coverage and response are adequate. A nonrandom sample gives weaker grounds for broad generalization; similar-population applications require judgment and explicit limits. Even excellent random sampling does not turn an observational association into a causal conclusion.

KEY IDEA

Bias is a systematic problem in the process; more observations do not automatically remove it.

WATCH FOR

Nonresponse means no answer from a selected person; response bias means the answer is systematically inaccurate.

OPENSTAX WORKED EXAMPLE

Can restricted samples represent all part-time students?

A college has 10,000 part-time students. OpenStax considers estimating their mean fall-term book expenditure with either of two samples: ten students in a first-term organic chemistry class, many also taking calculus, or ten older students selected from a physical-education class list. Explain why neither sampling frame necessarily represents all part-time students and describe a plausible direction of bias for the chemistry sample.

1. The population parameter is the mean fall-term book expenditure of all 10,000 part-time students. A class list is a restricted sampling frame, even if some chance mechanism is used to select names from that list.

2. The chemistry sample excludes students outside that class. Chemistry and calculus books may be relatively expensive, so this sample could overestimate the college-wide mean book expenditure. The direction is a plausible consequence of the course mix, not a certainty inferred from sample size alone.

3. The physical-education list excludes other part-time students. Its students may buy a different number or type of books. Systematic selection within that restricted list cannot repair its undercoverage of the target population.

4. A better plan starts with the full current part-time-student roster, uses random selection, asks the same expenditure question for the same term, and records nonresponse. Increasing the size of either restricted sample would not remove its coverage problem.

Conclusion. Neither class-based frame adequately covers the stated population. A precise calculation from either sample can still estimate the wrong population mean because the selection process is systematically restricted.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.2, Example 1.13, printed p. 23-24. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

What does the online poll estimate?

An invented city wants the proportion of all adult residents who support extending library hours. A library website poll receives 800 responses, 640 supporting the extension. Evaluate the result and improve the design.

The online poll answers a narrower question

Quantity	What is known from the illustrative poll?
Target population	All adult city residents.
Actual observed group	800 people who chose to answer on the library website.
Observed support	640/800 = 80% of respondents.
Citywide support proportion	Unknown from these counts alone.
Coverage / participation concerns	Website users may differ from other residents; participation is voluntary.

The numerator and denominator establish a respondent statistic. They do not establish that respondents represent all adult residents.

1. The target population is all adult city residents and the variable is support for the extension. The observed statistic is 640/800=80% among respondents.

2. The poll is voluntary response: people choose whether to participate. It also undercovers residents who do not use the website, including some who rarely use the library.

3. If regular library users are more supportive and more likely to respond, the observed percentage could overestimate citywide support. This direction is a reasoned possibility, not established solely by the large 80% result.

4. Use an appropriate comprehensive adult-resident frame, draw a random sample, and contact selected residents using accessible modes. Protect privacy and follow up with nonrespondents without replacing them with convenient respondents.

5. Use neutral wording that describes the actual proposed change and any relevant tradeoff. Report the response rate, coverage limits, and exact denominator.

Conclusion. The poll establishes 80% support among its 800 respondents. Its size does not justify treating that number as the citywide proportion.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.2, Example 1.13, printed p. 23-24. Adapted for this course; source data retained unless otherwise stated.

The source's restricted-course sampling scenario is retained. The numerical spending lists are unnecessary for the design question and are omitted. Claims about bias direction are stated as plausible, not established.

S02.03 | Designing a Completely Randomized Experiment

CED 1.10, 1.13

- Identify experimental units, factors, levels, treatments, and responses.
- Describe complete random assignment with comparison, replication, and direct control.
- Explain placebo effects, blinding, and the role of chance in experimental conclusions.

CONCEPT AND INTERPRETATION

Impose a treatment and measure a response

An experiment assigns conditions to experimental units and measures outcomes. The experimental unit is the smallest unit to which a treatment is assigned. Several measurements on that unit still represent one assignment. A factor is an explanatory variable whose levels are imposed. With one factor, its levels define the treatments. With multiple factors, each combination of levels is a treatment. The response variable records the outcome after assignment and should be defined precisely. If a classroom is assigned one method and all its students are measured, the classroom may be the experimental unit, not each student independently.

Build a fair comparison

A well-designed experiment compares at least two treatment groups, uses random assignment, replicates treatments across units, and directly controls relevant conditions when possible. A control group provides a baseline, such as the standard method, no treatment, or a placebo. Direct control keeps potential extraneous sources of variation the same, such as task length, instructions, and room conditions. Replication means more than one independently assigned experimental unit per treatment; repeated measurements on one unit do not provide the same replication. More units help distinguish a treatment effect from ordinary unit-to-unit variation. The source aspirin exercise illustrates three separate protections. The placebo provides a comparison for the active treatment, random assignment addresses systematic differences in who receives each pill, and blinding limits changes in behavior or measurement caused by knowing the assignment. None substitutes for the others.

Random assignment addresses confounding

In a completely randomized design, assign treatments using a chance mechanism across all eligible experimental units. For two equally sized groups, label all units uniquely, select half the labels at random without replacement for one treatment, and assign the remainder to the other. Random assignment tends to balance both measured and unmeasured extraneous variables across groups, preventing a deliberate link between those variables and treatment. It does not guarantee perfect balance in a particular experiment. Chance imbalances remain possible, which is why statistical evidence is still needed before interpreting an observed difference as a treatment effect.

Control expectations and measurement

A placebo resembles an active treatment while lacking its active component. A participant's response can change because they expect a treatment to help. Comparing a placebo group with a group receiving no treatment can help investigators study that effect. A placebo group can help separate those expectations from the active component. Blinding, also called masking, limits knowledge of treatment assignment. In a single-blind design, participants or interacting research personnel do not know assignment; state exactly who is masked. In a double-blind design, neither participants nor personnel who interact with them know. Some interventions cannot conceal treatment, but outcome assessors can often still be blinded.

Different design features address different problems

Feature	Purpose	Does not guarantee
Random assignment	Reduce systematic links between treatment and extraneous variables.	Perfectly identical groups in this experiment.
Replication	Use multiple independent experimental units per treatment.	Population representativeness.
Direct control	Keep known conditions consistent.	Removal of every source of variation.
Blinding	Limit effects of knowing the assigned treatment.	That all treatments can realistically be concealed.

State which feature is used and explain its role in the particular experiment.

What did Oregon's lottery actually assign?

In 2008, Oregon drew names from a waiting list to decide which households could apply for Medicaid. Eligibility still had to be established. To study the effect of the invitation, would you compare households invited and not invited, or adults who eventually had insurance and those who did not? Only the invitation was assigned by lottery. Comparing insurance holders with everyone else could mix the effect of coverage with differences in who completed the application. The waiting list also limits which population the comparison describes. Source: Finkelstein et al., Quarterly Journal of Economics (2012), doi:10.1093/qje/qjs020.

KEY IDEA

Random assignment creates a fair treatment comparison; random sampling answers a different question about generalization.

WATCH FOR

Repeated measurements of one unit do not replace independent experimental units.

OPENSTAX WORKED EXAMPLE

Identify the design of an aspirin experiment

In an OpenStax study scenario, 400 adults aged 50-84 are recruited and randomly divided into two equal groups. One group takes aspirin daily for three years; the other takes a placebo. Participants do not know which pill they receive. Researchers record whether each participant has a heart attack during the study. Identify the experimental units, explanatory variable, response variable, and treatments. Explain what the random assignment supports and whether the description establishes double blinding.

1. Each of the 400 adults is an experimental unit because treatment is assigned to the individual. The explanatory variable is assigned pill type. The two treatments are daily aspirin and daily placebo.

2. The response variable is whether an individual has a heart attack during the three-year period. Comparing the groups could use their sample proportions experiencing a heart attack.

3. The random assignment helps balance other characteristics between treatment groups in the long run. It permits assessment of whether an observed difference is consistent with chance assignment and can support a causal conclusion when the evidence is convincing.

4. The participants are blinded. The description does not say whether personnel who interact with participants or assess outcomes know treatment assignments, so double blinding cannot be concluded.

5. Recruitment is stated, but random sampling from all adults aged 50-84 is not. A causal result for the study does not automatically justify a population-wide estimate for every adult in that age range.

Conclusion. The design is a randomized, placebo-controlled experiment with participant blinding. It can support a causal comparison, subject to implementation and evidence, while its recruitment limits generalization.

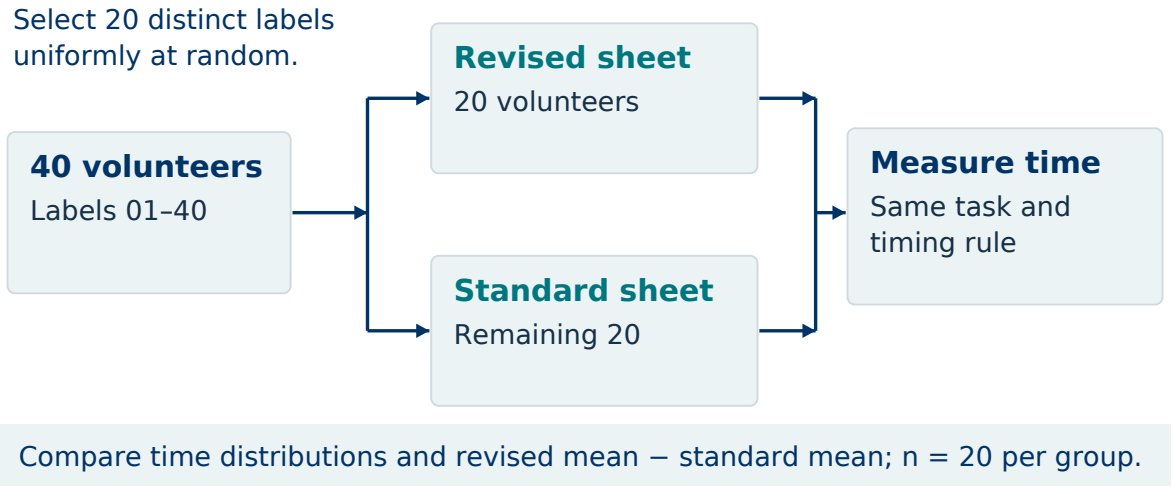
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.4, Example 1.19, printed p. 35. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Does a revised instruction sheet reduce completion time?

Forty volunteers will complete the same unfamiliar assembly task. Investigators want to compare standard instructions with revised instructions. Design a completely randomized experiment.

A complete randomized allocation



Illustrative worked design. The random choice determines treatment assignment; all other task conditions are held consistent.

1. The experimental units are the 40 volunteers, each assigned one instruction sheet. The factor is instruction format with two levels, standard and revised. The response is task completion time in minutes using a fixed stopping rule.

2. Label volunteers 01 through 40 and select 20 distinct labels uniformly at random for the revised sheet. Assign the remaining 20 to the standard sheet. This describes a reproducible chance assignment, not an alternating or ability-based split.

3. Give everyone the same materials, task, working time policy, and room conditions. Use standardized directions that do not reveal a preferred treatment. This direct control reduces avoidable variation.

4. Use 20 units per treatment for replication. Mask the person who records completion time to sheet type when feasible, or use an objective time stamp to limit assessor bias. Participants may recognize different sheets, so do not claim full double blinding without a workable procedure.

5. Compare group completion-time distributions and their mean difference. Use a randomization analysis to determine whether the observed difference is surprising under no treatment effect.

Conclusion. The random assignment supports a causal comparison if the result is convincing, but the volunteer sample limits broad population generalization.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.4, Example 1.19, printed p. 35. Adapted for this course; source data retained unless otherwise stated.

The source exercise is presented as a study scenario, not a verified historical event. Equal allocation is explicit. Recruitment is distinguished from population sampling; random assignment does not guarantee exact balance or a treatment effect.

S02.04 | Blocking, Matched Pairs, and Scope of Inference

CED 1.10, 1.13

- Choose and describe a randomized block or matched-pairs experiment.
- Explain why blocking can improve precision without replacing random assignment.
- Connect random sampling and assignment to appropriate conclusions.
- Describe randomization evidence for a treatment comparison.

CONCEPT AND INTERPRETATION

Block on a relevant source of variation

A randomized block design groups experimental units using a characteristic expected to affect the response, then randomly assigns treatments within each block. Every treatment occurs within each block. Units in a block should be similar with respect to the blocking variable so treatment comparisons are less affected by that source of variation. For a plant study, blocks might be initial size ranges. Blocking does not make the treatment groups observational: random assignment still occurs inside blocks. Nor is a block the same as a treatment; the block is an existing source of variation being accounted for.

Matched pairs is a special block design

A matched-pairs design compares two treatments using blocks of two similar units or repeated treatment on the same unit. With two matched people, randomly assign one to each treatment within the pair. When one person receives both treatments, randomize treatment order and consider learning, fatigue, carryover, and whether a washout or separate equivalent task is needed. Analyze within-pair differences rather than pretending the two sets are independent. Define the subtraction order before analysis; new minus standard and standard minus new have opposite signs but describe the same comparison when interpreted consistently. In a repeated-condition study, the two measurements from one person usually share that person's ability or baseline level. Treating them as independent loses that connection. A difference is a new variable with one value per pair, and it must carry the original measurement units.

Choose a design for a reason

Blocking is useful when the blocking variable is strongly related to the response and is available before treatment. A completely randomized design may be simpler when no useful blocking variable is known. Too many tiny blocks can complicate implementation, and a matched-pairs design can be unsuitable when treatment effects persist or when the first task changes performance on the second. A design justification should name the source of response variation and explain how comparison within blocks reduces its impact. "Blocking is more accurate" is too vague; state why this block matters for this outcome.

Separate two kinds of randomization

Random selection determines who enters a study and supports generalization to the population represented by the frame when participation and measurement are adequate. Random assignment determines who receives each treatment and supports cause-and-effect conclusions when the experiment is well designed and provides convincing evidence. A random sample without treatment assignment supports population association, not causation. An assigned experiment with volunteers supports a causal comparison for participants and may inform similar populations, but it does not automatically justify a population-wide probability claim. With neither form of randomization, both generalization and causal interpretation are limited.

Two randomization decisions support different conclusions

Random population sample?	Random treatment assignment?	What the design supports
Yes	No	Population association or description; causation is not established by sampling alone.
No	Yes	A causal comparison for the experiment with convincing evidence; broad population generalization is limited.
Yes	Yes	Population generalization and a causal comparison, given sound implementation and convincing evidence.
No	No	Descriptive findings for observed units; both broader generalization and causal interpretation need strong caution.

Coverage, response, ethical implementation, measurement, and statistical evidence still matter in every row.

Use randomization to assess chance

Even fair treatment groups can differ by chance. To model a no-treatment-effect explanation in a completely randomized experiment, keep the observed responses fixed and repeatedly shuffle treatment labels in the same group sizes used originally. For each shuffle, calculate the same comparison statistic, such as the mean of new minus standard. The simulated distribution shows differences compatible with chance assignment under that null model. For a prespecified lower-time alternative, count simulated differences at least as negative as observed and divide by the number of simulations. For a paired experiment, randomization must respect pairs; do not shuffle all labels freely across unrelated units.

KEY IDEA

Block first on a relevant characteristic, then randomize treatments within every block.

WATCH FOR

Random sampling supports generalization; random assignment supports a causal comparison. Neither guarantees a convincing result.

OPENSTAX WORKED EXAMPLE

Why have each participant complete both conditions?

In the maze-study scenario described in OpenStax, each participant completes pencil-and-paper mazes three times while wearing a floral-scented mask and three times while wearing an unscented mask. Participants are randomly assigned to use the floral mask during either the first three or the last three trials. Researchers record completion times. Identify the paired structure, explain why treatment order is randomized, and identify a feasible form of blinding.

1. The experimental unit for the treatment-order assignment is one participant. Each participant experiences both mask conditions, so the comparison is paired within person. The explanatory variable is mask scent and the response is maze-completion time.

2. Summarize each participant's three times under each condition, then calculate one floral-minus-unscented difference per participant. A negative difference indicates faster completion under the floral condition.

3. Practice can shorten later times and fatigue can lengthen them. Randomizing the order prevents one mask condition from always occurring later. Equivalent mazes and consistent instructions also help control task differences.

4. Participants can usually distinguish a scented mask from an unscented one, so participant blinding is not feasible as described. Personnel analyzing coded timing records could be blinded to which code represents each condition.

5. Randomized order does not guarantee that learning, fatigue, or carryover are perfectly balanced in the realized sample. Consider those possibilities when interpreting the comparison.

Conclusion. The matched-pairs comparison controls for stable differences in participants' maze-solving ability. Randomized order helps separate mask condition from practice and fatigue, but it does not guarantee that every other influence disappears.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.4, Example 1.20, printed p. 35-36. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Compare two display formats within experience blocks

An invented experiment has 24 volunteers: 12 experienced and 12 inexperienced users. It compares a standard and a revised display on task time. The revised group averages 3 minutes faster. Illustrative randomization output contains 27 of 1,000 differences at least as favorable to the revised display.

Block on experience, then randomize within blocks

02



Both treatments occur in each block. Reassign labels within blocks in simulation.

Illustrative worked design: 24 volunteers, 12 per experience block, with 6 assigned to each treatment inside each block.

1. Experience plausibly affects task time, so create two blocks before assignment. Within each 12-person block, randomly choose six for the revised display and assign the remaining six to standard.

2. Use the same task, timing rule, and equipment for everyone. Each treatment has six units in each experience block, preventing a difference in block composition from being mistaken for a display effect.

3. Define the statistic as revised mean minus standard mean. The observed value is -3 minutes. The direction was specified as faster before examining the results.

4. Under a no-effect model, shuffle assignments within each experience block while preserving six per treatment in each block. Recompute the overall revised-minus-standard mean difference for each shuffle.

5. The illustrative estimated tail proportion is $27/1000=0.027$. Such a result is relatively unusual under the stated chance model and provides evidence that the revised display reduces time for these volunteers. This output is supplied for interpretation, not claimed as an actual simulation run.

Conclusion. The revised display was 3 minutes faster on average for these volunteers. Only 27 of 1,000 within-block shuffles produced a difference this low or lower. The experiment supports a display effect here; the volunteers do not represent all users.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 1.4, Example 1.20, printed p. 35-36. Adapted for this course; source data retained unless otherwise stated.

The source procedure is retained and presented as a source-described scenario. Its claim that random assignment eliminates lurking-variable problems is corrected. The within-person mean difference and feasible assessor masking are stated explicitly.

Unit synthesis

1. Describe an SRS of 20 from a 300-person roster and explain why taking one randomly selected 20-person club is generally not an SRS of people.

2. Distinguish undercoverage, nonresponse, and response bias using a survey about library use.

3. Explain why self-selected tutoring and exam score can be confounded by motivation.

4. Design a randomized block experiment with 30 novice and 30 experienced users comparing two interfaces.

5. An experiment uses volunteers but random treatment assignment. State a defensible conclusion if convincing evidence favors the treatment.

6. In a paired experiment, what must a no-effect randomization preserve?

Unit 03 | Probability

Probability connects a chance model to the patterns expected from repeated observations. This unit begins with simulation so students must define a trial, an event, and a defensible model before using formulas. It then develops complements, overlapping events, conditional probabilities, independence, and multi-stage trees. Calculations are paired with interpretations and checks on the model. The same habits return in inference: specify the assumption, determine how unusual a result would be, and limit the conclusion to what the study supports. All supplied classroom datasets and model rates are illustrative rather than factual claims about the school or region.

CED topics: 2.3, 2.4, 2.5, 2.6, 2.7

IN THIS UNIT

Lesson	Focus
S03.01	Chance models and reproducible simulation
S03.02	Events, complements, and overlapping categories
S03.03	Conditional probability and independence
S03.04	Multi-stage chance processes and probability trees



Observing activity along Bear Creek

LOCAL QUESTION

If a model gives each passing rider a fixed chance of wearing a helmet, what assumptions would a five-rider simulation require?

S03.01 | Chance models and reproducible simulation

CED 2.3

- Interpret probability as a long-run relative frequency.
- Design and critique a reproducible simulation.
- Use simulated results to assess whether an observed result is unusual under a specified model.



One attempt, an uncertain result

CONCEPT AND INTERPRETATION

What probability describes

A random process has uncertain individual outcomes but may have a stable pattern across many repetitions. An outcome is one possible result; an event is a collection of outcomes. For example, one trial might be five attempted free throws, the outcome might be the sequence miss-make-make-miss-make, and the event might be at least four makes. Define the trial before counting outcomes. The probability of an event is the proportion of trials in which it would occur in the long run under the stated model. A probability of 0.60 does not promise six successes in the next ten trials.

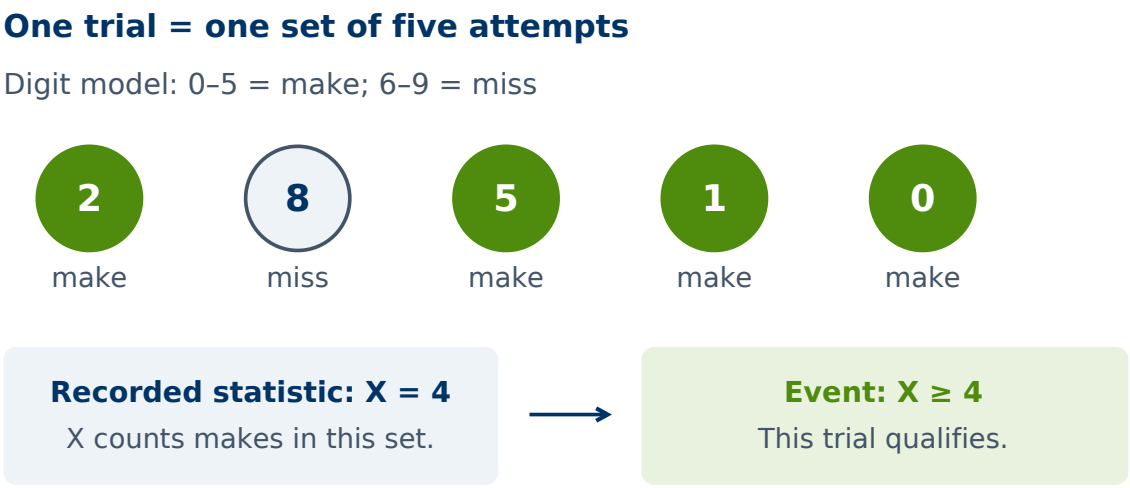
The law of large numbers

For independent repetitions under the same chance model, the relative frequency tends to settle near the underlying probability as the number of trials grows. It need not move steadily closer after every repetition. A long run of failures does not make success due on the next independent trial. More trials improve the stability of a simulation estimate, but they do not fix an incorrect model. In a real shooting study, a changing skill level or fatigue could make independent trials with a constant success probability questionable.

Build the model before generating results

A complete simulation states the question, assumptions, random device, outcome assignment, one trial, recorded result, repetition rule, and final calculation. To model a 60% success rate with digits, assign 0-5 to success and 6-9 to failure. Generate digits independently with replacement. If the question concerns five attempts, use five digits for one trial and record its number of successes. Repeat the entire five-attempt process many times. Counting individual successful attempts estimates a different probability from counting five-attempt sets containing at least four successes. An outcome assignment must match the intended probabilities. For a fair coin, five of ten digits represent heads. For a two-toss event, two new independent digits form one trial. Write an example trial and its recorded result before running a long simulation so that the estimated event is unambiguous.

Define the trial before simulating



Illustrative five-digit trial. The digit model has success probability 0.60; one recorded statistic summarizes all five attempts.

Turn a simulation into evidence

Estimate an event probability by dividing the number of trials meeting the event definition by the total number of trials. Report both counts, not just a decimal. To assess a claim, simulate under that claim and define results at least as extreme in the relevant direction before looking at the simulation. A small estimated probability makes the observed result surprising under the model. This is evidence against the model, not proof that it is false. Later units formalize the decision with a significance level and p-value.

Separate real data from simulated data

Observed data come from the process being studied. Simulated data come from a model chosen by the analyst. A simulation can estimate a model probability even when no physical experiment is possible, but it cannot establish that the real process satisfies the model assumptions. Label demonstration data as illustrative. Different groups will get different simulated estimates; reasonable variation is expected. Pool compatible repetitions to improve precision, while keeping each trial and event definition unchanged.

KEY IDEA

A simulation answers the question encoded in its trial, event, and assumptions.

WATCH FOR

More repetitions reduce simulation variability; they do not remove bias in the model.

OPENSTAX WORKED EXAMPLE

Two coin tosses: exact model and simulation

OpenStax considers two tosses of a fair coin. Assume the tosses are independent. Determine the probability of at most one tail. Then describe how a simulation could estimate that same probability and explain what one simulated trial must represent.

1. Distinguish the first and second toss. The four equally likely outcomes are HH, HT, TH, and TT. At most one tail includes zero tails, so the event consists of HH, HT, and TH.

2. Three of the four equally likely outcomes satisfy the event. The model probability is $3/4=0.75$.

3. For a digit simulation, assign 0-4 to heads and 5-9 to tails. Generate two independent random digits to represent one complete trial, and record whether at most one digit represents a tail.

4. Repeat the two-digit trial many times. Divide the number of qualifying two-toss trials by the total number of two-toss trials. Do not divide by the number of individual tosses.

5. The simulated fraction should tend to be near 0.75 for a large number of repetitions, but it need not equal 0.75 or move steadily closer after every additional trial.

Conclusion. Under the independent fair-coin model, the probability of at most one tail in two tosses is 0.75. A simulation estimates this probability by counting complete pairs of tosses.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.2, Example 3.6, printed p. 174. Adapted for this course; source data retained unless otherwise stated.

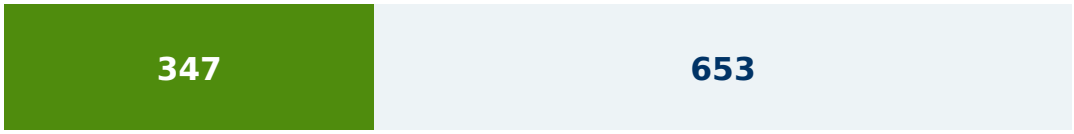
LOCAL APPLICATION

At least four makes

An illustrative model gives each free throw an independent 0.60 chance of success. Estimate the chance of at least four makes in five attempts. A simulation of 1,000 five-attempt sets produces 347 sets with four or five makes.

Count qualifying sets

1,000 simulated sets of five attempts



347 sets: four or five makes
653 sets: fewer than four makes

Estimated probability

347 / 1,000 = 0.347

The denominator is 1,000 complete five-attempt sets. The bar shows the supplied illustrative simulation results, not the exact model probability.

1. Define one trial as five independent attempts. Let X be the number of makes in that trial; the event is X at least 4.

2. Assign digits 0-5 to make and 6-9 to miss. Generate five independent digits, count the makes, and record whether the event occurred. Repeat 1,000 times without changing the assignment.

3. The estimated event probability is 347/1,000 = 0.347. In many five-attempt sets under this model, roughly 35% would contain at least four makes.

4. The estimate concerns sets of five, not the success rate of individual shots. An observed set of four makes would not be particularly surprising under this model.

5. A different valid simulation need not give 0.347. The exact model probability, developed with the binomial distribution later, is 0.33696, which provides a useful check.

Conclusion. The simulation supports an estimated probability near one third, subject to the independent, constant -0.60 model.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.2, Example 3.6, printed p. 174. Adapted for this course; source data retained unless otherwise stated.

The exact two-coin model is retained. A reproducible random-digit simulation is added to teach the AP simulation objective, which the source example itself does not fully develop.

S03.02 | Events, complements, and overlapping categories

CED 2.4, 2.5, 2.7

- Calculate probabilities from equally likely outcomes or frequency tables.
- Use complements and the general addition rule.
- Justify whether events are mutually exclusive.

CONCEPT AND INTERPRETATION

Define the sample space

The sample space contains every possible outcome of a random process, with outcomes that do not overlap. Its probability is 1. An event is any collection of those outcomes. If all elementary outcomes are equally likely, divide the number in the event by the total number. The equal-likelihood condition matters. The possible sums from rolling two fair dice are 2 through 12, but these eleven sums are not equally likely. The 36 ordered pairs of die faces are equally likely, so they provide an appropriate counting basis.

Complements make some questions easier

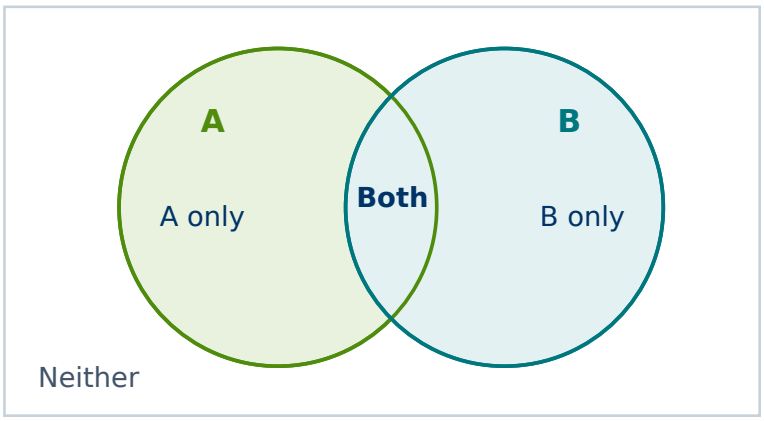
The complement of event A consists of outcomes outside A. Write it as A complement or A^c. Because A and its complement cover the sample space without overlap, $P(A^c) = 1 - P(A)$. Translate wording carefully: the complement of at least one is none, the complement of more than three is at most three, and the complement of at least three is fewer than three. This translation often prevents a long list of cases. A probability must lie between 0 and 1; a result outside that range signals a counting or arithmetic error.

And, or, and overlap

The intersection, written $A \cap B$, means both A and B occur. The union, written $A \cup B$, means A or B or both occur. Statistical or is inclusive unless the problem explicitly says exactly one. Adding P(A) and P(B) counts outcomes in their overlap twice. Subtract one copy: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. In a table, calculate each probability using the full population total unless the question conditions on a subgroup. A Venn diagram can organize counts, but its circle areas need not be proportional unless designed that way. An alternative calculation splits a Venn diagram into four disjoint regions: A only, both, B only, and neither. Fill the overlap first, subtract it from each marginal total, and use the remaining probability for neither. The four regions must total 1, providing a check on the addition-rule calculation.

Account for every outcome once

Four nonoverlapping regions exhaust the sample space.



Union A ∪ B:
A only + Both
+ B only

P(A) + P(B)
counts Both twice.
Subtract it once.

Schematic Venn diagram: region areas are not proportional to probabilities. Inclusive “or” contains all three inside regions.

Mutually exclusive is a structural claim

Two events are mutually exclusive, or disjoint, when they cannot occur on the same trial. Their intersection is empty, so its probability is zero. A single card cannot be both a heart and a club. A single card can be both a heart and a face card, so those events are not disjoint. When events are disjoint, the addition rule simplifies to P(A or B) = P(A) + P(B). Do not assume disjointness because the events have different names or because their observed overlap happens to be small.

Check the denominator and the claim

A probability calculated from a complete set of recorded cases describes random selection from those cases. Generalizing it to a larger population requires a defensible sampling process. Counts are additive; percentages are additive only when they share the same denominator and overlap is handled correctly. Before calculating, identify the unit selected, write the event in words, and check that every counted case satisfies the event. After calculating, interpret the result in the original context rather than treating it as an isolated decimal.

KEY IDEA

For an inclusive or, add both probabilities and subtract their overlap once.

WATCH FOR

Different labels do not imply mutually exclusive events.

OPENSTAX WORKED EXAMPLE

Club membership and part-time work

In the OpenStax college scenario, 40% of students belong to a club, 50% work part time, and 5% do both. One student is selected at random. Calculate the probability that the student belongs to a club or works part time, the probability of neither, and the probability of exactly one of the two activities.

1. Let C denote club membership and W denote part-time work. The overlap is P(C and W)=0.05. The word or includes students who do both.

2. Apply the addition rule: P(C or W)=0.40+0.50−0.05=0.85. The overlap was included in both marginal probabilities, so subtract it once.

3. Neither is the complement of the union, giving 1−0.85=0.15.

4. Club only has probability 0.40−0.05=0.35 and work only has probability 0.50−0.05=0.45. Exactly one therefore has probability 0.35+0.45=0.80.

5. The four disjoint regions have probabilities 0.35, 0.05, 0.45, and 0.15. Their total is 1, which checks the allocation.

Conclusion. For this stated distribution, 85% of students do at least one of the activities, 80% do exactly one, and 15% do neither. Exactly one excludes the 5% who do both.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.5, Example 3.29, printed p. 194-195. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Two activities, one student

An illustrative roster of 200 students shows 80 in a sport, 70 in a music group, and 30 in both. Select one roster student uniformly at random. Find the chance the student does at least one activity, exactly one activity, and neither.

Check the union against a complete table

Illustrative roster: 200 students

	Music	No music	Total
Sport	30	50	80
No sport	40	80	120
Total	70	130	200



The three cells for at least one activity total 120; the remaining cell contains 80 students. The bar widths match those counts.

1. Let A be participation in a sport and B participation in music. The unit selected is one of the 200 roster students; each has probability 1/200.

2. The overlap is 30. Therefore sport only is 80 - 30 = 50, music only is 70 - 30 = 40, and both is 30.

3. At least one includes all three groups: 50 + 40 + 30 = 120. Equivalently, 80 + 70 - 30 = 120. Its probability is 120/200 = 0.60.

4. Exactly one excludes the overlap: (50 + 40)/200 = 0.45. Neither is the complement of at least one: 1 - 0.60 = 0.40.

5. The events are not mutually exclusive because 30 students satisfy both. The two original activity percentages cannot simply be added.

Conclusion. For this roster, 60% participate in at least one activity, 45% in exactly one, and 40% in neither.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.5, Example 3.29, printed p. 194-195. Adapted for this course; source data retained unless otherwise stated.

Explanations expand the source reasoning and use AP terminology. The local application uses a separate illustrative data set.

S03.03 | Conditional probability and independence

CED 2.6, 2.7

- Calculate conditional probabilities with a correct restricted denominator.
- Distinguish joint, marginal, and conditional probabilities.
- Check independence with numerical evidence.

CONCEPT AND INTERPRETATION

Conditioning changes the reference group

A conditional probability answers a question after information restricts the relevant cases. Write $P(A \mid B)$ for the probability of A given B. The cases satisfying B become the new reference group. If 54 of 90 walkers support a proposed schedule, the support probability given walking is 54/90. Dividing 54 by all 200 students instead answers the joint question, the chance a randomly selected student both walks and supports the schedule. Read the expression after the vertical bar first to identify the denominator. The phrase given B need not describe an event that happened earlier in time. It describes information being supplied. For example, learning that a visitor checked out a DVD restricts the relevant library visitors even if the book and DVD were checked out together.

The conditional probability formula

Provided P(B)>0, $P(A \mid B) = P(A \cap B)/P(B)$. Both numerator and denominator on the right refer to the original sample space, so their common total cancels when counts are used. Conditioning is directional: P(A given B) generally differs from P(B given A). A high proportion of walkers might support the schedule even though relatively few supporters walk. A statement about conditional probability is not a causal statement; a subgroup difference may reflect other variables or the way the data were collected.

Independence means unchanged probability

Events A and B are independent when knowing one occurred does not change the probability of the other. For events with positive probabilities, check whether P(A given B)=P(A), whether P(B given A)=P(B), or whether P(A and B)=P(A)P(B). One valid equality establishes independence in a fully specified probability model. In a real sample, proportions often differ slightly by chance, so an observed inequality is descriptive evidence of association, not by itself a formal conclusion about independence in the population. The numerical comparison must use probabilities for the same events. Comparing P(A given B) with P(B) does not test the independence condition.

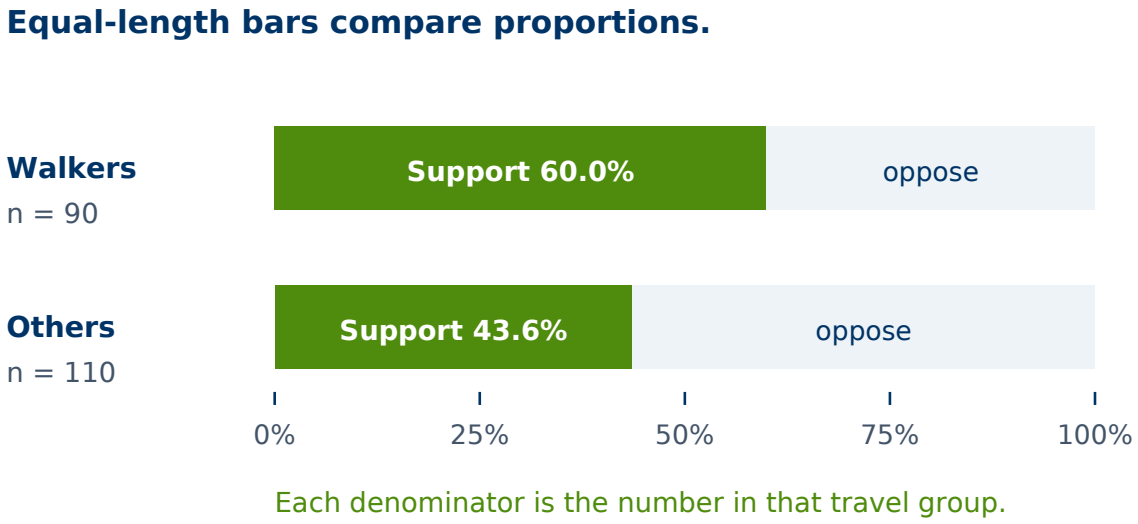
Disjoint and independent answer different questions

Disjoint events cannot occur together; independent events do not alter one another's probabilities. If A and B each have positive probability and are disjoint, then learning that B occurred makes the probability of A zero. That differs from the original positive P(A), so they are dependent. The words unrelated and separate are too vague to replace either mathematical definition. Always identify whether the claim concerns simultaneous occurrence or a change in conditional probability.

Use conditional comparisons responsibly

Compare proportions within each group when group sizes differ. Counts alone can make a large group appear to have a stronger tendency just because it contains more people. In a two-way table, each row or column of conditional percentages should sum to 100% within the chosen denominator. State the condition in the interpretation: among walkers, 60% support the schedule. For a claim about the whole school, also evaluate coverage, random selection, nonresponse, and measurement. Probability calculations cannot repair an unrepresentative survey.

Compare rates within groups



Each full bar represents 100% of its own group, although group sizes differ. Support is 60.0% among walkers and about 43.6% among other commuters.

KEY IDEA

The event after given determines the denominator.

WATCH FOR

P(A given B) and P(B given A) usually answer different questions.

OPENSTAX WORKED EXAMPLE

Independence and mutually exclusive events

Let C be the event that a randomly selected college student takes an English class and D the event that the student takes a speech class. Suppose P(C)=0.75, P(D)=0.30, P(C given D)=0.75, and P(C and D)=0.225. Determine whether C and D are independent and whether they are mutually exclusive. Calculate P(D given C). Justify each answer numerically.

1. The probability of taking English among speech students is P(C given D)=0.75. It equals the marginal P(C)=0.75, so knowing the student takes speech does not change the probability of taking English. The events are independent under the stated model.

2. The product check gives P(C)P(D)=0.75(0.30)=0.225, equal to P(C and D). This is an equivalent check of independence.

3. The events are not mutually exclusive because their intersection has positive probability, 0.225. Some students take both classes.

4. The reverse conditional is P(D given C)=P(C and D)/P(C)=0.225/0.75=0.30. The denominator is the probability of the condition, English enrollment.

Conclusion. English and speech enrollment are independent in the stated model, but they are not mutually exclusive. Of the English students, 30% also take speech.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.2, Example 3.10, printed p. 177. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

Support within travel groups

An illustrative table has 90 walkers: 54 support a schedule change and 36 oppose it. It also has 110 other commuters: 48 support and 62 oppose. Select one of the 200 respondents uniformly. Compare support within the groups and assess independence for this respondent distribution.

The condition selects the reference group

Use the reference group named after “given.”

	Support	Oppose	Total
Walkers	54	36	90
Others	48	62	110
Total	102	98	200

Given walker: use the 90-person row.

P(support | walker) = 54 / 90 = 0.60
P(walker | support) = 54 / 102 ≈ 0.529

Illustrative worked data. The highlighted row is the denominator group for support given walking; reversing the condition changes the denominator.

1. Let W mean walker and S mean supporter. The totals are 90 walkers, 110 other commuters, 102 supporters, and 98 opponents.

2. P(S given W)=54/90=.60, whereas P(S given not W)=48/110≈.4364. The support proportion is larger among walkers.

3. The overall P(S)=102/200=.51. Because .60 differs from .51, support and walking are not independent in this respondent distribution.

4. P(S and W)=54/200=.27. Independence would require P(S)P(W)=.51(.45)=.2295, providing the same descriptive conclusion.

5. The reversed conditional is P(W given S)=54/102≈.5294. Its denominator is supporters, so it is not the same as .60.

6. The comparison uses proportions rather than the raw support counts because the two travel groups have different sizes.

Conclusion. Among these respondents, 54 of 90 walkers (60%) supported the change, compared with 48 of 110 other commuters (about 44%). Walking and support were associated in this group. The survey cannot establish that walking caused support.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.2, Example 3.10, printed p. 177. Adapted for this course; source data retained unless otherwise stated.

Explanations expand the source reasoning and use AP terminology. The local application uses a separate illustrative data set.

S03.04 | Multi-stage chance processes and probability trees

CED 2.6, 2.7

- Use the general multiplication rule for sequential events.
- Combine mutually exclusive paths in a probability tree.
- Distinguish sampling with and without replacement.

CONCEPT AND INTERPRETATION

Multiply probabilities along a path

A multi-stage outcome can be represented by a path through a tree. Each branch gives the probability of the next result conditional on the path already followed. The general multiplication rule is $P(A \cap B) = P(A)P(B \mid A)$. It works whether or not the stages are independent. For three stages, continue with the probability of the third result conditional on both earlier results. Branch probabilities leaving a single node sum to 1. Label branches with conditions rather than assuming every stage uses the original probability.

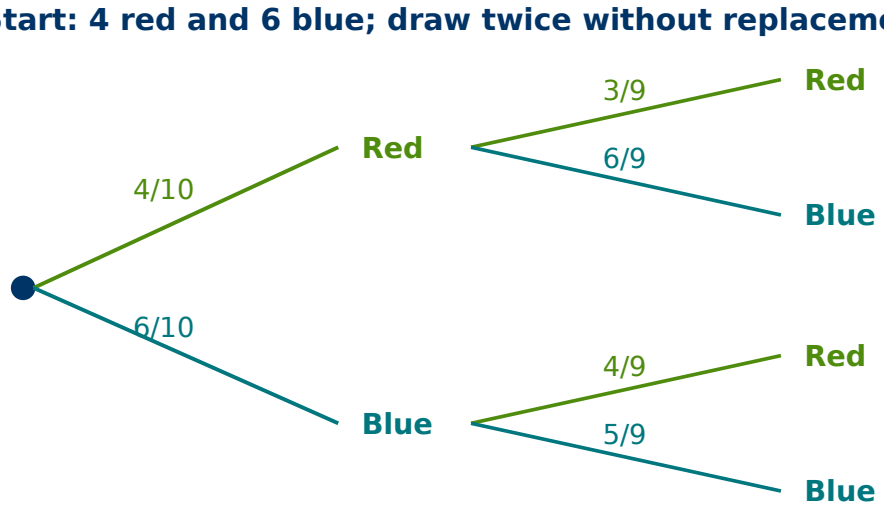
Add probabilities across alternative paths

Different complete paths represent mutually exclusive outcomes. If an event can happen through several paths, multiply along each relevant path, then add those path probabilities. Exactly one success in two trials includes success-failure and failure-success. At least one success also includes success-success. When there are many paths, a complement may be easier: P(at least one)=1-P(none). Count all relevant orders; a sequence of red then blue is a different outcome from blue then red even if both satisfy exactly one red. To interpret a tree, separate stages from alternatives. Red then blue is one complete path requiring both stages, so multiply its conditional branch probabilities. Exactly one red can occur by either red-blue or blue-red, so add those two disjoint path probabilities. Branch probabilities leaving the same node must total 1.

Replacement affects the model

When a selected item is replaced and the population is thoroughly mixed, successive draws can be modeled as independent with unchanged probabilities. Without replacement, the available counts change. From a bag with 4 red and 6 blue counters, a red first draw leaves 3 red among 9 counters; a blue first draw leaves 4 red among 9. The second-stage probabilities depend on the first result. For sampling a small fraction of a large population, an independence approximation can be useful, but an exact small-bag calculation must use the changed counts.

Update the second-stage probabilities



Exact probability tree for the investigation's four-red, six-blue bag. Branches leaving each node sum to 1.

Conditioning can reverse the tree question

A tree may describe probabilities of an observed signal given a true category. The reverse question, category given signal, requires combining paths first. Find the joint probability of the category and signal, then divide by the total probability of the signal across all categories. This is simply the conditional probability formula. An accurate screening process can still produce many false alerts if the underlying condition is rare. The starting category frequencies are part of the calculation; ignoring them changes the question.

A tree is a model, not an explanation of cause

Trees organize conditional chance information but do not by themselves establish why probabilities differ. The branch values may be stipulated model assumptions, estimated from historical data, or measured in a study. Say which. Verify that complete path probabilities sum to 1 and that a requested event is represented by all and only the relevant paths. Interpret the final probability for a clearly defined random unit. Avoid turning a model-based probability into certainty about a particular individual item.

KEY IDEA

Multiply along each path; add across disjoint paths.

WATCH FOR

Without replacement, update both the numerator and the denominator.

OPENSTAX WORKED EXAMPLE

Two draws without replacement

An urn contains 3 red and 8 blue marbles. Two marbles are selected at random, one at a time, without replacement. Calculate the probability that exactly one of the two marbles is red and the probability that the second marble is red given that the first is blue.

1. The first draw is red with probability 3/11 and blue with probability 8/11. After a red first draw, 2 red and 8 blue remain among 10 marbles. After a blue first draw, 3 red and 7 blue remain among 10.

2. Exactly one red has two disjoint paths. For red then blue, the path probability is $(3/11)(8/10)=24/110$. For blue then red, it is $(8/11)(3/10)=24/110$.

3. Add the two alternative paths: $P(\text{exactly one red})=48/110=24/55\approx0.4364$. Multiplication applies along a path; addition combines disjoint complete paths.

4. Given that the first marble is blue, only the branch from a blue first draw matters. Three red marbles remain among 10, so $P(\text{second red given first blue})=3/10=0.30$.

5. As a check, the four path probabilities are RR=6/110, RB=24/110, BR=24/110, and BB=56/110. They total 1.

Conclusion. The probability of exactly one red marble is about 43.6%. Given a blue first marble, the probability the second is red is 30%, reflecting the changed contents after the first draw.

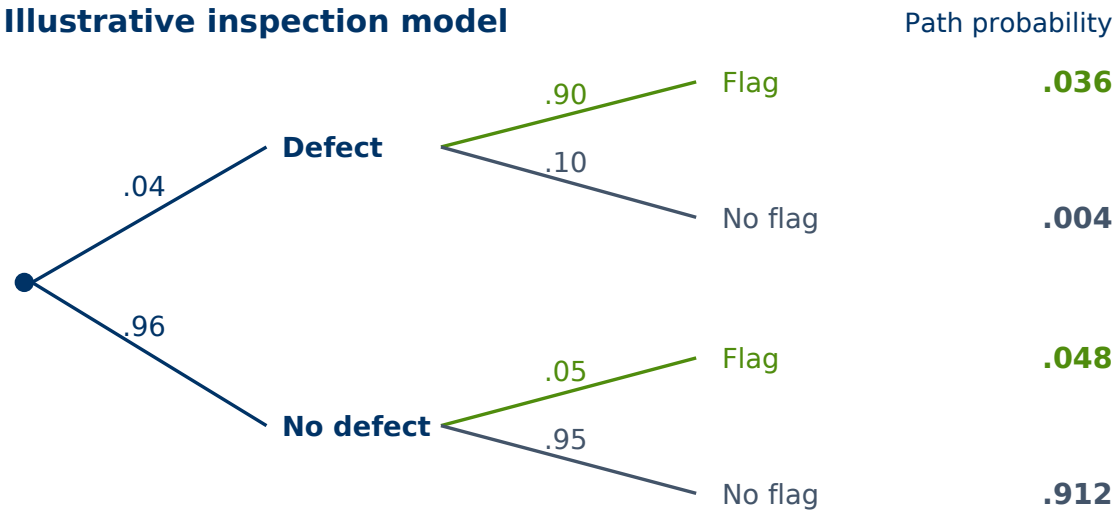
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.5, Example 3.25, printed p. 190-192. Adapted for this course; source data retained unless otherwise stated.

LOCAL APPLICATION

A quality-control alert

In an illustrative factory model, 4% of items have a defect. An inspection flags 90% of defective items and 5% of nondefective items. For a randomly selected item, find the probability of a flag and the probability of a defect given a flag.

Multiply along paths, add flagged alternatives



Flagged paths: .036 + .048 = .084. Defect given flag: .036 / .084 ≈ .429.

The four path probabilities sum to 1. The two green flagged paths determine the denominator when conditioning on a flag.

1. Let D mean defective and F mean flagged. The starting branches are P(D)=.04 and P(not D)=.96. The second-stage probabilities are P(F given D)=.90 and P(F given not D)=.05.

2. The defective-and-flagged path has probability .04(.90)=.036. The nondefective-and-flagged path has probability .96(.05)=.048.

3. These paths are disjoint and cover all flagged items, so P(F)=.036+.048=.084. About 8.4% of items would be flagged under the model.

4. To reverse the condition, P(D given F)=P(D and F)/P(F)=.036/.084≈.4286.

5. In an imagined 10,000-item population, about 360 would be defective-and-flagged and 480 nondefective-and-flagged. Only 360 of 840 flagged items would be defective.

6. The false-flag path matters because nondefective items are much more common in the starting population.

Conclusion. About 42.9% of flagged items would have defects, despite the 90% detection rate. The two conditional probabilities answer different questions.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 3.5, Example 3.25, printed p. 190-192. Adapted for this course; source data retained unless otherwise stated.

The source's red/blue model and without-replacement assumptions are retained. Ambiguous repeated event letters are replaced by explicit first-draw and second-draw wording.

Unit synthesis

1. Define one simulation trial for the chance of at least two successes in four independent attempts with success probability .30.

2. P(A)=.45, P(B)=.35, and P(A and B)=.15. Find P(A or B), P(neither), and P(A given B).

3. Can the events in the previous question be independent or disjoint?

4. From 4 green and 5 purple counters, find P(one of each) in two draws without replacement.

5. A simulation gives 9 extreme results in 3,000 repetitions under a claim. Interpret this and give one limitation.

6. Explain why equal observed counts in two differently sized groups do not establish independence.

Unit 04 | Random Variables

A chance model becomes more useful when its outcomes are organized as a numerical random variable. Students construct discrete and cumulative distributions, interpret probability-weighted averages, and measure variability. They then study two major models: the binomial distribution for counts and the normal distribution for continuous measurements. Conditions are part of each model, not an afterthought. Technology supports exact probabilities and inverse calculations, while sketches, event notation, and contextual interpretation make the work reviewable. The examples use illustrative model values. Geometric distributions and a standalone general unit on combinations of random variables are outside this core sequence. Unit 5 explicitly teaches the means and standard deviations of differences of independent sample proportions and sample means, including why their variances add.

CED topics: 2.8, 2.9, 2.10, 2.11

IN THIS UNIT

Lesson	Focus
S04.01	Random variables, distributions, and expected value
S04.02	The binomial model
S04.03	Continuous models and the normal distribution
S04.04	Normal probabilities, percentiles, and cutoffs



Local question

Suppose X counts hikers who pass a fixed trail point during one half hour. What values can X take? What would you need to know before assigning probabilities to those values?

Table Rock and the oak-lined trail

S04.01 | Random variables, distributions, and expected value

CED 2.8, 2.9

- Construct and validate a discrete probability distribution.
- Calculate cumulative probabilities, expected value, and standard deviation.
- Interpret the center and variability of a random variable in context.

CONCEPT AND INTERPRETATION

From an outcome to a numerical variable

A random variable assigns a number to each outcome of a random process. Let X be the number of late buses among three selected trips. Individual sequences such as late-on time-late and on time-late-late are different outcomes, but both give X=2. The possible values of X are 0, 1, 2, 3. A discrete random variable has separated possible values that can be listed or counted. Define the variable with its units and time frame before writing a distribution; number of late trips differs from total minutes late.

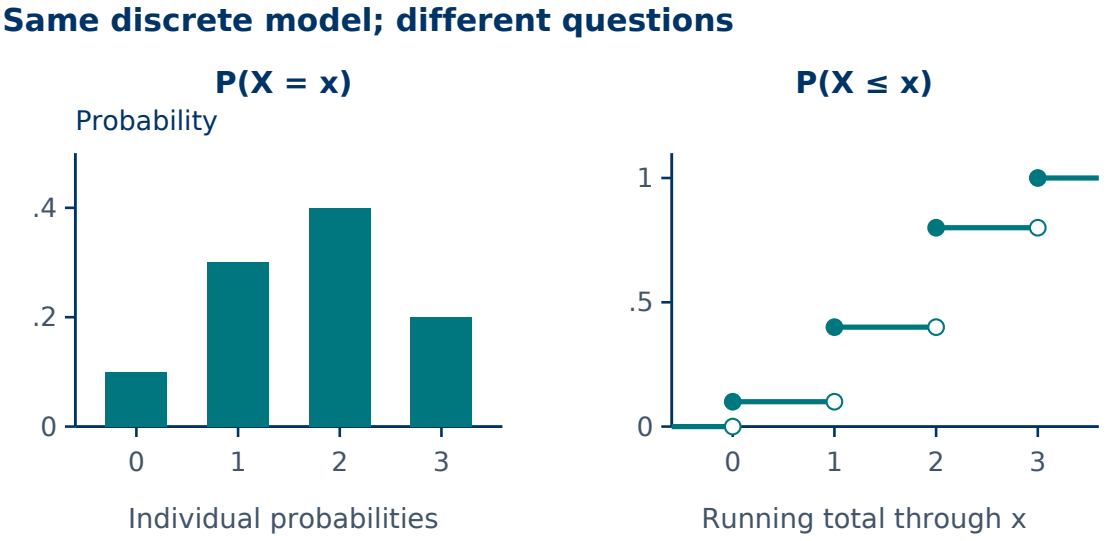
A distribution assigns probability to every value

A discrete probability distribution lists each possible value x and its probability P(X=x). Each probability is between 0 and 1, and all probabilities sum to 1. A table, bar graph, or formula can represent the same distribution. Bar height gives the probability at its labeled value; the x-axis is numerical and its spacing should reflect that. Describe shape, center, and spread in context. A valid distribution can come from equally likely outcomes, a stated model, or simulation; if estimated from simulation, its values carry sampling variation.

Cumulative probabilities answer boundary questions

The cumulative distribution gives P(X at most x), or $P(X \leq x)$. Add probabilities for all values meeting the boundary condition. At most 2 includes 2, whereas fewer than 2 does not. For discrete variables this distinction changes the answer. To find P(X at least 2), subtract P(X at most 1) from 1. A cumulative table must never decrease and must reach 1 at the largest possible value for a finite distribution. It does not usually list the individual value probabilities directly; adjacent differences recover those probabilities.

Read a mass function and a cumulative distribution



Illustrative delayed-delivery model. Open circles exclude the endpoint on the lower step; filled circles include it on the upper step. Cumulative probabilities are .10, .40, .80, and 1.00.

Expected value is a probability-weighted average

The expected value is the long-term average or mean of a random variable. To find it, multiply each possible value by its probability and add the products: $\mu_X = \sum xP(X=x)$. This is the probability-weighted mean described in OpenStax section 4.2. For a team that plays on 0, 1, or 2 days per week with probabilities 0.20, 0.50, and 0.30, the products are 0, 0.50, and 0.60, so the mean is 1.10 days per week. Averaging only 0, 1, and 2 would incorrectly give each possible outcome equal weight. Expected value describes the average over many repetitions under the same model. It need not be a possible individual result and need not be the most likely result.

Measure variability around that expected value

To calculate the standard deviation of a probability distribution, find each deviation from the expected value, square it, multiply by its probability, add the products, and take the square root. In symbols, $\sigma_X^2 = \sum(x-\mu_X)^2P(X=x)$ and $\sigma_X = \sqrt{\sigma_X^2}$. This is the method used in OpenStax Table 4.12. Squaring prevents positive and negative deviations from canceling. The probability weights account for how often each squared deviation occurs in the long run. Use the complete probability distribution, rather than a sample-data formula with divisor n-1. Variance has squared units; standard deviation returns to the original units. Distributions with the same mean can have different standard deviations.

KEY IDEA

Expected value averages possible outcomes using their probabilities as weights.

WATCH FOR

A long-run mean need not be a possible single outcome.

OPENSTAX WORKED EXAMPLE

How many days does the soccer team play?

A soccer team plays on zero, one, or two days per week. The probabilities are 0.20, 0.50, and 0.30, respectively. Let X be the number of days the team plays during a randomly selected week under this model. Calculate and interpret the mean and standard deviation of X.

1. The possible values and their probabilities form a valid distribution: 0.20 + 0.50 + 0.30 = 1.

2. Calculate the probability-weighted mean: $\mu_X = 0(0.20) + 1(0.50) + 2(0.30) = 1.10$ days per week.

3. Use that mean in each squared deviation: $\sigma_X^2 = (0 - 1.10)^2(0.20) + (1 - 1.10)^2(0.50) + (2 - 1.10)^2(0.30)$.

4. The three contributions are 0.242, 0.005, and 0.243. They sum to 0.490, so $\sigma_X = \sqrt{0.490} = 0.70$ day.

5. A week has an integer number of playing days. The mean of 1.10 describes the average over many weeks, not the number of days the team must play in one week.

Conclusion. Under this model, the soccer team averages 1.10 playing days per week over many weeks. The weekly number of playing days has a standard deviation of 0.70 day.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.2, Example 4.3 and Table 4.12, printed pp. 231, 235 (PDF pp. 243, 247).

LOCAL APPLICATION

How many delayed deliveries?

An illustrative model for the number X of delayed deliveries on a route gives values 0, 1, 2, 3 with probabilities .10, .30, .40, .20. Find P(X at least 2), the mean, and the standard deviation.

Probability weights determine the mean

Each outcome contributes x × P(X = x).

x	P(X = x)	x × P(X = x)	Cumulative
0	.10	0.00	.10
1	.30	0.30	.40
2	.40	0.80	.80
3	.20	0.60	1.00

Mean = 0 + .30 + .80 + .60 = 1.70 deliveries

The weighted contributions sum to the long-run mean. The final column also makes the boundary P(X≥2)=1−P(X≤1) visible.

1. The probabilities are nonnegative and sum to 1, so this is a valid distribution. The event X at least 2 contains values 2 and 3.

2. P(X at least 2)=.40+.20=.60. Equivalently, 1-P(X at most 1)=1-(.10+.30)=.60.

3. Compute the weighted mean: 0(.10)+1(.30)+2(.40)+3(.20)=1.70 delayed deliveries per route.

4. The variance is $\sigma_X^2 = (0-1.7)^2(0.10) + (1-1.7)^2(0.30) + (2-1.7)^2(0.40) + (3-1.7)^2(0.20) = 0.81$ squared deliveries. The four weighted squared deviations are 0.289, 0.147, 0.036, and 0.338.

5. The standard deviation is $\sigma_X = \sqrt{0.81} = 0.90$ delivery. The mode is 2, which is distinct from the expected value 1.7.

6. Check the calculation by imagining 100 routes: 10 have no delays, 30 have one, 40 have two, and 20 have three. Those 100 routes have 170 delayed deliveries in total, confirming the mean 1.7.

Conclusion. Across many routes under this model, the average delayed count would approach 1.7, with a standard deviation of .9 delivery; 60% of routes would have at least two delays.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.2, Example 4.3 and Table 4.12, printed pp. 231, 235 (PDF pp. 243, 247). The soccer distribution, weighted-mean method, and variance calculation are retained. Interpretations are clarified for AP Statistics.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.2, Try It 4.3 and Try It 4.7, printed pp. 231, 236 (PDF pp. 243, 248). Practice uses the source water-bottle and card-selection distributions. Cumulative and error-analysis questions extend those source models.

S04.02 | The binomial model

CED 2.10

- Justify a binomial model using its conditions.
- Calculate exact and cumulative binomial probabilities.
- Interpret binomial mean and standard deviation.

CONCEPT AND INTERPRETATION

A count with four defining features

A binomial experiment has a fixed number of trials, two possible outcome categories on each trial, independent trials, and the same success probability on every trial. These are the characteristics described in OpenStax section 4.3, with independence and constant probability stated separately so both must be checked. Define the random variable X as the number of successes in n trials; let p be the probability of success on one trial. Write $X \sim \text{Binomial}(n, p)$. The possible values are 0, 1, ..., n. “Success” means the outcome being counted, even when it is an undesirable outcome such as a damaged component. Counting successes among eight attempts can be binomial. Counting how many attempts are needed to obtain three successes does not have a fixed n and is not a binomial count.

Check assumptions in the actual context

Independence means knowing one trial's outcome does not change another's probability. A shared machine problem, changing weather, or learning across attempts can violate that assumption. Sampling without replacement creates dependence, although when the sample is at most 10% of a large population, a binomial approximation is often reasonable. The 10% check does not create random selection or guarantee a constant underlying success rate. Explain each assumption in context; the label binomial by itself is not a justification.

A binomial model counts successes

One trial set: 12 independent components

Success = unusable; the model fixes p = .25 on every trial.



Illustrative outcome: X = 4 unusable components

Fixed n
12 components before counting

Two outcomes
Usable or unusable

Independent
Knowing one outcome does not change another’s probability.

Constant p
Same .25 chance on every trial

The displayed sequence is one illustrative outcome, not a promise of four unusable components per set. All four conditions belong to the model.

Calculate the chance of exactly x successes

For x=0, 1, ..., n, $P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$. The factor $p^x(1-p)^{n-x}$ gives the probability of one specified ordering with x successes. The binomial coefficient counts the distinct orderings with that success count. Technology may evaluate the formula, but the written work should identify the variable, model parameters, and target event. A probability of at least x requires summing several possible values, not evaluating the single-value formula once.

Translate boundaries before entering technology

A cumulative binomial function typically returns P(X at most k). For P(X at least k), calculate 1-P(X at most k-1). For P(X greater than k), calculate 1-P(X at most k). For a through b inclusive, subtract P(X at most a-1) from P(X at most b). These one-unit adjustments matter because X is discrete. Sketch or list the included integer values before calculating. Check whether the answer is plausible using the expected count and a rough sense of spread. When technology reports a cumulative probability, record the last included integer. This makes off-by-one errors visible and lets another reader reproduce the calculation.

Describe the binomial distribution

Its mean is $\mu_X=np$ and its standard deviation is $\sigma_X=\sqrt{np(1-p)}$. The mean is the long-run average success count in repeated sets of n trials, while the standard deviation describes typical count variation around that mean. Shape depends on both n and p: small p often gives a right-skewed count distribution, p near one gives left skew, and p=.5 gives symmetry. An optional normal approximation requires at least 10 expected successes and 10 expected failures: $np\geq 10$ and $n(1-p)\geq 10$. Exact binomial calculations do not require a normal approximation or these large-count thresholds. Simulation provides a separate way to estimate and check probabilities.

KEY IDEA

A binomial model counts successes in a fixed number of independent, constant-probability trials.

WATCH FOR

At least k is 1 minus the cumulative probability through k-1, not through k.

OPENSTAX WORKED EXAMPLE

An altered coin and an upper-tail event

An altered coin lands heads with probability 0.25 on each of five independent flips. Let X be the number of heads. Calculate the probability of more than three heads. Also calculate the mean and standard deviation of X.

1. There are five fixed trials, two outcomes per trial, independence, and constant heads probability 0.25. Thus $X\sim\text{Binomial}(5,0.25)$. The altered coin is not fair.

2. More than three heads means X = 4 or X = 5. These outcomes are mutually exclusive, so add their probabilities.

3. $P(X>3)=\binom{5}{4}(0.25)^4(0.75)+(0.25)^5=0.0146484375+0.0009765625=0.015625$.

4. A cumulative function gives the same result: P(X > 3) = 1 - P(X ≤ 3). Subtracting through 4 would exclude four-head outcomes and would answer a different question.

04

5. The mean is $5(0.25)=1.25$ heads per five-flip set. The SD is $\sqrt{5(0.25)(0.75)}\approx 0.9682$ head.

Conclusion. Under this model, a five-flip set has a 1.5625% chance of containing more than three heads. Across many such sets, the mean number of heads is 1.25.

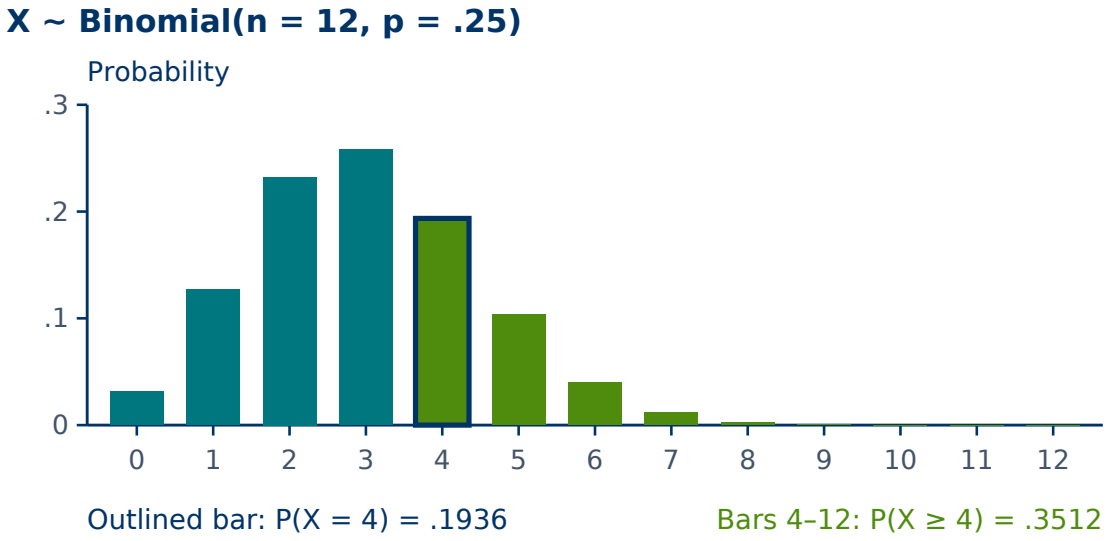
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.3, Example 4.11, printed pp. 238 (PDF pp. 250).

LOCAL APPLICATION

Twelve independently inspected components

An illustrative process makes a usable component with probability .75. Let X count unusable components among 12 independent components. Find P(X=4), P(X at least 4), and the distribution mean and standard deviation.

Exact count and upper-tail probability



Bar heights are exact binomial probabilities. The outlined bar is included in the upper tail beginning at 4.

1. Define success as unusable. Then n=12 and p=.25. Each component is classified usable/unusable, and the model stipulates independent trials with a constant unusable probability.

2. For exactly 4, $P(X=4)=\binom{12}{4}(0.25)^4(0.75)^8\approx 0.1936$. The combinatorial factor is 495.

3. At least 4 includes 4 through 12. Use 1-P(X at most 3)≈.3512, not 1-P(X at most 4), which would omit exactly 4.

4. The mean is 12(.25)=3 unusable components. The standard deviation is the square root of 12(.25)(.75), or 1.5 components.

5. Check the event against the distribution: 4 is only one count above the mean, so an upper-tail probability around .35 is plausible; a probability near .001 would deserve scrutiny.

6. The mean 3 describes repeated sets of 12, not a claim that each set must contain exactly three unusable components.

Conclusion. Under this model, about 19.4% of 12-component sets have exactly 4 unusable components, and about 35.1% have at least 4.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.3, Example 4.11, printed pp. 238 (PDF pp. 250). Retains the altered coin, five independent flips, and event more than three heads. Corrects the source's stray reference to a fair coin. Mean/SD interpretation extends the source problem.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.3, Try It 4.11 and Example 4.16, printed pp. 238, 242 (PDF pp. 250, 254). Used for independent practice. Source language is sharpened to distinguish a fixed set of trials from a simulation repetition.

S04.03 | Continuous models and the normal distribution

CED 2.11

- Interpret areas under a density curve as probabilities.
- Describe a normal distribution using its mean and standard deviation.
- Use the empirical rule and standardized positions.

CONCEPT AND INTERPRETATION

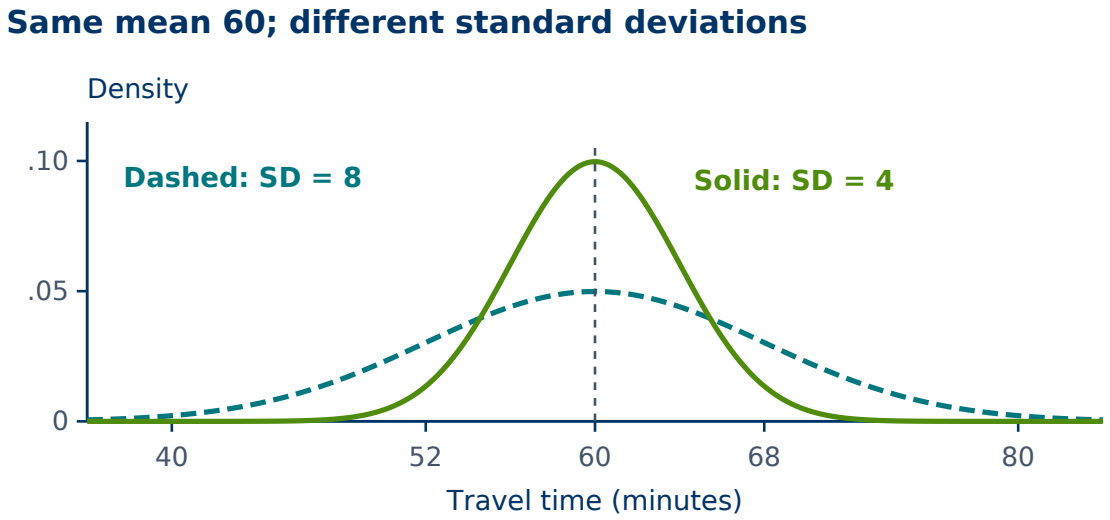
Continuous outcomes are modeled with areas

A continuous random variable can take any value in an interval. Measurements such as travel time are modeled this way even when the recording device rounds to seconds. A probability density curve is nonnegative and has total area 1. Probability for an interval is the area over that interval, not the curve height at one point. In a continuous model, $P(X=x)=0$ for an exact single value because a point has zero width. A recorded value of 60 seconds may represent a small interval of unrounded times, which can have positive probability.

Shape and parameters of a normal model

A normal distribution is continuous, unimodal, symmetric, and bell-shaped. Its mean, median, and peak occur at the same central value. The mean μ locates the center, and the positive standard deviation σ controls spread. Larger σ produces a broader, lower curve because its total area stays 1. State whether notation $N(\mu, \sigma)$ lists standard deviation or variance; in this package the second quantity is standard deviation. A graph that looks roughly bell-shaped may support a normal model, but the model is an assumption to assess, not a default for every variable.

Standard deviation controls the spread



Exact normal densities for N(60,8) and N(60,4), with the second parameter defined as standard deviation. The shared axis uses density, not point probability.

The empirical rule uses standard-deviation distances

OpenStax calls this the empirical rule, or 68-95-99.7 rule. For a normal distribution with mean μ and standard deviation σ, about 68% of values lie from μ - σ to μ + σ, about 95% from μ - 2σ to μ + 2σ, and about 99.7% from μ - 3σ to μ + 3σ. Start by labeling the mean and moving in equal increments of σ along the horizontal axis. Symmetry divides excluded probability equally between the two tails. Outside the central 68%, for example, 32% remains, so each tail has about 16%. Between one and two standard deviations above the mean lies about (95% - 68%)/2 = 13.5%. These are normal-model benchmarks, not universal percentages for arbitrary distributions or guaranteed percentages in every sample.

Standardization keeps relative position

For a value x, define its standardized score as $z = (x - \mu)/\sigma$. A positive z is above the mean, a negative z below, and |z| measures distance in standard-deviation units. Standardizing a normal variable produces a standard normal variable with mean 0 and standard deviation 1. Standardizing any distribution sets its center and spread in these units but does not change its shape into a normal curve. The inverse transformation is x=μ+zσ. Always identify which distribution supplies μ and σ.

Link the model to the original question

A normal probability can describe a proportion of a modeled population or the chance a randomly selected observation falls in an interval. For N(60, 8) travel times in minutes, P(52<X<68) is the modeled proportion between 52 and 68 minutes. It is not the probability that an average of many trips falls there; that requires a sampling distribution. Check whether impossible values, skewness, or extreme outliers make the normal assumption doubtful. Calculated probabilities inherit every important limitation of the model and the data collection.

KEY IDEA

Normal probabilities are areas; the mean locates the curve and standard deviation controls its spread.

WATCH FOR

A standardized distribution is not automatically normal.

OPENSTAX WORKED EXAMPLE

Locate the three empirical-rule intervals

Suppose X has a normal distribution with mean 50 and standard deviation 6. Identify the central intervals containing approximately 68%, 95%, and 99.7% of the distribution. Estimate $P(X > 62)$.

1. One standard deviation on either side of the mean gives 50 - 6 = 44 and 50 + 6 = 56. Approximately 68% of values lie between 44 and 56.

2. Two standard deviations give $50 \pm 2(6)$, or 38 to 62. Approximately 95% of values lie in this interval.

3. Three standard deviations give $50 \pm 3(6)$, or 32 to 68. Approximately 99.7% of values lie in this interval.

4. The boundary 62 is two standard deviations above the mean. The approximately 5% outside the central 95% is divided equally between the two tails by symmetry. Therefore $P(X > 62) \approx 0.025$.

Conclusion. For this normal model, the central empirical-rule intervals are 44-56, 38-62, and 32-68. Approximately 2.5% of values exceed 62. These percentages are approximations, not exact normal-CDF results.

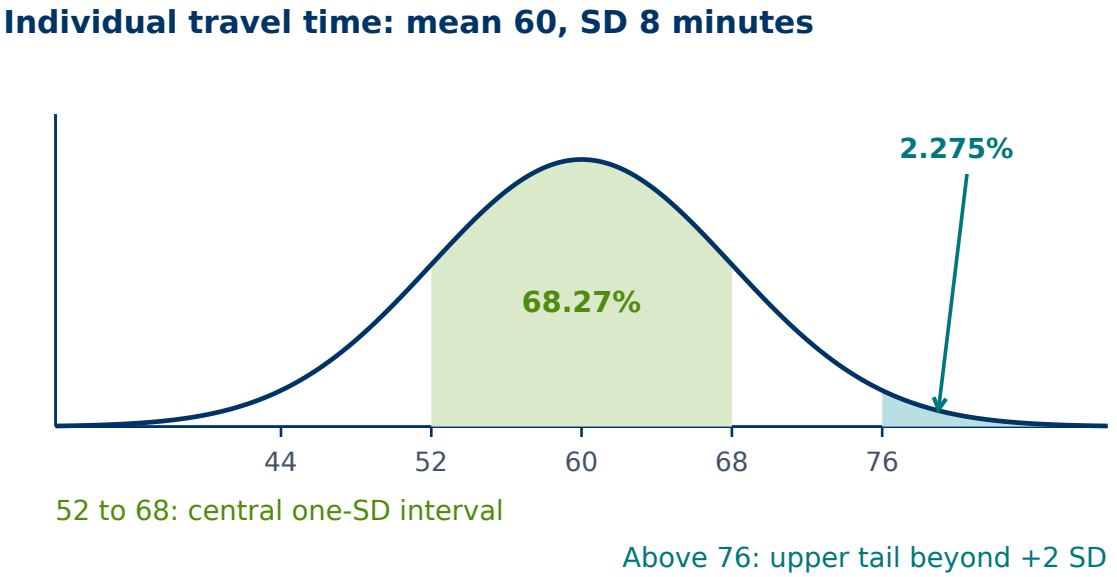
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 6.1, Example 6.5, printed pp. 339 (PDF pp. 351). Upper-tail application added.

LOCAL APPLICATION

A travel-time model

An illustrative distribution of travel times is normal with mean 60 minutes and standard deviation 8 minutes. Estimate the proportion between 52 and 68 minutes and above 76 minutes. Locate a 70-minute trip within the distribution.

Translate boundaries into areas



The shaded areas use precise normal probabilities. The empirical-rule approximations are 68% and 2.5%, respectively.

1. Write X for one trip's travel time and X approximately N(60, 8), with the second parameter explicitly defined as the standard deviation.

2. The bounds 52 and 68 are 60±8, one standard deviation on either side. The empirical rule gives about 68% between them. A precise calculation gives .6827.

3. The bound 76 is 60+2(8). About 5% lie outside 44 to 76, with half in the upper tail, so about 2.5% exceed 76. A precise calculation gives .0228.

4. For 70 minutes, z=(70-60)/8=1.25. This trip is 1.25 standard deviations above the modeled mean. It is slower than average, but the z-score alone is not a percentile.

5. Interpret the probabilities across comparable individual trips. If selecting trips is nonrandom or conditions change, the model may not represent the selected trips.

Conclusion. Using the empirical rule, the normal travel-time model places about 68% of trips between 52 and 68 minutes and about 2.5% above 76 minutes. A precise normal calculation gives about 2.28% for the latter tail. A 70-minute trip is 1.25 standard deviations above the mean travel time.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 6.1, Example 6.5, printed pp. 339 (PDF pp. 351). Retains mean 50 and SD 6. The standard-normal positions and empirical-rule reasoning follow the source.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 6.1, Try It 6.6, printed pp. 340 (PDF pp. 352). Exam-score model used in independent practice. The distinction between density and probability also follows section 5.1.

S04.04 | Normal probabilities, percentiles, and cutoffs

CED 2.11

- Calculate normal probabilities for one-tail and interval events.
- Find normal percentiles and central-interval boundaries.
- Compare relative positions across normal distributions.

CONCEPT AND INTERPRETATION

Translate words into a shaded event

OpenStax section 6.2 represents the area to the left of a boundary x by P(X < x). The area to the right is 1 - P(X < x). Begin each question by defining X, stating its normal mean and standard deviation, and translating the requested event into an inequality. Shade the relevant part of a labeled normal curve before calculating. For an interval, subtract the cumulative area below the lower boundary from the cumulative area below the upper boundary. Endpoints have probability zero in a continuous model, so P(X < x) and P(X ≤ x) agree. Check the result against symmetry: a boundary above the mean must leave less than half the distribution to its right.

Standardize or supply the original parameters

Technology may calculate normal areas with the original mean and standard deviation, or a standard normal table can be used after computing z=(x-μ)/σ. Both routes must produce the same probability when entered correctly. State the inputs and requested event so the calculation can be checked; a calculator command name alone is not enough. Keep extra digits in intermediate z-scores and round the final probability sensibly. A rough empirical-rule comparison often reveals a reversed tail or an incorrect standard deviation.

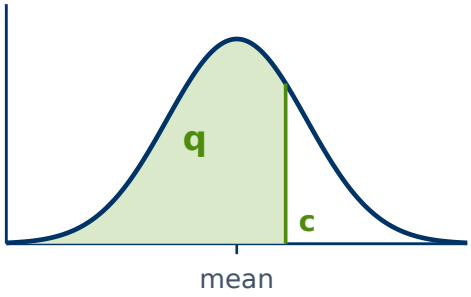
Inverse questions start with area

A percentile cutoff c satisfies P(X at most c)=q, where q is the desired left-tail probability. First obtain the standard-normal percentile z_q, then calculate $c = \mu + z_q\sigma$. A top 10% cutoff has 90% below it, so use q=.90, not .10. A bottom 10% cutoff uses q=.10. Inverse functions require probabilities as proportions between 0 and 1; entering 90 instead of .90 is not meaningful. Interpret the cutoff in original units and preserve the requested direction.

The value and the area exchange roles

Identify what is given and what must be found.

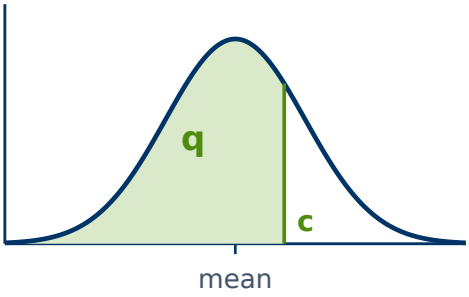
Forward probability



Given boundary c

Find shaded area q

Inverse percentile



Given area q

Find boundary c

Schematic normal diagrams. q denotes the cumulative area to the left of c; numerical parameters are deliberately unspecified.

Central intervals split the excluded area

A central 90% normal interval leaves 10% outside, divided into 5% in each tail by symmetry. Its boundaries therefore use the 5th and 95th percentiles. The central 80% interval uses the 10th and 90th percentiles. In contrast, the most extreme 10% consists of both outside tails, not the middle 10%. Explain which values the interval includes. If a question specifies unequal tail areas, do not impose symmetry; use the actual left cumulative area for each boundary.

Compare relative standing, not raw values alone

To compare observations from different normal distributions, use z-scores or percentiles. A larger raw score can have a lower relative position if its distribution has a much larger mean or spread. For normal distributions, larger z means larger percentile. For differently shaped nonnormal distributions, z-scores still compare standard-deviation distance but do not guarantee the same percentile ordering. Context also matters: a high percentile is desirable for some variables, such as performance scores, and undesirable for others, such as waiting time.

KEY IDEA

Forward problems start with a value and find area; inverse problems start with area and find a value.

WATCH FOR

A top-tail percentage must be converted to the matching left-tail cumulative probability.

OPENSTAX WORKED EXAMPLE

An exam probability and a percentile cutoff

Final exam scores in a statistics class are modeled by a normal distribution with mean 63 points and standard deviation 5 points. Calculate the probability that a randomly selected student scores more than 65. Then find the 90th percentile of the score distribution.

1. Let X be the selected student's score. State the event $P(X > 65)$ and shade the region to the right of 65 on a normal curve centered at 63.

2. Standardize the boundary: $z = (65 - 63)/5 = 0.40$. The cumulative area to its left is about 0.6554, so $P(X > 65) = 1 - 0.6554 \approx 0.3446$.

3. For the percentile, let c satisfy $P(X \leq c) = 0.90$. The input area is given and the boundary is unknown. The standard-normal 90th percentile is approximately 1.28155.

4. Convert to points: $c = 63 + 1.28155(5) \approx 69.41$. Check that the cutoff is above the mean and leaves 10% above it.

Conclusion. Under the normal model, about 34.46% of students score more than 65 points. The 90th-percentile cutoff is approximately 69.4 points, with about 90% of scores at or below that value.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 6.2, Example 6.8, parts (a) and (c), printed pp. 341-343 (PDF pp. 353-355).

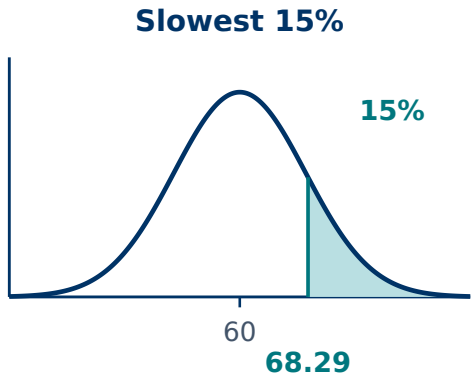
LOCAL APPLICATION

A cutoff for the slowest 15%

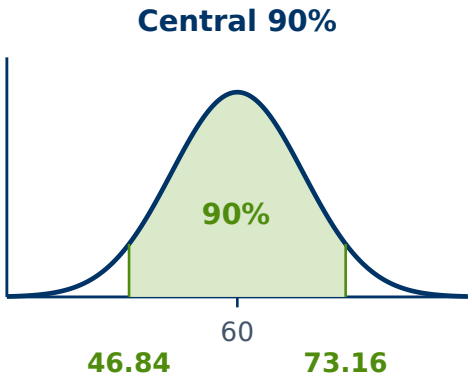
For the illustrative $N(60, 8)$ travel-time model, find the cutoff separating the slowest 15% of trips. Also find the central 90% interval.

Place one-sided and central cutoffs

N(60,8): the second parameter is the SD in minutes.



85% lies below the cutoff.



5% remains in each tail.

Computed cutoffs use the supplied mean 60 and SD 8. The one-sided 85th percentile differs from the two boundaries of a central 90% interval.

1. Let X be travel time in minutes. The slowest 15% are the largest times, so seek c satisfying $P(X>c)=.15$. This means $P(X\leq c)=.85$.

2. The 85th standard-normal percentile is about 1.03643. Convert back: $c=60+1.03643(8)\approx 68.29$ minutes.

3. The answer is above the 60-minute mean, consistent with keeping only 15% in the upper tail. A cutoff below the mean would signal a reversed percentile.

4. For a central 90% interval, split the remaining 10% into 5% in each tail. The standard-normal cutoffs are approximately -1.64485 and +1.64485.

5. Convert both endpoints: $60\pm 1.64485(8)$, giving 46.84 to 73.16 minutes. Each outside tail has probability .05.

6. The central interval is centered at 60 because normal symmetry gives equal distances to the two tail cutoffs. Its total width is about 26.32 minutes; the upper 15% cutoff is a separate one-sided boundary.

Conclusion. The model puts the slowest 15% above about 68.3 minutes and the central 90% between about 46.8 and 73.2 minutes.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 6.2, Example 6.8, parts (a) and (c), printed pp. 341-343 (PDF pp. 353-355). Exam scores, parameters, probability boundary, and percentile are retained. Solutions show both area and inverse calculations.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 6.2, Try It 6.8 and Try It 6.9, printed pp. 343-344 (PDF pp. 355-356). The school golf-score model is used for independent and homework practice. Additional percentile tasks use that same source model.

Unit synthesis

1. A discrete distribution assigns probabilities .1, .6, .3 to values 0, 1, 2. Find its mean and $P(X \geq 1)$.

2. For that distribution, find variance and SD.

3. Explain why counting attempts until two successes is not a binomial count.

4. For X binomial with $n=10$, $p=.2$, find $P(X \geq 1)$, mean, and SD.

5. A population of test scores is modeled by a normal distribution with mean 80 points and standard deviation 12 points. Let X be the score of a randomly selected student. Calculate $P(X > 104)$ and interpret it in context.

6. For the score model in Question 5, calculate the central 80% interval. Identify the percentiles used and give endpoints in points.

Unit 05 | Sampling Distributions

Inference depends on understanding how statistics change across repeated samples. This unit separates individual data from sample summaries, distinguishes bias from variability, and makes repeated sampling visible through small exact examples and simulation. It also introduces randomization distributions, which hold observed responses fixed while reallocating treatment labels under a no-effect model. Students then develop sampling models for proportions, means, and independent differences. Every calculation identifies the statistic and checks design, independence, and shape separately. These models lead directly to confidence intervals and hypothesis tests, where unknown population quantities must be estimated or specified by a null hypothesis.

CED topics: 2.12, 3.1, 3.2, 3.9, 4.1, 4.6

IN THIS UNIT

Lesson	Focus
S05.01	Sampling distributions and randomization distributions
S05.02	Sampling distributions of proportions and their differences
S05.03	Sample means, differences in means, and the CLT

Three illustrative samples on the same trail

Group	Users interviewed	Local residents
A	20	8
B	20	11
C	20	9

Each fraction uses one group’s own 20 interviews. The three sample proportions vary.

LOCAL QUESTION

If different groups sample trail users, how will their estimates vary? Distinguish the trail-user data from the distribution of sample statistics.

S05.01 | Sampling distributions and randomization distributions

CED 2.12, 3.1

- Distinguish a population, one sample, and a sampling distribution.
- Explain estimator bias and the effect of sample size on variability.
- Construct a randomization distribution by reallocating fixed responses.

CONCEPT AND INTERPRETATION

Three distributions answer three different questions

A population distribution describes individual values for every observational unit in the population. A sample distribution describes individual values in one selected sample. A sampling distribution describes a statistic, such as a sample mean, across all possible random samples of the same size from the same population using the same sampling process. Each point in a simulated sampling distribution represents a whole sample summarized by one number. Increasing the number of simulated samples improves the approximation; increasing the size within each sample changes the sampling distribution itself.

Parameters are fixed; statistics vary

A parameter describes the population, such as μ for a population mean or p for a population proportion. A statistic is calculated from sample data, such as a sample mean or sample proportion. Before selection, the statistic varies because different random samples contain different units. A point estimator is the rule used to estimate a parameter; its realized value is a point estimate. A single estimate can miss the target even when the estimator is unbiased. Unbiased means the sampling distribution is centered at the true parameter, not that every estimate equals it.

Bias and variability are different problems

Bias is systematic overestimation or underestimation across repetitions. Variability is the spread of the estimator across repetitions. A narrow distribution centered away from the target is precise but biased. A wide distribution centered at the target is unbiased but imprecise. Larger random samples generally reduce the variability of a sample mean or proportion; they do not repair undercoverage, misleading questions, or a biased estimator. Evaluate both the center and spread rather than judging a method from one unusually good or bad estimate.

Approximate a sampling distribution by repeated samples

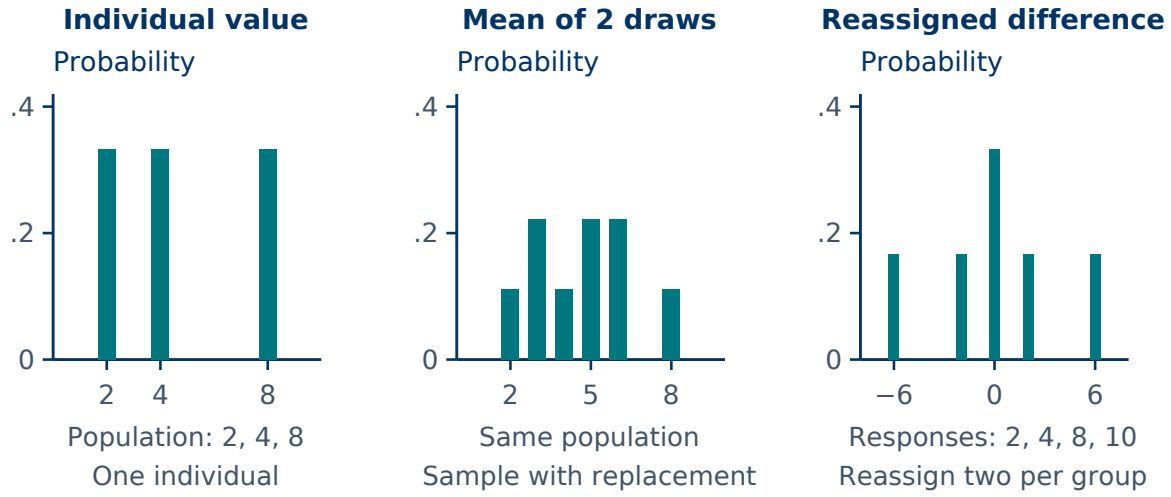
The OpenStax dice experiment makes the repetition explicit: roll a fixed number of dice, calculate their mean, and record that one mean. Repeat the whole process many times. The number of dice in each set is the sample size; the number of sets is the number of repetitions. For any sampling simulation, state the population or model, sample size, sampling method, and statistic. Generate one sample, calculate one statistic, and record it. Repeat without changing the model or sample size, then describe the shape, center, and variability of the recorded statistics. With replacement, each draw uses the same population. Without replacement within a sample, restore the full population before beginning the next independent sample.

A randomization distribution reassigns fixed outcomes

In a randomized experiment, a null model of no treatment effect allows the observed response values to be reallocated to treatment groups. Keep those responses and group sizes fixed; repeatedly shuffle group assignments, calculate a treatment-comparison statistic, and plot the results. This distribution represents variation due to assignment under the null model. It differs from drawing new observations from a population. Compare the observed statistic with equally or more extreme reallocations in the question's direction. Random assignment supports a causal comparison, while population generalization still depends on who entered the study.

Three distributions, three meanings

Identify what one plotted value represents.



The middle panel summarizes all nine ordered samples of size 2. The right panel summarizes all six equal-size reallocations of four fixed responses; it uses a different illustrative experiment.

KEY IDEA

Each point in a sampling distribution summarizes a whole sample; each point in a randomization distribution summarizes one reassignment.

WATCH FOR

A large number of simulated samples is different from a large sample size.

OPENSTAX WORKED EXAMPLE

Five die results produce one sample mean

In the OpenStax dice experiment, one set of five fair die rolls gives 2, 2, 3, 4, and 6. Each set is a sample of size five. Explain how the observations from this set differ from the values plotted when many groups repeat the experiment and record their sample means.

1. The five observations are the individual die results. Their sample mean is $(2 + 2 + 3 + 4 + 6)/5 = 17/5 = 3.4$.

2. A graph of this one sample would contain five observations. A graph of means from repeated five-roll sets contains one number for each complete set. The value 3.4 contributes one plotted mean.

3. Repeat all five independent rolls for each new set and calculate a new mean. Keep the sample size at five. Repeating the experiment more times adds more sample means; it does not increase the size of any one sample.

4. For a fair die, the population mean is $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$. A particular sample mean such as 3.4 can differ from 3.5. Across all possible five-roll samples, the sample means are centered at 3.5.

Conclusion. One dot at 3.4 in the sampling-distribution display summarizes five die rolls. It does not represent one roll or the mean of all the groups combined.

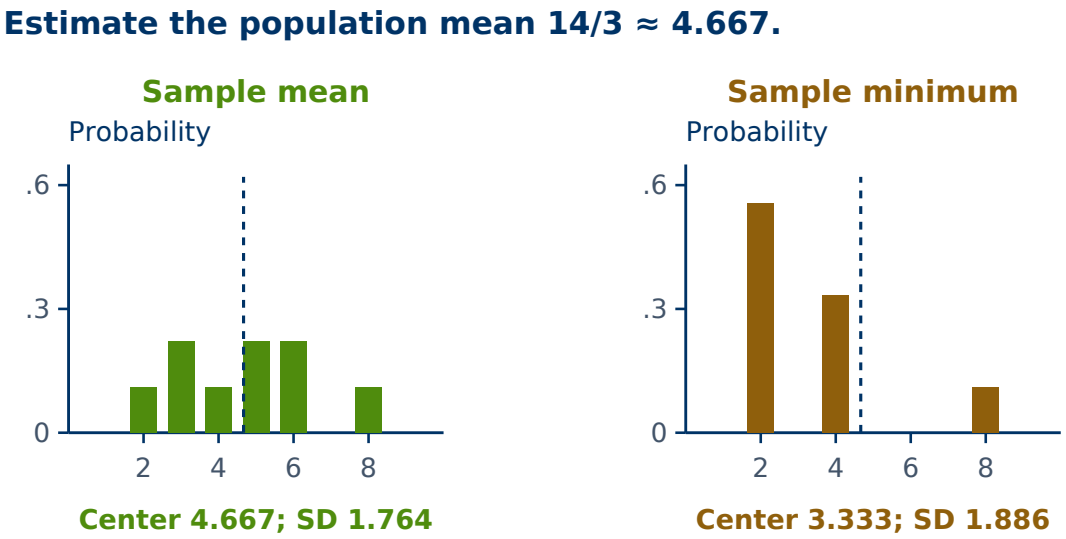
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 7 Introduction, Collaborative Exercise, printed pp. 365 (PDF pp. 377).

LOCAL APPLICATION

Enumerate a small sampling distribution

An illustrative population contains three values: 2, 4, and 8. Select one uniformly at random, replace it, and repeat for a second draw. Use the mean of the two draws to estimate the population mean. Enumerate the ordered samples.

Compare bias and variability using the same samples



Exact distributions from the worked example. The mean estimator is centered at the target; the minimum estimator is biased downward and has a slightly larger sampling SD.

1. The population mean is $(2+4+8)/3=14/3\approx4.6667$. There are nine equally likely ordered samples because each draw has three possibilities.

2. The ordered samples (2, 2), (2, 4), (2, 8), (4, 2), (4, 4), (4, 8), (8, 2), (8, 4), (8, 8) give means 2, 3, 5, 3, 4, 6, 5, 6, 8.

3. The mean of these nine equally likely sample means is 42/9=14/3, equal to the population mean. Therefore the sample-mean estimator is unbiased for this sampling process.

4. The population SD is approximately 2.4944. The SD of the nine possible sample means is approximately 1.7638, smaller than the population SD.

5. If the estimator were the smaller of the two sampled values, its nine values would be 2, 2, 2, 2, 4, 4, 2, 4, 8. Their mean is 30/9≈3.3333, below the target 4.6667.

Conclusion. The sample mean is unbiased in this example, but individual sample means vary. The sample minimum systematically underestimates the population mean.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 7 Introduction, Collaborative Exercise, printed pp. 365 (PDF pp. 377). Uses the source's five-die example 2, 2, 3, 4, 6 to distinguish observations from a sample mean. Repeated-sample interpretation is developed before normal approximation.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 7.1, opening explanation, printed pp. 366-367 (PDF pp. 378-379). Supplies the distinction between sample size and number of repetitions. Exact small-population and randomization examples are original AP bridges, because these required conceptual tasks are not supplied as worked examples in the cited OpenStax sections.

S05.02 | Sampling distributions of proportions and their differences

CED 3.2, 3.9

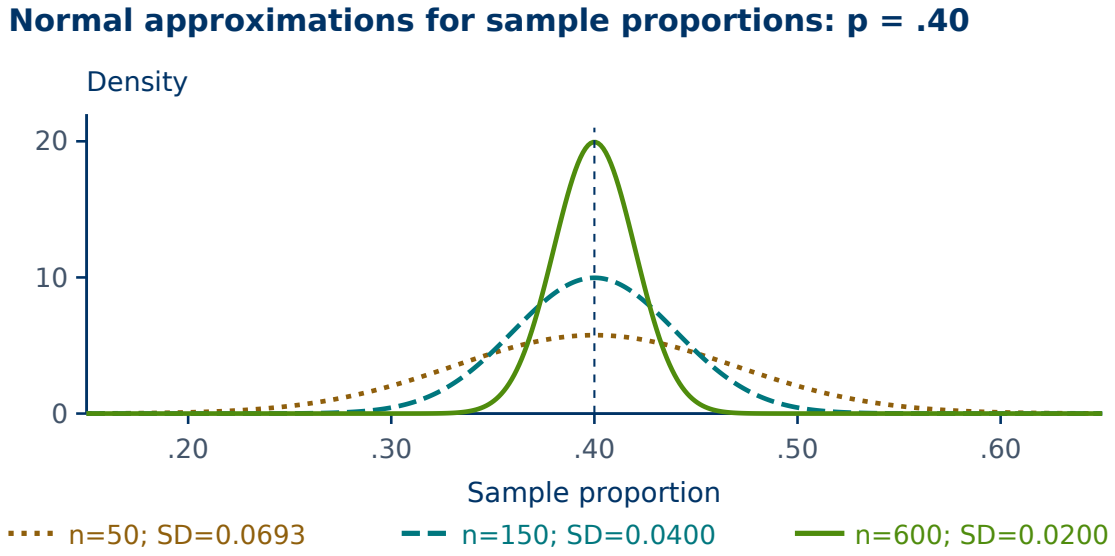
- Calculate the mean and SD of one sample proportion or a difference of independent sample proportions.
- Check randomization, 10%, and large-count conditions.
- Calculate and interpret normal-model probabilities for sample proportions.

CONCEPT AND INTERPRETATION

A sample proportion varies around the population proportion

For a binary outcome, let p be the population success proportion, n the sample size, and X the observed success count. The sample proportion is $\hat{p} = X/n$. Under independent random sampling its sampling distribution has mean $\mu_{\hat{p}} = p$ and SD $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$. The mean says the estimator is unbiased. The SD describes typical sample-to-sample fluctuation in proportion units, not individual binary outcomes. A decimal SD of .04 corresponds to four percentage points. A success count and a success proportion have different units and different standard deviations. Dividing a count by n divides both its mean and its standard deviation by n, which explains the proportion formula.

Larger samples narrow the sampling distribution



All three approximations are centered at .40. Increasing n from 150 to 600 divides the sampling SD by 2; it does not change the population proportion.

Separate three conditions and their purposes

Random sampling supports a probability-based connection to the population. When sampling without replacement, require n≤.10N, where N is population size, to make independence a reasonable approximation for the stated SD formula. An approximately normal shape also requires expected successes np≥10 and expected failures n(1-p)≥10. These are different checks: a sample can meet the 10% condition but fail large counts, or meet large counts while being a biased convenience sample. For a probability calculation with a stipulated p, use that p in the expected counts and SD.

Compare two independent samples

Fix the group order before calculating. Let $\hat{p}_1$ and $\hat{p}_2$ be the sample proportions from independent random samples, and let p1 and p2 be the corresponding population proportions. The statistic $D = \hat{p}_1 - \hat{p}_2$ estimates p1 - p2, so its mean is $\mu_D = p_1 - p_2$. A positive D means the sample proportion in Group 1 is larger. For independent groups, $\sigma_D = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$. Each term inside the square root is a variance, not a standard deviation. Use the sizes of the two samples separately. Repeated responses from the same people are paired and do not satisfy between-group independence.

Why subtraction still adds independent variances

Subtracting Group 2 changes the direction of its deviations, but it does not make those deviations smaller. If Group 1 happens to be high and Group 2 low, their difference is especially high. The opposite combination makes it especially low. Neither random fluctuation is removed by subtraction. More formally, for independent quantities A and B, $\mathrm{Var}(A-B)=\mathrm{Var}(A)+\mathrm{Var}(B)$. Independence justifies this variance rule; without independence, adding variances is not guaranteed. Means subtract whether or not the quantities are independent. Therefore, square the two sampling SDs, add the variances, and take a square root. For example, independent estimates with SDs 0.03 and 0.04 have difference SD $\sqrt{0.03^2+0.04^2}=0.05$, not 0.01 or 0.07. Reversing group order negates the mean difference and leaves its SD unchanged.

Check each group before using a normal difference model

For two random samples without replacement, check n1≤.10N1 and n2≤.10N2 separately. For normal shape, check all four expected counts: n1p1, n1(1-p1), n2p2, n2(1-p2), each at least 10. Randomized experiments use random assignment as their randomization condition; a finite-population 10% sampling check is not automatically required merely because there are treatment groups. The large-count requirement for a normal approximation still matters. State which design produced the data rather than reciting an irrelevant checklist.

Standardize the statistic, not a single observation

To approximate a probability, subtract the sampling-distribution mean from the specified sample statistic boundary and divide by its sampling SD. Use a normal area only after justifying shape. If the statistic is a difference, keep the same group order in the parameter difference, mean, boundary, and interpretation. These calculations assume population proportions are known model values. Confidence intervals and tests later handle unknown proportions with estimated or null-model standard errors; do not silently substitute a sample proportion into this known-parameter probability problem.

KEY IDEA

A proportion's sampling SD measures variation among sample proportions, not among individual binary outcomes.

WATCH FOR

Use expected counts from the stated model and check both groups separately.

OPENSTAX WORKED EXAMPLE

Turn a binomial count into a sample proportion

In the OpenStax model, each of 50 statistics students independently completes homework on time with probability 0.70. Let X be the number who do so and let p-hat = X/50. Find the mean and standard deviation of p-hat. Express “at least 40 students” as an event involving p-hat.

1. X counts successes in 50 independent, constant-probability trials, so X ~ Binomial(50, 0.70). Its mean is 50(0.70) = 35 students and its SD is √[50(0.70)(0.30)] ≈ 3.2404 students.

2. Dividing X by 50 changes counts into proportions. Divide both its mean and SD by 50: mean 35/50 = 0.70; SD 3.2404/50 ≈ 0.06481.

3. Equivalently, $\sigma_{\hat{p}} = \sqrt{0.70(0.30)/50} \approx 0.06481$. This is about 6.48 percentage points. It describes variation among the proportions from repeated 50-student sets.

4. The event X ≥ 40 is the same as p-hat ≥ 40/50 = 0.80. The exact binomial probability is P(X ≥ 40) = 1 - P(X ≤ 39) ≈ 0.07885.

5. A normal approximation would require checking both expected counts. They are 35 on-time students and 15 not-on-time students, each at least 10. The exact binomial calculation remains available and does not require that approximation.

Conclusion. The sample proportion is centered at 0.70 with sampling SD about 0.06481. Under the source model, about 7.89% of repeated 50-student sets contain at least 80% who complete homework on time. These are assumed model probabilities, not measured SMS completion rates.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.3, Example 4.12, printed pp. 238-239 (PDF pp. 250-251). Count-to-proportion bridge added.

WORKED AP CONNECTION

A fully worked difference of independent proportions

Two large school populations have true walking-to-school proportions p1 = 0.55 and p2 = 0.40. Independent simple random samples of 100 students are selected, one from each population of 5,000 students. Let D be the first sample proportion minus the second. Describe its approximate sampling distribution and approximate P(D > 0.25).

1. The mean difference is p1 - p2 = 0.55 - 0.40 = 0.15. Thus repeated sample differences are centered at a 15-percentage-point advantage for Group 1.

2. Random selection and independence between groups are stated. Each sample meets the 10% condition: 100 ≤ 0.10(5,000) = 500.

3. For normal shape, the expected success and failure counts are 55 and 45 in Group 1, and 40 and 60 in Group 2. All four are at least 10.

4. Add the two sampling variances: $\sigma_D = \sqrt{0.55(0.45)/100 + 0.40(0.60)/100} = \sqrt{0.004875} \approx 0.06982$. The SD is about 6.98 percentage points.

5. The event D > 0.25 asks for a sample gap exceeding 25 percentage points. Standardize using the sampling distribution of D: z = (0.25 - 0.15)/0.06982 ≈ 1.4322. The upper-tail normal area is approximately 0.0760.

Conclusion. Under the stated model, roughly 7.60% of independent pairs of samples of 100 students would show a walking-proportion difference greater than 25 percentage points in favor of Group 1.

Source: Original AP bridge for CED 3.9, developing the independent-proportion comparison in OpenStax section 10.3. The supplied textbook does not give this known-parameter worked probability example.

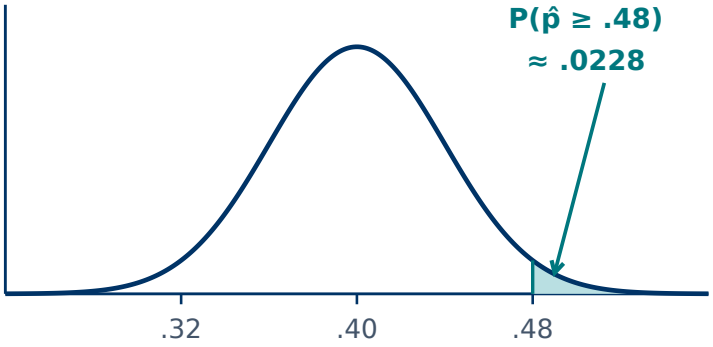
LOCAL APPLICATION

How much can a random sample fluctuate?

An illustrative population of 5,000 students has 40% who walk to school. Select an SRS of 150 without replacement. Approximate the probability that at least 48% of the sample walks.

Use the sampling SD for a sample proportion

SRS of 150 from 5,000 students; population p = .40



Model checks

150 ≤ 500

np = 60

n(1−p) = 90

Mean = .40

SD = .04

Boundary z = (.48 − .40) / .04 = 2

One plotted value represents a complete sample of 150.

The shaded area is a normal approximation to the event in the worked example. Its variability is .04 proportion units, or four percentage points.

1. Let p=.40 and let p-hat be the proportion of walkers in the 150-person sample. Its mean is .40. The requested event is p-hat≥.48, not an event about an individual student.

2. Random selection is stated. The 10% check is 150≤.10(5,000)=500. Expected walkers are 150(.40)=60 and nonwalkers 150(.60)=90, both at least 10.

3. Ignoring the small dependence from sampling without replacement, the sampling SD is approximately √[0.40(0.60)/150] = 0.04. Sample proportions from repeated samples of 150 typically vary by about four percentage points around 0.40.

4. Standardize the boundary: z=(.48-.40)/.04=2. Use the upper-tail normal area, approximately .0228.

5. The actual count of walkers is discrete. The normal area approximates its upper tail; it is not the exact probability for a sample drawn without replacement. An individual result near 0.40 cannot establish how the sample was selected.

Conclusion. Under the stated sampling model, roughly 2.3% of samples of 150 would contain at least 48% walkers.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 4.3, Example 4.12, printed pp. 238-239 (PDF pp. 250-251). Retains 50 students and homework-on-time probability 0.70. The count-to-proportion calculation is an AP bridge added to the original binomial example.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, introductory comparison of independent proportions, printed pp. 523 (PDF pp. 535). Develops the comparison idea, but its pooled null standard error is reserved for hypothesis testing. The known-parameter sampling formulas, large-count threshold of 10, and worked two-population probability below follow CED 3.2 and 3.9.

S05.03 | Sample means, differences in means, and the CLT

CED 2.12, 4.1, 4.6

- Describe the sampling distribution of a sample mean and its dependence on sample size.
- Use the CLT without claiming the original data become normal.
- Model a difference of independent sample means with appropriate conditions.

CONCEPT AND INTERPRETATION

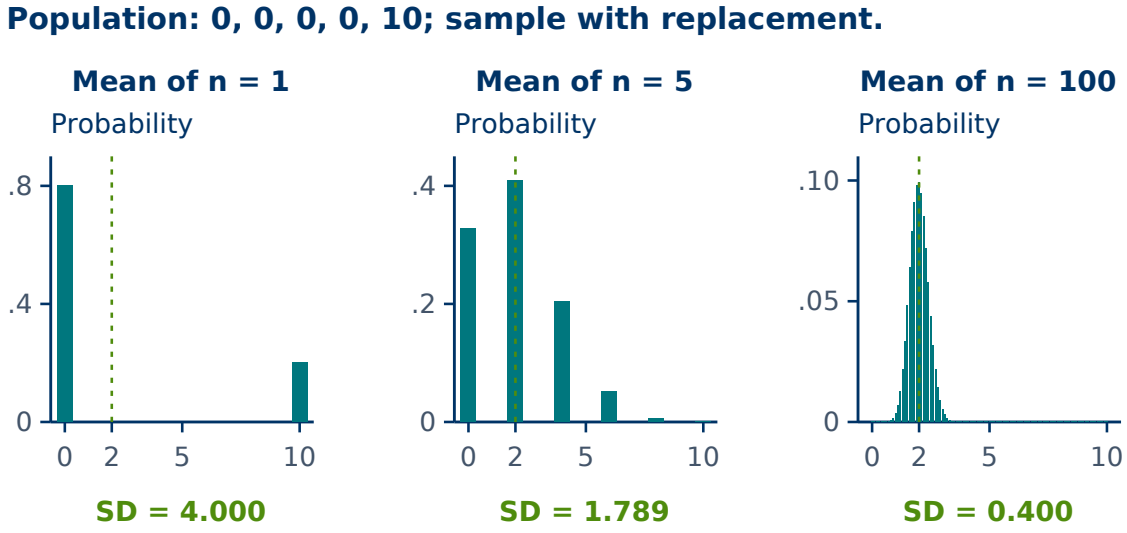
Average variation is smaller than individual variation

For a quantitative population with mean μ and standard deviation σ, the mean of n independent observations has mean μ and variance σ^2/n. As OpenStax section 7.1 explains, standard deviation is the square root of variance, so the sampling SD is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. The sample mean is an unbiased estimator of μ. The symbol n is the number of observations averaged in each sample, not the number of repeated samples in a simulation. A population SD of 12 minutes describes variation among individual times. With n = 36, the sampling SD is $12/\sqrt{36} = 2$ minutes and describes variation among 36-observation averages. Both quantities use minutes. Larger samples reduce the variability of the average without changing the spread of individual values.

Shape depends on the population and sample size

If the population distribution is normal, the sampling distribution of the mean is normal for every sample size under independent sampling. For a nonnormal population with finite mean and variance, the central limit theorem says the sampling distribution of the mean becomes approximately normal as n grows. The course uses $n \geq 30$ as a common guideline, but extremely skewed or heavy-tailed populations may require much larger samples. The CLT changes the shape of the distribution of sample means, not the shape of individual observations in the population or each sample.

The distribution of averages changes with sample size



Each distribution is centered at 2. Vertical probability scales differ.

These are exact sampling distributions: the number of 10s is Binomial(n,.20), and the sample mean is 10X/n. The original population remains skewed while averages become more concentrated and approximately normal.

Randomization and finite-population checks remain necessary

Use a random sample to connect probability calculations to a population. For sampling without replacement, check n≤.10N before using the independence-based SD as an approximation. A large n can help with normal shape, but it does not fix selection bias or dependence. Keep these conditions separate. When a population SD is supplied as a model value, use it for sampling-distribution probability calculations. Later inference procedures use a sample SD to estimate an unknown σ and account for that extra uncertainty with a t distribution.

Independent sample means can be compared

For independent samples from populations with means μ1, μ2 and SDs σ1, σ2, the statistic sample mean 1 minus sample mean 2 has mean μ1-μ2 and SD $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Check random selection and the 10% condition for each population when sampling without replacement. A normal model is justified when both populations are normal, or approximately when both samples are sufficiently large for their respective shapes. Checking only the combined sample size is insufficient. Randomized treatment comparisons rely on the assignment mechanism rather than automatically requiring a population-sampling 10% check.

Build the difference from the two sampling variances

First calculate the variance of each sample mean: $\mathrm{Var}(\bar{X}_1) = \sigma_1^2/n_1$ and $\mathrm{Var}(\bar{X}_2) = \sigma_2^2/n_2$. If the samples are independent, their fluctuations both contribute to the variability of $\bar{X}_1 - \bar{X}_2$. Subtract the means, add the variances, and then take a square root. For example, if each sample mean has SD 2 minutes, their difference has SD $\sqrt{2^2 + 2^2} = \sqrt{8}$ minutes. It does not have SD zero. Equal spread does not mean the two independently selected averages will rise and fall together. In paired data, however, they may move together, so the independent-sample rule does not apply. Form each within-pair difference and analyze those differences instead.

Pairing changes the unit being averaged

If the same individual is measured twice, or individuals are deliberately matched, analyze within-pair differences. A sample of 20 pairs has 20 differences, not 40 independent responses. Define the subtraction direction, then treat the differences as one quantitative variable with its own mean, SD, and shape. Do not use the independent-two-sample formula just because there are two columns of data. For every probability calculation, name the random statistic, check conditions, standardize using its own mean and SD, and interpret the area across repeated samples or assignments.

KEY IDEA

The CLT describes averages across repeated samples; it does not make individual observations normal.

WATCH FOR

For two groups, check each sample and identify pairing before choosing a formula.

OPENSTAX WORKED EXAMPLE

Average duration of fifty soccer matches

The time to play a soccer match in a particular league is modeled by a normal distribution with mean 2 hours and standard deviation 0.5 hour. Consider 50 independently selected matches from this model. Calculate the probability that their mean duration is between 1.8 and 2.3 hours.

1. Let X be one match duration in hours and let x-bar be the mean duration of the 50 matches. The question concerns the sampling distribution of x-bar.

2. Because the population is normal and observations are independent, the sample mean is normal. Its mean is 2 hours and SD is $0.5/\sqrt{50} \approx 0.07071$ hour.

3. Standardize both endpoints using this sampling SD: $(1.8 - 2)/0.07071 \approx -2.8284$ and $(2.3 - 2)/0.07071 \approx 4.2426$.

4. The area between those standard-normal boundaries is approximately 0.99765. A normal-CDF calculation with mean 2 and SD 0.07071 gives the same result.

5. Using SD 0.5 in this calculation would answer a question about one match rather than the mean of 50 matches.

Conclusion. Under the stated model, about 99.765% of independent sets of 50 matches would have a mean duration between 1.8 and 2.3 hours.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 7.1, Example 7.2, printed pp. 368-369 (PDF pp. 380-381).

LOCAL APPLICATION

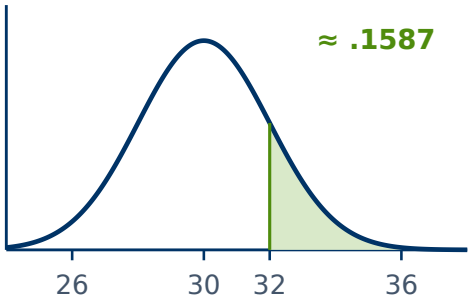
Individual times and average times

A hypothetical normal population of service-visit waiting times has mean 30 minutes and standard deviation 12 minutes. Select a simple random sample of 36 from 10,000 recorded visits. Approximate the probability that its mean exceeds 32 minutes. Then compare two independent normal route populations with mean travel times of 50 and 45 minutes and standard deviations of 10 and 12 minutes. Independently select simple random samples of 25 and 36 trips from populations of at least 1,000 trips each. Approximate the probability that the first sample mean exceeds the second by more than 8 minutes.

Two statistics require two variability calculations

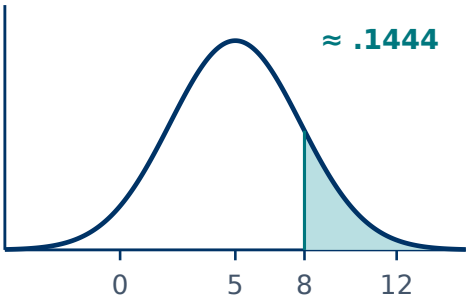
Match the distribution to the statistic.

Mean of 36 waits



Center 30; SD $12/\sqrt{36} = 2$
Event: sample mean > 32

Difference of means



Center $50-45 = 5$; SD $= \sqrt{8}$
Event: mean 1 − mean 2 > 8

Both panels use the normal populations and independent samples in the worked example. The difference SD is $\sqrt{(10^2/25 + 12^2/36)} = \sqrt{8}$, not the difference of the two SDs.

1. The random sample of 36 visits is below 10% of the 10,000 recorded visits. Using the stated normal model and treating the draws as approximately independent, the sampling distribution of the mean is approximately normal with mean 30 minutes and SD about $12/\sqrt{36} = 2$ minutes.

2. The boundary 32 has $z=(32-30)/2=1$. Therefore $P(\text{sample mean}>32)\approx.1587$. Using SD 12 would answer a question about one individual visit.

3. For the difference problem, both populations are normal and samples independent. Each sample is at most 10% of its population, so the standard model is justified.

4. For the route difference D = sample mean 1 minus sample mean 2, the mean is 50 - 45 = 5 minutes. Each sample mean has SD 2 minutes: $10/\sqrt{25} = 2$ and $12/\sqrt{36} = 2$. Add their variances: $SD(D) = \sqrt{(10^2/25 + 12^2/36)} = \sqrt{(4+4)} \approx 2.8284$ minutes.

5. For a difference above 8, $z=(8-5)/2.8284 \approx 1.0607$. The upper-tail probability is approximately .1444.

Conclusion. About 15.87% of samples of 36 service visits have mean waiting time above 32 minutes under the first model. About 14.44% of independent route-sample pairs have a first-minus-second mean travel-time difference above 8 minutes under the second model. Each probability uses the sampling distribution of the statistic named in its question.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 7.1, Example 7.2, printed pp. 368-369 (PDF pp. 380-381). Retains the normal soccer-match model, mean 2 hours, SD 0.5 hour, n = 50, and probability bounds 1.8 and 2.3 hours. Independence is explicit. Normality of sample means follows from normal populations without needing the CLT.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 7.1, opening explanation and Try It 7.1, printed pp. 366-368 (PDF pp. 378-380). Adapted explanation retains the source distinction between sample size and repetitions. The Try It model is used only for mean and SD unless a shape assumption is supplied.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.1, opening discussion of independent sample means, printed pp. 511 (PDF pp. 523). Supports the difference-of-means explanation. Known-population-SD calculations are developed as an AP sampling-distribution bridge; no pooled variance assumption is used.

Unit synthesis

1. Explain the difference between sampling more individuals per sample and generating more simulated samples.

2. In repeated independent samples of 150 students from a population with walking proportion p = 0.40, let p-hat be the sample walking proportion. Calculate its sampling mean and standard deviation, then verify the large-count condition.

3. Independent random samples of sizes 100 and 120 are drawn from two large populations with proportions p1 = 0.60 and p2 = 0.45. Treat draws within each sample as independent. Calculate the mean and standard deviation of p-hat 1 - p-hat 2.

4. A normal population of waiting times has mean 20 minutes and standard deviation 6 minutes. For nine independent randomly selected waits, calculate the probability that their sample mean exceeds 22 minutes. Explain why the normal sampling model is appropriate.

5. Explain why two measurements per person require a paired analysis rather than an independent-two-sample SD formula.

6. Describe a randomization distribution for a randomized experiment comparing two mean responses.

Unit 06 | Confidence Intervals for Proportions

A sample proportion gives a useful estimate, but another random sample would produce another value. This unit turns that sampling variability into an interval estimate. Students first investigate what confidence means across repeated samples, then connect standard error, critical value, and margin of error. They learn to justify conditions, choose a sample size, compare two population proportions, and communicate what an interval supports. Every conclusion must name the population and response being estimated. The calculations address random sampling variation; they cannot repair a biased sampling frame, leading question, or unrepresentative set of volunteers.

CED topics: 3.3, 3.4, 3.10, 3.11

IN THIS UNIT	
Lesson	Focus
S06.01	What a confidence interval says
S06.02	Build and justify a one-proportion interval
S06.03	Plan precision before collecting data
S06.04	Estimate a difference in population proportions

One roster, one annual survey date

Sampling frame	Day's measurement	Summary
List of river sites	Above cutoff: yes or no	Fraction of sampled sites above the cutoff

Define the cutoff, observation time, and eligible sites before sampling.

LOCAL QUESTION

For a specified sampling date and temperature threshold, what proportion of sites on a monitoring roster exceed the threshold? Define the measurement and how you would select the sites.

S06.01 | What a confidence interval says

CED 3.3, 3.4

- Distinguish a point estimate from an interval estimate.
- Interpret a confidence level using repeated random sampling.
- Interpret an interval and assess a contextual claim.

CONCEPT AND INTERPRETATION

Estimate a population quantity

A population proportion, p, is the fraction of a specified population with a specified response or characteristic. A sample proportion, $\hat{p} = x/n$, is calculated from x successes among n sampled individuals. The word success simply names the response being counted; it need not describe a desirable event. Define that response before collecting data. The statistic changes from sample to sample, while the population parameter stays fixed during the investigation. A point estimate reports one value. An interval estimate reports a range of plausible values and makes sampling uncertainty visible.

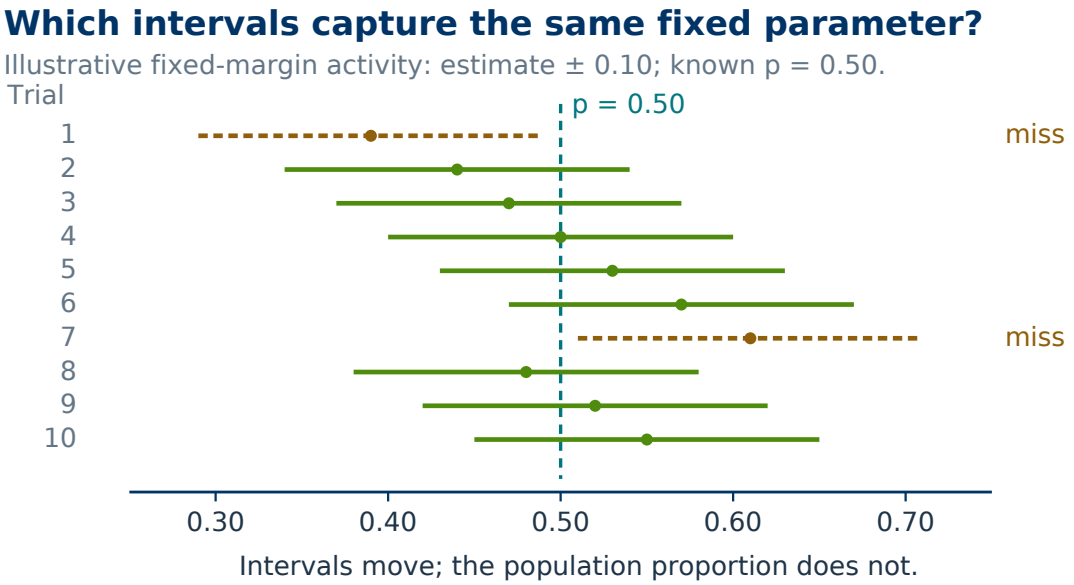
From a statistic to an interval

Many confidence intervals have the form point estimate plus or minus margin of error. If a poll reports 0.62 plus or minus 0.09, the interval is (0.53, 0.71). Its midpoint is 0.62, its margin of error is 0.09, and its width is 0.18. Proportions and percentages describe the same quantity on different scales: a margin of 0.09 is nine percentage points. It is not a nine-percent relative change. The margin reflects an estimation method and a chosen confidence level; it is not the observed distance between the statistic and the unknown parameter.

Confidence describes a repeated method

A 95% confidence method is designed so that, over many repetitions of the same random-sampling process and interval calculation, approximately 95% of the resulting intervals contain the true parameter. The sample and endpoints vary across repetitions; p does not. In a classroom simulation with known p, we can check which intervals capture it. In an actual study, p is unknown, so we cannot identify whether our particular interval is one of the successful intervals. A small simulation need not produce exactly 95% coverage, and a normal-approximation method has approximate rather than exact coverage.

Confidence across repetitions



Eight of these ten illustrative fixed-margin intervals capture 0.50. The small-batch capture rate is not required to equal a nominal confidence level; this simplified activity is not the formal estimated-SE interval.

Interpret the particular interval

An appropriate report is: We are 95% confident that between 53% and 71% of all adults in the defined town support the proposed service. This statement names the confidence level, population, characteristic, and endpoints. It does not say that 95% of adults have proportions in that interval. Individuals have responses, while the population has a proportion. Nor does it say that 95% of the sampled adults support the proposal. The frequentist confidence level describes the method; after the interval is calculated, the fixed p either lies in it or does not.

Claims and limits

An interval entirely above 0.50 supports a claim that a population majority has the characteristic, at the stated confidence level. An interval containing 0.50 leaves a majority and a nonmajority plausible. It does not prove equality or prove that a majority is absent. Before interpreting either result, examine how data were collected. A narrow interval from a voluntary online poll can still miss p because of selection bias. The usual margin of error addresses random sampling variability, not every possible source of error. Confidence should therefore be accompanied by a clear description of the study design.

Read the uncertainty in a poll

OpenStax introduces proportion intervals with a poll reporting 40% support within three percentage points. The point estimate is $\hat{p} = 0.40$ and the margin of error is 0.03, so the endpoints are $0.40 - 0.03 = 0.37$ and $0.40 + 0.03 = 0.43$. This construction answers how precisely the population support proportion has been estimated. It does not divide individual voters into a 95% group and a 5% group. To explain a confidence level, imagine repeating the entire random-sampling and interval procedure. To interpret the interval already calculated, name the population, the response being counted, the endpoints, and the confidence level. These are two related but different explanations.

KEY IDEA

The confidence level describes the long-run capture rate of a method; the computed interval estimates a specific population parameter.

WATCH FOR

Do not move the 95% from the interval-generating method to individual people, sample responses, or a probability assigned to a fixed parameter.

OPENSTAX WORKED EXAMPLE

A poll reports a margin of error

A textbook election-poll example reports an estimated support proportion of 0.40 with a margin of error of 0.03 at the 95% confidence level. Assume that the sampling design and interval conditions are satisfied. Interpret the interval and explain whether it provides convincing evidence that a majority of the population supports the candidate.

1. Let p be the proportion of voters in the population represented by the poll who support the candidate. The sample estimate is 0.40.

2. Calculate the endpoints: 0.40 ± 0.03 gives (0.37, 0.43). The full width is 0.06, or six percentage points.

3. We are 95% confident that between 37% and 43% of the voters in this population support the candidate.

4. Across many random samples collected and analyzed by this same method, approximately 95% of the resulting intervals would contain the population support proportion.

5. Every value in the interval is below 0.50. The interval provides convincing evidence that fewer than half of the voters support the candidate.

Conclusion. A 40% sample estimate and a three-percentage-point margin describe both the estimate and its uncertainty. Neither the 95% confidence level nor the interval describes the percentage of individual responses captured by the interval.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, opening election-poll example, printed p. 420 (PDF p. 432). Numerical example retained. Population and design assumptions, majority question, and AP interpretation are made explicit.

LOCAL APPLICATION

Would a majority support later library hours?

Illustrative study: a simple random sample of 100 adults from a town of 5,000 adults contains 62 supporters of later library hours. A previously checked 95% interval is (0.525, 0.715). Interpret it and evaluate the majority claim.

Read the worked interval accurately

Feature	Library-hours example	Meaning
Population parameter	p = town-wide support proportion	The fixed, unknown quantity being estimated
Point estimate	62 / 100 = 0.620	Observed sample support
95% interval	(0.525, 0.715)	Plausible population support values
Margin and width	0.095; 0.190	9.5-point half-width; 19-point full width
Majority threshold	0.50 lies below the interval	The interval supports majority approval

The endpoints here are the rounded values supplied in the lesson. Confidence describes the repeated interval method, not the percentage of people who support the proposal.

1. Define p as the proportion of all 5,000 adults in this town who support later library hours. The population and the yes/no response identify the quantity of interest.

2. The point estimate is $62/100 = 0.62$. The interval midpoint is $(0.525 + 0.715)/2 = 0.620$; the margin of error is $(0.715 - 0.525)/2 = 0.095$.

3. Write a contextual interpretation: We are 95% confident that the proportion of all adults in the town who support later library hours is between 0.525 and 0.715.

4. Compare the entire interval with 0.50. Even its lower endpoint exceeds 0.50, so the interval provides evidence of majority support. The sample proportion alone would not communicate this uncertainty.

5. Explain confidence separately: Across many SRSs of 100 adults and corresponding intervals constructed this way, about 95% would contain the fixed population proportion. This does not establish whether this particular interval contains p.

Conclusion. The interval supports majority approval in this town, subject to the sampling and measurement assumptions. It does not estimate the views of all Oregon adults.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, opening election-poll example, printed p. 420 (PDF p. 432). Numerical example retained. Population and design assumptions, majority question, and AP interpretation are made explicit.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Example 8.11, printed p. 423 (PDF p. 435). Source interval and context retained. Sampling assumptions are stated explicitly, and the task separates interval from level.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S06.02 | Build and justify a one-proportion interval

CED 3.3, 3.4

- Identify a one-sample z-interval for a population proportion.
- Verify randomization, the 10% condition when applicable, and observed success/failure counts.
- Calculate standard error, margin of error, and a contextual interval.

CONCEPT AND INTERPRETATION

Choose a method from the question

Use a one-sample z-interval for a population proportion when the target is the fraction of a population with a specified categorical outcome. Define p in words and calculate $\hat{p} = x/n$, where x counts successes and n counts sampled individuals. A count, a proportion, and a mean are different quantities. For example, the fraction of households with a rain barrel is a proportion; their average water use is a mean. The desired confidence level must be supplied or chosen before reporting the interval. A calculator does not decide whether the method matches the question.

Justify the sampling basis

The usual interval requires a random sample from the population being described. State the sampling procedure given in the study rather than merely writing random. When sampling without replacement, check $n \le 0.10N$, where N is the population size. This makes the dependence caused by sampling without replacement small enough for the usual standard-error formula. It does not make a convenience sample random. If N is not supplied, state the population-size requirement and any justified assumption explicitly. Random sampling supports generalization; no confidence interval can extend the sampling frame automatically.

Check the normal approximation

For the interval method, the observed numbers of successes and failures must each be at least 10: $n\hat{p} = x \ge 10$ and $n(1-\hat{p}) = n - x \ge 10$. Write both counts and connect them to the condition. This checks the adequacy of the normal approximation used by this particular method. It is not the same as the 10% condition, which concerns approximate independence. A sample of 300 with only three successes fails the observed-count requirement despite being large overall. Do not replace a failed condition with an unsupported assertion that the calculator can handle it.

Three conditions protect different parts of the argument

Condition	Evidence in the worked study	What it supports
Randomization	Stated SRS of district households	A sampling basis for generalization
10% condition	200 ≤ 0.10(8,000) = 800	Approximate independence without replacement
Observed counts	138 successes; 62 failures; both ≥ 10	Normal approximation for this interval
Correct standard error	sqrt [0.69(0.31)/200] = 0.03270	Estimated variation of the sample proportion

A large sample size does not substitute for any of these separate justifications.

Estimate sampling variability

The standard error is $SE = \sqrt{\hat{p}(1-\hat{p})/n}$. It estimates the standard deviation of the sampling distribution of the sample proportion because the true p is unknown. A C% confidence interval is $\hat{p} \pm z^* SE$, where z^* encloses the middle C% of a standard normal distribution. Common critical values are approximately 1.645 for 90%, 1.960 for 95%, and 2.576 for 99%. The margin of error is $z^* SE$. Keep extra digits through the calculation and round endpoints at the end. Standard error and margin of error have the same units as the proportion.

Report evidence in a connected argument

An inference response should identify the parameter and method, justify conditions with study details, show meaningful numerical work, and interpret the result in context. Calculator output can support the calculation, but report the entered success count, sample size, confidence level, and resulting interval so another reader can reproduce it. Separate a descriptive sample fact from the inferred population conclusion. If a condition fails, explain which one and why the standard interval is not justified. That diagnosis is a statistical result, not an invitation to force a numerical answer. Use the complete inference structure when a question asks you to construct and interpret an interval. If a question asks only for a parameter, a condition, or an interpretation, provide that requested component clearly. Repeating all five template headings is unnecessary for a one-part task.

Why this standard error uses the sample proportion

OpenStax develops the proportion interval by dividing a binomial success count X by the sample size n. If X counts independent successes with probability p, then the sample proportion has mean p and standard deviation $\sqrt{p(1-p)/n}$. In an estimation problem p is the unknown quantity, so substitute $\hat{p}$ to obtain the estimated standard error $\sqrt{\hat{p}(1-\hat{p})/n}$. The multiplier z^* then extends the interval by the required number of estimated standard errors on each side. For 95% confidence, 2.5% remains in each standard-normal tail, so $z^* = 1.96$. The AP large-count check uses at least 10 observed successes and 10 observed failures. This is stricter than the five-count rule in the source text.

KEY IDEA

An interval is defensible only when the method, study design, conditions, calculation, and contextual conclusion agree.

WATCH FOR

Use observed successes and failures for a proportion interval; a later hypothesis test will use counts calculated under its null hypothesis.

OPENSTAX WORKED EXAMPLE

Estimate smartphone ownership

A market research firm surveys 500 randomly selected adult residents of a large city. Of those surveyed, 421 report owning a smartphone. Assume the city has at least 5,000 adult residents and all selected adults respond. Construct and interpret a 95% confidence interval for the proportion of the city's adult residents who own a smartphone.

1. Let p be the proportion of all adult residents of this city who own a smartphone. Use a one-sample z interval for a population proportion.

2. The residents were randomly selected. The sample size 500 is at most 10% of the population of at least 5,000. The observed counts are 421 owners and 79 nonowners, both at least 10.

3. The point estimate is $\hat{p} = 421/500 = 0.842$. The estimated standard error is $\sqrt{0.842(0.158)/500} = 0.016312$ and the 95% margin of error is $1.959964(0.016312) = 0.031970$.

4. The confidence interval is 0.842 ± 0.031970, or approximately (0.8100, 0.8740).

5. We are 95% confident that between 81.0% and 87.4% of all adult residents of this city own a smartphone.

Conclusion. This is the textbook's numerical scenario, not a statement about smartphone ownership in Medford today. The conclusion concerns the city population represented by the random sample.

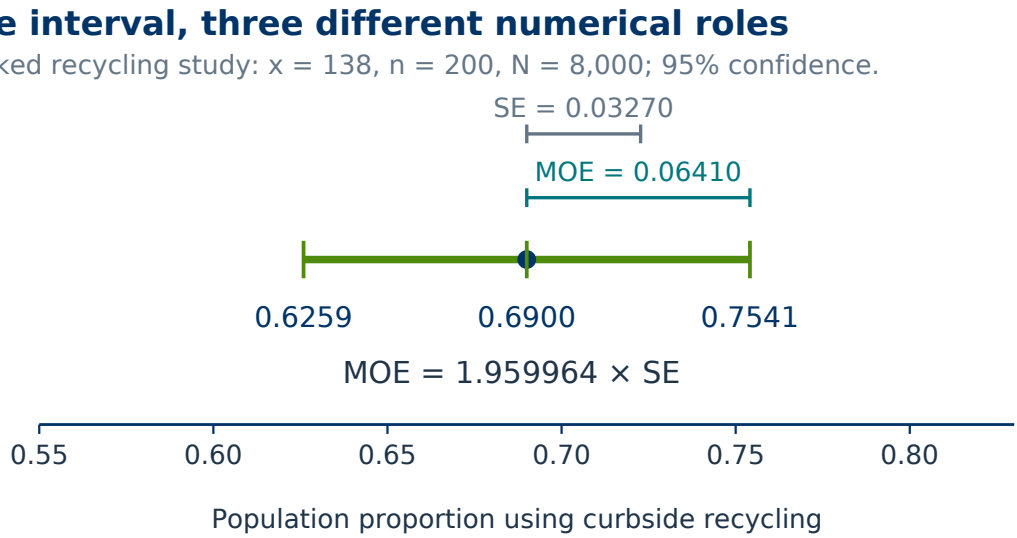
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Example 8.10, printed pp. 421-422 (PDF pp. 433-434). Counts and context retained. Population-size and response assumptions and AP condition checks are explicit.

LOCAL APPLICATION

Household recycling participation

Illustrative study: an SRS without replacement of 200 of a district's 8,000 households finds 138 using curbside recycling. Construct and interpret a 95% interval for the district proportion.

Standard error is not the margin of error



The point estimate is 0.6900. Multiplying its estimated standard error by the 95% critical value gives the half-width of the interval. Both endpoints were calculated from the worked data.

1. Classify. Each household is recorded as using or not using curbside recycling. The district sampled households without assigning a treatment.

2. Choose. Let p be the proportion of all 8,000 district households using curbside recycling. Use a one-sample z interval for p at 95% confidence.

3. Check. The sample was selected by SRS. The sample fraction is 200/8,000 = 2.5%. The observed counts are 138 using recycling and 62 not using it. Both exceed 10.

4. Calculate. The estimate is 138/200 = 0.690. Its estimated standard error is sqrt(0.690 × 0.310 / 200) = 0.03270. The interval is 0.690 ± 1.960(0.03270), or (0.6259, 0.7541).

5. Conclude. We are 95% confident that between 62.6% and 75.4% of all district households use curbside recycling.

Conclusion. All endpoints exceed 0.50, providing evidence that a majority participate. The conclusion assumes the recorded responses measure the intended behavior accurately.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Example 8.10, printed pp. 421-422 (PDF pp. 433-434). Counts and context retained. Population-size and response assumptions and AP condition checks are explicit.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Try It 8.10, printed pp. 422-423 (PDF pp. 434-435). Counts and ownership context retained. A population size and complete response are supplied so all AP conditions can be evaluated.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S06.03 | Plan precision before collecting data

CED 3.3, 3.4

- Explain effects of confidence level and sample size on interval width.
- Solve for a planning sample size and round upward.
- Distinguish precision, confidence, bias, and practical cost.

CONCEPT AND INTERPRETATION

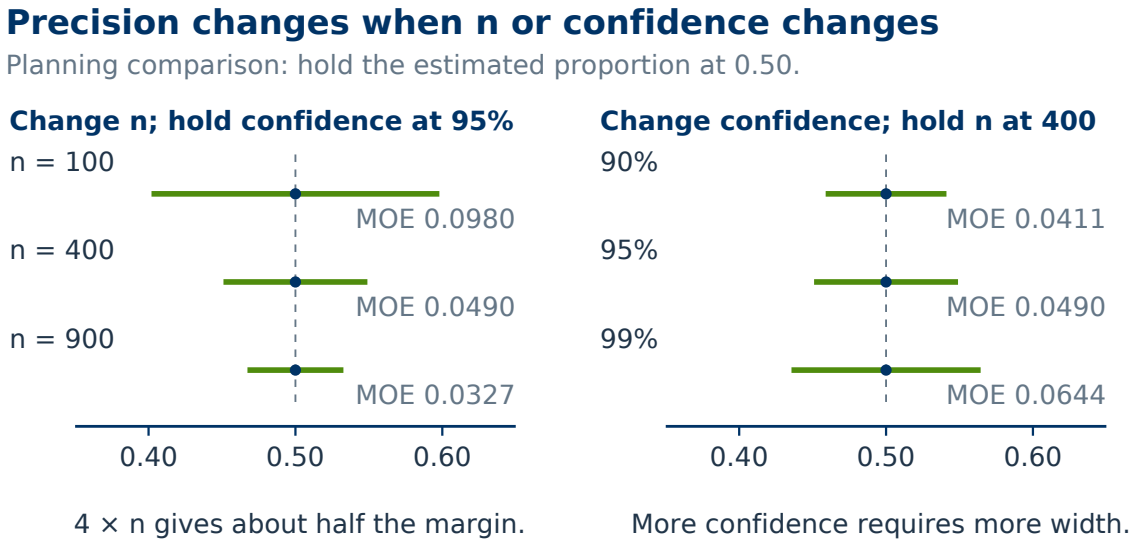
Precision is a width question

The margin of error describes the half-width of an interval, not whether its midpoint is free of bias. A more precise interval has a smaller margin, provided the population quantity and method are comparable. A higher-confidence interval from the same data is wider because its critical value is larger. Thus confidence and precision are different goals. State both when planning a study: for example, estimate the population proportion within four percentage points using a 95% confidence method. A target margin must use proportion units, so four percentage points means E = 0.04, not E = 4.

How sample size changes uncertainty

For an approximately fixed proportion and confidence level, $MOE = z^* \sqrt{\hat{p}(1-\hat{p})/n}$. The margin therefore scales approximately as $1/\sqrt{n}$. Doubling n divides the margin by the square root of two, not by two. To halve the margin, multiply n by four. Different random samples can produce different sample proportions, so actual intervals will not have exactly the predicted width. The relationship describes planning with other components held fixed. Once the 10% condition is safely satisfied, population size does not appear in the usual formula; the absolute sample size drives precision.

Separate the two precision tradeoffs



All intervals are centered at 0.50 for comparison. The sample-size panel changes n only; the confidence panel changes the critical value only.

Solve the planning equation

Let E be the desired margin of error, z^* the critical value for the selected confidence level, and p_{plan} a planning estimate of the population proportion. Solve $E = z^* \sqrt{p_{plan}(1-p_{plan})/n}$ to obtain $n \geq (z^*)^2 p_{plan}(1-p_{plan})/E^2$. A pilot study or relevant earlier study may provide the planning estimate, but it remains uncertain. Round the computed n upward to the next whole number: rounding downward fails the stated inequality. This gives the needed number of completed, usable observations under the model, not necessarily the number of people who must be contacted.

Why use one-half when p is unknown?

The product p (1-p) is largest at p = 0.50, where it equals 0.25. Choosing $p_{plan} = 0.50$ therefore gives the conservative normal-approximation planning size when no reliable estimate is available. It does not assert that exactly half of the population has the characteristic. A planning value near zero or one produces a smaller required n, but may underestimate the required size if the actual proportion is closer to one-half. After collection, check the actual success/failure counts and other conditions. A planning formula cannot promise that every realized sample will satisfy them.

A defensible plan includes more than n

Increasing the sample size reduces random sampling variability but does not fix undercoverage, nonresponse bias, or misleading wording. Plan a sampling frame, random selection process, consistent response definition, and follow-up procedures as well as a sample size. If anticipated nonresponse is 20%, contacting n/0.80 people estimates the required contact count, but this arithmetic does not guarantee an unbiased responding sample. If the proposed sample exceeds 10% of a small population, reconsider the usual approximation or use a method that accounts for finite-population sampling. A wider honest interval is preferable to an unjustified narrow one.

Derive the required sample size

OpenStax obtains the sample-size formula by solving the proportion margin-of-error equation. Starting with $E = z^* \sqrt{p_{plan}(1-p_{plan})/n}$, square both sides and multiply by n. This gives $nE^2 = (z^*)^2 p_{plan}(1-p_{plan})$. Dividing by E^2 gives the minimum unrounded sample size. When no planning estimate is available, set $p_{plan} = 0.50$ because $p(1-p) = 0.25 - (p-0.50)^2$ is never larger than 0.25. Finally round upward. This protects the target margin under the stated approximation; it does not guarantee that the sample will be representative.

KEY IDEA

Choose the confidence level and target margin first; determine a sufficient sample size and a defensible way to collect it.

WATCH FOR

Halving the margin requires about four times the sample, and a calculated planning size must be rounded upward.

OPENSTAX WORKED EXAMPLE

Plan a survey of customers aged 50 and older

A mobile phone company wants to estimate the proportion of its customers aged 50 and older who use text messaging. Determine how many completed responses are needed for 90% confidence and a margin of error no greater than three percentage points. No reliable prior estimate is available.

1. Let p be the proportion of the company's customers aged 50 and older who use text messaging. Use E = 0.03, not 3.

2. For 90% confidence, use z* = 1.644854. With no reliable planning estimate, use p_plan = 0.50.

3. Calculate $n \geq (1.644854)^2(0.50)(0.50)/(0.03)^2 = 751.5398$. Round upward to 752 completed responses.

4. Select customers randomly from an appropriate customer list. If sampling without replacement and using the ordinary formula, the population should contain at least 7,520 eligible customers. Check actual success and failure counts after collection.

Conclusion. The plan requires at least 752 usable responses under these assumptions. Sending 752 invitations alone does not guarantee that many responses or eliminate nonresponse bias.

Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Example 8.14, printed pp. 426-427 (PDF pp. 438-439). Question, confidence level, precision target, and answer retained. AP sampling and response qualifications are added.

LOCAL APPLICATION

How large should a trail survey be?

A county wants to estimate the proportion of its 30,000 adult residents who favor a new trail connection. It plans an SRS and wants a 95% margin no greater than 0.04. No reliable prior estimate exists. Find the required completed sample size.

Work backward from a precision target

Planning choice	Worked trail-survey input	Result or reason
Confidence	95%; z* = 1.959964	Chooses the critical value
Target margin	E = 0.04	Four percentage points
Planning proportion	0.50	Maximizes p(1 − p) at 0.25
Required completed n	$(1.959964^2)(0.25)/(0.04^2)$ = 600.2279	Round upward: n = 601
Sampling fraction	601 / 30,000 ≈ 0.0200	Below 10%; random sampling still required

The calculation determines completed observations under the stated model. It does not solve nonresponse or sampling-frame problems.

1. Define p as the proportion of all 30,000 adult residents favoring the connection. Use E=0.04, the 95% critical value z*=1.959964, and planning p=0.50.

2. The one-half value maximizes p (1-p). It protects against choosing too small a sample merely because an unsupported planning estimate is near 0 or 1.

3. Solve n >= (1.959964)^2(0.50)(0.50)/(0.04)^2 =600.2279. Round upward to 601 completed responses, rather than 600.

4. Check the sampling fraction: 601/30,000 is about 2.0%, below 10%. The intended SRS provides a random-sampling basis; actual success/failure counts must still be checked after collection.

5. If only 80% of contacted residents respond, 601/0.80=751.25 suggests contacting at least 752. That is a workload estimate, not a cure for nonresponse bias; follow-up remains essential.

Conclusion. Plan for 601 usable randomly selected responses. Report the sampling frame and follow-up process along with the numerical target, because precision alone does not establish validity.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Example 8.14, printed pp. 426-427 (PDF pp. 438-439). Question, confidence level, precision target, and answer retained. AP sampling and response qualifications are added.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 8.3, Try It 8.14, printed p. 427 (PDF p. 439). Source question and values retained; sampling requirements are requested explicitly.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S06.04 | Estimate a difference in population proportions

CED 3.10, 3.11

- Define an ordered difference of two population proportions.
- Verify independent sampling and observed counts for both groups.
- Construct, interpret, and use a two-proportion z-interval.

CONCEPT AND INTERPRETATION

Define the comparison before subtracting

Let p_1 and p_2 be proportions for two clearly identified populations or treatment conditions, measured using a compatible yes/no outcome. The parameter of interest is $p_1 - p_2$. The order matters: a positive value means the first proportion is larger; a negative value means the second is larger. The point estimate is $\hat{p}_1 - \hat{p}_2$, where each sample proportion uses its own numerator and denominator. Different group sizes do not prevent comparison. However, comparing counts alone can be misleading when denominators differ, and comparing differently defined outcomes can be meaningless.

Justify independence and conditions

The ordinary two-proportion z-interval requires two independent random samples or a randomized experiment with independent groups. When sampling without replacement, check the 10% condition separately for each population: $n_1 \le 0.10N_1$ and $n_2 \le 0.10N_2$. This population sampling check is not needed solely because treatment groups were formed by random assignment. For the normal approximation, each group must have at least 10 observed successes and 10 observed failures. Write all four counts. Repeated yes/no measurements on the same individuals form paired categorical data and do not justify this independent-groups interval.

Estimate the variability of the difference

For independent samples, the estimated variance of the difference is the sum of the two estimated sampling variances. Thus $SE = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$. The interval is $(\hat{p}_1 - \hat{p}_2) \pm z^* SE$. Use the separate sample proportions here; do not pool them. The purpose is to estimate a potentially nonzero difference, not to assume equality. A larger sample from either group reduces its contribution to uncertainty, but uncertainty from the other group remains. Keep enough digits to avoid changing whether an endpoint crosses zero.

Keep the two interval contributions separate

Quantity	Service A	Service B
Observed successes / n	84 / 120	66 / 120
Observed failures	36	54
Sample proportion	0.700	0.550
Variance contribution	0.70(0.30)/120 = 0.001750	0.55(0.45)/120 = 0.0020625
Combined interval SE	sqrt(0.001750 + 0.0020625)	0.0617454

The interval estimates a potentially nonzero difference. Its standard error therefore uses each group's own estimated proportion.

Interpret the sign and size

An interval of (0.03,0.17) for $p_1 - p_2$ estimates that the first population proportion is between 3 and 17 percentage points higher than the second. An interval of (-0.12,-0.02) estimates that it is 2 to 12 percentage points lower. If the interval is (-0.04,0.10), plausible differences include both directions and no difference. Such an interval does not prove equality. Reversing the comparison negates and reverses the endpoints: if the interval for $p_1 - p_2$ is (a, b), then the interval for $p_2 - p_1$ is (-b,-a). Percentage-point differences are not relative percentage increases.

Connect inference to the study design

A difference interval excluding zero supplies evidence that population proportions differ at the stated confidence level. Its sign supports a direction, and its width shows how precisely the difference is estimated. The confidence level still refers to repeated intervals constructed by the same procedure. Random samples support generalization to the respective populations. A randomized experiment can support a treatment-effect conclusion for its participants when design conditions hold. Observational comparisons alone do not establish causation. A statistically supported difference may be too small to matter practically, so compare the interval with relevant decision thresholds as well as zero.

Why subtracting proportions still adds variability

Write a sample proportion as its population proportion plus its sampling error: $\hat{p}_1 = p_1 + e_1$ and $\hat{p}_2 = p_2 + e_2$. Their difference is $(p_1 - p_2) + (e_1 - e_2)$. If the first estimate is unusually high and the second unusually low, the error in the difference is larger. Subtracting the second estimate therefore does not cancel its uncertainty. For independent samples the two sampling variances add. Estimate each variance using that group's sample proportion, then take one square root of the sum. This connects the two-proportion interval to the independent-difference sampling model developed in Unit 5.

KEY IDEA

Specify the subtraction order, use separate sample proportions in the interval standard error, and interpret the whole interval in percentage points.

WATCH FOR

A confidence interval uses unpooled standard error. Do not import the pooled standard error used for an equal-proportions test.

OPENSTAX WORKED EXAMPLE

Compare outcomes for two hives medications

In an OpenStax textbook study, 20 of 200 randomly sampled adults who received medication A still have hives 30 minutes later, compared with 12 of 200 independently sampled adults who received medication B. For this interval exercise, assume each treated-adult population contains at least 2,000 people. Construct and interpret a 95% confidence interval for pA − pB, the difference in the population proportions still having hives.

1. Use a two-sample z interval for pA − pB. A positive difference means a higher proportion still has hives after medication A, not a better outcome for A.

2. The samples are random and independent. Each sample is at most 10% of its treated-adult population. Observed success/failure counts are 20/180 and 12/188; all four are at least 10.

3. The sample proportions are $20/200 = 0.10$ and $12/200 = 0.06$, so the difference is 0.04. The unpooled standard error is $\sqrt{0.10(0.90)/200 + 0.06(0.94)/200} = 0.0270555$.

4. The interval is $0.04 \pm 1.959964(0.0270555)$, or approximately $(-0.0130, 0.0930)$.

5. We are 95% confident that the proportion still having hives after medication A is between 1.3 percentage points lower and 9.3 percentage points higher than the proportion after medication B in the populations represented by the samples.

Conclusion. Because zero lies within the interval, these data do not provide convincing evidence of a population difference at this confidence level. The wording gives random samples, not random treatment assignment, so a causal conclusion is not justified.

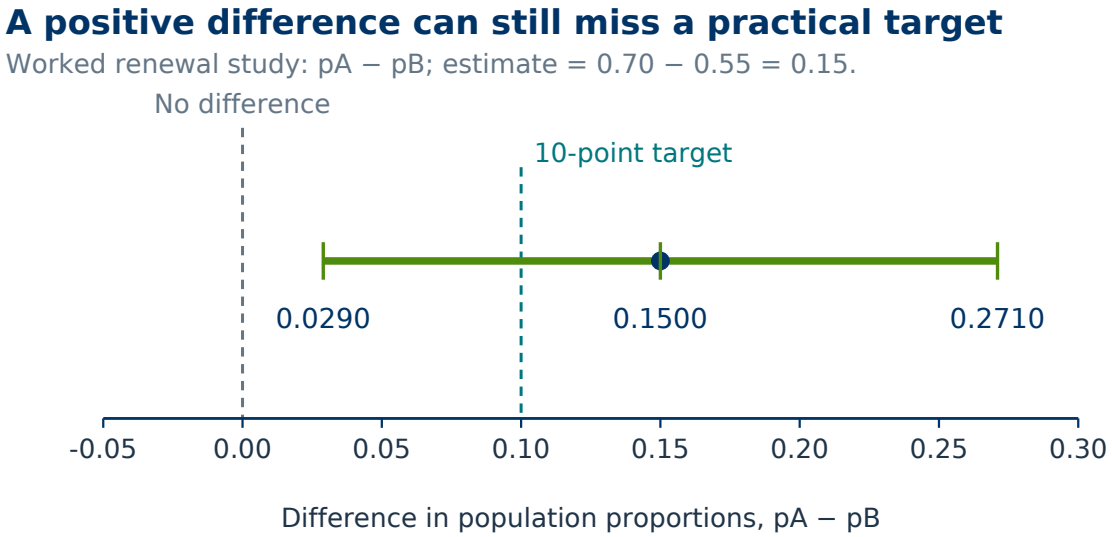
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, Example 10.8, printed pp. 523-524 (PDF pp. 535-536). Original medication outcome and counts retained. The source test is adapted to an interval to teach CED 3.10-3.11; population-size assumptions and unpooled calculation are explicit.

LOCAL APPLICATION

Compare two renewal systems

Illustrative study: independent SRSs select 120 users from each of two regional services with at least 5,000 users each. 84 service-A users and 66 service-B users favor online renewal. Estimate pA - pB with 95% confidence.

Compare the entire interval with the relevant threshold



The 95% interval (0.02898, 0.27102) excludes zero, but contains differences below 0.10. It supports a positive population difference without establishing an advantage of at least ten percentage points.

1. Define pA and pB as the proportions of all users of services A and B who favor online renewal. We seek pA - pB, so positive values favor service A.

2. Independent SRSs satisfy randomization. Each sample 120 is below 10% of at least 5,000. Counts are 84 successes/36 failures for A and 66/54 for B, all at least 10.

3. The estimates are 84/120=0.70 and 66/120=0.55. Their difference is 0.15, or 15 percentage points. Use separate proportions to estimate SE=sqrt (0.70(0.30)/120+0.55(0.45)/120)=0.061745.

4. The 95% margin is 1.959964 times 0.061745=0.121019. The interval is 0.15 +/-0.121019, approximately (0.02898,0.27102).

5. We are 95% confident that the proportion of service-A users favoring online renewal is between about 2.9 and 27.1 percentage points higher than the proportion of service-B users. Every plausible difference is positive.

Conclusion. The data support a population difference, but the wide interval leaves substantial uncertainty about its size. Because the services were not randomly assigned, this does not show that a service caused support.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, Example 10.8, printed pp. 523-524 (PDF pp. 535-536). Original medication outcome and counts retained. The source test is adapted to an interval to teach CED 3.10-3.11; population-size assumptions and unpooled calculation are explicit.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, Try It 10.9, printed pp. 525-526 (PDF pp. 537-538). Original counts and beverage context retained. An interval replaces the source test, and independent random samples and population sizes are stated.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

Unit synthesis

1. Explain how a 95% confidence level differs from a 95% confidence interval.

2. An SRS of 120 of a district's 6,000 adults finds 72 favoring later library hours. Construct and interpret a 95% confidence interval for the population support proportion. Verify the conditions.

3. An analyst plans 95% confidence with margin 0.025 and no prior proportion estimate. Find n.

4. Why are separate observed proportions used in a confidence interval for p1-p2?

5. A difference interval (-0.02,0.07) is reported. Explain why neither equality nor a positive difference is established.

6. A survey's confidence interval is extremely narrow but half the sampled people did not respond. State the additional concern and why the margin does not resolve it.

Unit 07 | Hypothesis Tests for Proportions

A hypothesis test asks whether observed data are sufficiently unusual under a stated null model to support an alternative claim. This unit builds that logic with simulation before formal z procedures. Students test one proportion, compare two proportions, analyze errors and power, and investigate categorical tables with chi-square tests. The study design determines the scope of every conclusion. Statistical significance is evidence against a model, not proof of a claim or a measure of practical importance. Students practice defining parameters, checking the correct conditions, interpreting p-values, and connecting numerical evidence to a careful contextual conclusion.

CED topics: 3.5, 3.6, 3.7, 3.8, 3.12, 3.13, 3.14, 3.15

IN THIS UNIT

Lesson	Focus
S07.01	Hypotheses, null models, and p-values
S07.02	Carry out a one-proportion z test
S07.03	Errors, power, and consequences
S07.04	Test a difference in population proportions
S07.05	Compare categorical distributions with chi-square
S07.06	Investigate association with chi-square

Three stages in using historical records

Enrolled	Surviving records	Usable entries
All students in the target era	Only records retained in the archive	Only records with a value for the stated variable

A missing record and a blank field remove different kinds of information.

LOCAL QUESTION

An archive holds student records from different eras. Which students could those records represent? What evidence would justify testing a claim about a broader population?

S07.01 | Hypotheses, null models, and p-values

CED 3.5, 3.6, 3.7

- Write hypotheses about a population proportion.
- Describe a reproducible simulation under a null model.
- Interpret a p-value and connect it to a predetermined significance level.

CONCEPT AND INTERPRETATION

Start with the investigative question

A significance test evaluates a claim about a population parameter using sample evidence. Define p as a proportion of a specified population with a specified outcome. The null hypothesis, H_0, supplies a reference model; for a one-proportion test write $H_0 : p = p_0$. The alternative, H_a, identifies the departure for which evidence is sought: p greater than, less than, or unequal to p_0. Choose that direction from the research question before examining results. Hypotheses concern the unknown population proportion, not the already observed sample proportion. The data are evidence about the hypotheses; they are not the hypotheses themselves.

What would happen if the null were true?

A null distribution describes how a statistic would vary under the null model. To simulate a sample of n independent binary outcomes under $H_0 : p = p_0$, generate each success with probability p_0, count successes or calculate the sample proportion, and repeat the entire n-outcome trial many times. A complete simulation states the outcome-generating rule, the trial size, the statistic recorded, the number of repetitions, and the extremeness rule. A randomization test for an experiment uses a different mechanism, reassigning treatment labels under a no-effect model. Both approaches ask whether chance variation under the null can plausibly explain the observed result.

Define the p-value conditionally

A p-value is the probability, assuming the null hypothesis is true, of obtaining a test statistic as extreme or more extreme than the observed statistic in the direction specified by the alternative. For a greater-than alternative, use the upper tail; for a less-than alternative, use the lower tail. For a two-sided standardized z test, combine both tails beyond the observed absolute z value. A simulation estimates this probability as the number of qualifying repetitions divided by the total number of repetitions. Small p-values indicate that the observed result would be unusual under the null model and provide evidence for the alternative.

Keep the conditional direction clear

Question	Valid statement	Common reversal
Null model	Assume p = 0.50.	Assume the sample proportion is the population value.
Event being measured	A count of at least 60 in n = 100.	A probability that the null is true.
Simulation estimate	29 qualifying simulated samples / 1,000 simulated samples.	29 / 100 people in the original sample.
Decision at alpha = 0.05	0.029 < 0.05; reject H0.	The null has a 2.9% chance of being true.

A p-value is a probability of data at least this extreme under an assumed model. It is not a probability assigned to a hypothesis.

Make a decision without reversing the probability

The significance level, α, is a decision threshold chosen before analyzing the evidence. Reject H_0 when p-value is at most alpha; otherwise fail to reject it. The p-value is not the probability that H_0 is true, that chance caused the outcome, or that the conclusion is wrong. A large p-value means the data are not unusual enough under the null to provide convincing evidence for the alternative. It does not prove the null. In a contextual conclusion, connect the p-value comparison to evidence about the population parameter, using language such as convincing evidence or insufficient evidence.

Keep the model and the claim in view

A small p-value is meaningful only with an appropriate study design and model. Selection bias, dependence, or a misleading measurement can undermine an inference even if the calculation is correct. Simulations introduce Monte Carlo variability: two runs need not give identical estimated p-values. If no simulated trial is as extreme, report zero out of the stated number and recognize that the true model probability need not be zero. Increasing repetitions improves the probability estimate; increasing the actual study sample changes the evidence. Those are different actions. Statistical significance also does not describe the size or practical importance of a difference.

Follow the source's evidence-to-conclusion sequence

OpenStax organizes a test around a population claim, a model that assumes the null claim, a probability for the observed departure or a more extreme one, and a contextual conclusion. Keep those statements in that order. In a two-sided proportion test, a sample above the null value still requires both equally extreme tails because the research question allows departures in either direction. For example, a sample proportion of 0.53 differs from a null proportion of 0.50 by 0.03; the symmetric lower boundary is 0.47. A p-value of 0.5485 for those tails would be weak evidence against the null, but would not establish that the population proportion equals 0.50.

KEY IDEA

A p-value describes unusual data under a specified null model; it is not a probability assigned to the truth of the null hypothesis.

WATCH FOR

Set the alternative before inspecting the data. Do not switch from two-sided to one-sided merely to obtain a smaller p-value.

OPENSTAX WORKED EXAMPLE

Interpret a two-sided test before calculating one

In an OpenStax textbook scenario, 53 of 100 sampled first-time brides report being younger than their grooms. The research question is whether the population proportion differs from 0.50. Assume the sample was randomly selected from a sufficiently large population. A two-sided normal-approximation test gives a p-value of 0.5485. Use a significance level of 0.01.

1. Let p be the proportion of all first-time brides in the population who are younger than their grooms. State H_0: $p = 0.50$ and H_a: $p \neq 0.50$.

2. The sample proportion is 0.53. Its distance from the null value is 0.03, so equally extreme proportions are at least 0.53 or at most 0.47 under this symmetric normal model.

3. If the population proportion is 0.50, the probability of a sample proportion at least this far from 0.50 in either direction is approximately 0.5485 for random samples of 100.

4. Because $0.5485 > 0.01$, fail to reject H_0. The data do not provide convincing evidence that the population proportion of first-time brides younger than their grooms differs from 0.50.

Conclusion. The nonrejection does not prove that the proportion is one-half. The p-value is conditional on the null model, rather than a probability that the null claim is correct.

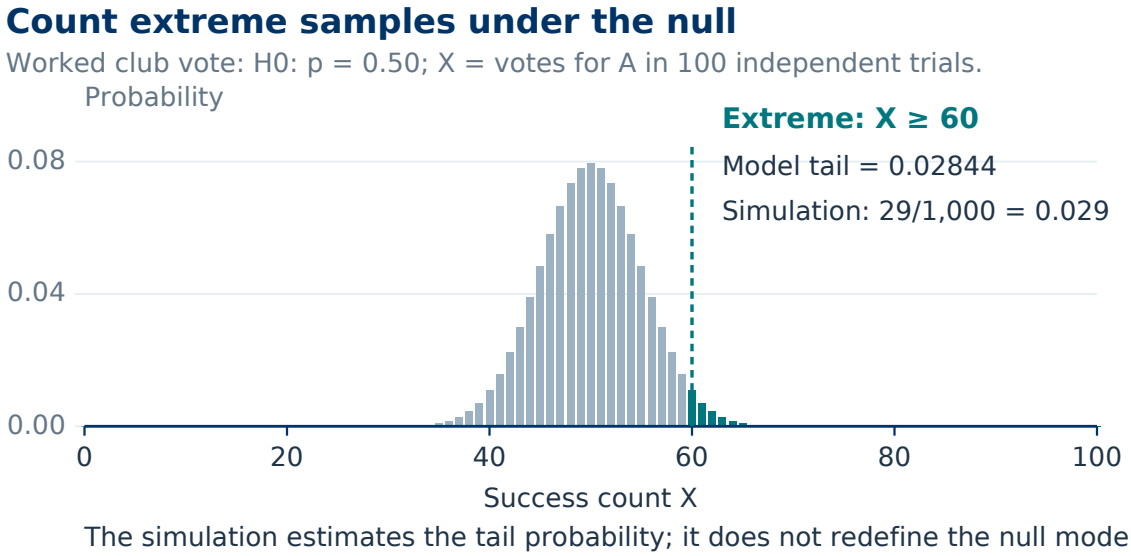
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.5, Example 9.17, printed pp. 476-477 (PDF pp. 488-489). Original counts, benchmark, significance level, and p-value retained. Random sampling is made an explicit classroom assumption; focus is interpretation before the z-test mechanics.

LOCAL APPLICATION

Does a design get more than half the votes?

Illustrative study: an SRS of 100 of a club's 3,000 members gives 60 votes for design A. The question was whether more than half of all members prefer A. In an illustrative 1,000-trial null simulation with $p = 0.50$, 29 trials produce at least 60 A votes. Use alpha = 0.05.

What “as extreme or more extreme” means



The bars show the theoretical Binomial(100, 0.50) model. The shaded bars include 60 through 100 successes. The worked simulation’s 0.029 is close to, but need not equal, the exact model tail of 0.02844.

1. Define p as the proportion of all 3,000 club members who prefer design A. State H0: p = 0.50 and Ha: p > 0.50. The SRS supports inference to these club members, subject to accurate responses.

2. Under H0, one simulated trial generates 100 independent votes, each with probability 0.50 of favoring A, and records the count favoring A. Repeat this full procedure 1,000 times.

3. The observed count is 60. Because the alternative is greater than 0.50, count simulated results at least 60. The estimated p-value is 29/1,000 = 0.029.

4. Interpret this conditionally: If exactly half of all club members prefer A, a random sample of 100 would produce at least 60 A supporters about 2.9% of the time under this simulation model.

5. Since 0.029 < 0.05, reject H0. There is convincing evidence that more than half of the club’s members prefer A. The simulation probability is an estimate, not an exact probability or proof.

Conclusion. The conclusion concerns the club population and the preference question. The calculation does not say there is a 2.9% probability that half the members prefer A.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.5, Example 9.17, printed pp. 476-477 (PDF pp. 488-489). Original counts, benchmark, significance level, and p-value retained. Random sampling is made an explicit classroom assumption; focus is interpretation before the z-test mechanics.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.1, Example 9.1 and Try It 9.4, printed pp. 462-463 (PDF pp. 474-475). Benchmarks retained; locations personalized to southern Oregon. Nulls use the equality form expected in AP work.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S07.02 | Carry out a one-proportion z test

CED 3.5, 3.6, 3.7

- Verify null-based counts for a one-proportion z test.
- Compute z and the appropriate normal-tail probability.
- Write a complete contextual inference response.

CONCEPT AND INTERPRETATION

Match the method and the null model

Use a one-sample z test for a population proportion when a random sample supplies binary outcomes and the question compares the population proportion p with a benchmark p_0. Define p with the population and response. State $H_0 : p = p_0$ and the prespecified alternative. The sample statistic is $\hat{p} = x/n$, but hypotheses concern p. The test asks how far the observed sample proportion is from the benchmark relative to the variation expected under that benchmark. It does not ask whether the sample proportion equals the benchmark exactly; random sampling almost always produces some difference.

Check conditions using the null value

State the random-sampling basis. When sampling without replacement, verify $n \le 0.10N$ using the population and sample sizes. For the normal approximation, use expected counts under the null: $np_0 \ge 10$ and $n(1 - p_0) \ge 10$. These differ from the observed counts used for a confidence interval. The reason is conceptual: a test constructs the distribution that would apply if its null model were true. A valid confidence-interval count check does not automatically establish a valid test check, or the reverse. Report both null expected counts rather than only a sample-size slogan.

Measure the departure in null standard deviations

Under H0, the standard deviation of the sample proportion is approximately $\sqrt{p_0(1 - p_0)/n}$. The test statistic is $z = (\hat{p} - p_0)/\sqrt{p_0(1 - p_0)/n}$. Positive z means the sample proportion exceeds the null value; negative z means it falls below. A z value of 2 means a departure of two null standard deviations above the benchmark, not a two-percentage-point difference. Under the null and appropriate conditions, the z distribution is approximately standard normal. Keep p0 in the denominator: replacing it with the observed proportion calculates a different statistic.

Use the tail selected by the alternative

For $H_a : p > p_0$, the p-value is the area to the right of the observed z. For $H_a : p < p_0$, it is the area to the left. For a two-sided alternative, it is $2P(Z \ge |z|)$. A sample result opposite the one-sided alternative should produce a large p-value, not a small one. Use a calculator's normal distribution or one-proportion test functionality, but record p0, x, n, the alternative, and resulting z and p-value. Never report a very small displayed probability as exactly zero; retain scientific notation or a justified upper bound.

Same statistic, different prespecified questions

Component	Value or rule	Why
Null expected counts	120(0.50) = 60; 120(0.50) = 60	Test conditions use p0, not p-hat.
Ha: p > 0.50	Right tail: 0.01423	Departure in the positive direction
Ha: p < 0.50	Left tail: 0.98577	Observed result is opposite this alternative.
Ha: p ≠ 0.50	Both tails: 0.02846	Departure in either direction

These are different questions that must be selected before inspecting the sample. Choosing the smallest p-value afterward changes the inferential procedure.

Connect the decision to the evidence

Compare the p-value with the previously selected alpha. Reject H0 if p-value is at most alpha; otherwise fail to reject it. Then answer the original question with the population and direction of the alternative. For example: Because 0.014 < 0.05, there is convincing evidence that more than half of the club population favors the plan. A statistically significant result is not proof and does not establish a large practical effect. A nonsignificant result does not prove equality. Study design still controls generalization and causation. Report the observed proportion as useful descriptive context alongside the inferential conclusion.

Keep the population model separate from the observed sample

The one-proportion examples in OpenStax use the benchmark from H_0 to construct a normal model for $\hat{p}$. The observed sample proportion enters the numerator of the standardized statistic, while the null proportion enters its standard deviation. For a benchmark of 0.92 and n = 200, the null expected count of successes is 184 even if the observed count is 174. It is that ten-count departure which the test evaluates. The confidence-interval calculation instead uses the observed proportion to estimate uncertainty about an unknown p. Naming the question first prevents the two denominators from becoming interchangeable rules.

KEY IDEA

A one-proportion test uses the null value in both its expected-count check and its standard-deviation denominator.

WATCH FOR

The test denominator is based on p0; the interval standard error is based on the observed sample proportion.

OPENSTAX WORKED EXAMPLE

Test a claimed cell-phone ownership rate

In an OpenStax textbook scenario, a manufacturer investigates whether the proportion of U.S. adults who own a cell phone is below 0.92. Of 200 surveyed adults, 174 report owning one. For the classroom analysis, assume the adults form an SRS, all respond, and the population is at least 2,000. Do the data provide convincing evidence for the manufacturer's claim at the 0.05 significance level?

1. Let p be the proportion of all U.S. adults in the textbook scenario who own a cell phone. State H_0: p = 0.92 and H_a: p < 0.92. Use a one-proportion z test.

2. Random selection is assumed. The sample is at most 10% of the population. Under H_0, expected successes and failures are 200(0.92) = 184 and 200(0.08) = 16, both at least 10.

3. The observed proportion is $174/200 = 0.87$. The null standard deviation is $\sqrt{0.92(0.08)/200} = 0.0191833$, giving $z = (0.87 - 0.92)/0.0191833 = -2.60643$.

4. The alternative is less than, so the p-value is $P(Z \le -2.60643) \approx 0.00457$. If the population ownership rate were 0.92, a random sample of 200 would have an ownership proportion of 0.87 or lower about 0.46% of the time under the normal approximation.

5. Because 0.00457 < 0.05, reject H_0. The data provide convincing evidence that the population ownership proportion in this scenario is below 0.92.

Conclusion. The 0.92 benchmark belongs to the textbook problem. It is not presented as a verified current ownership rate. Evidence that a proportion is lower does not by itself establish whether the difference matters to a business decision.

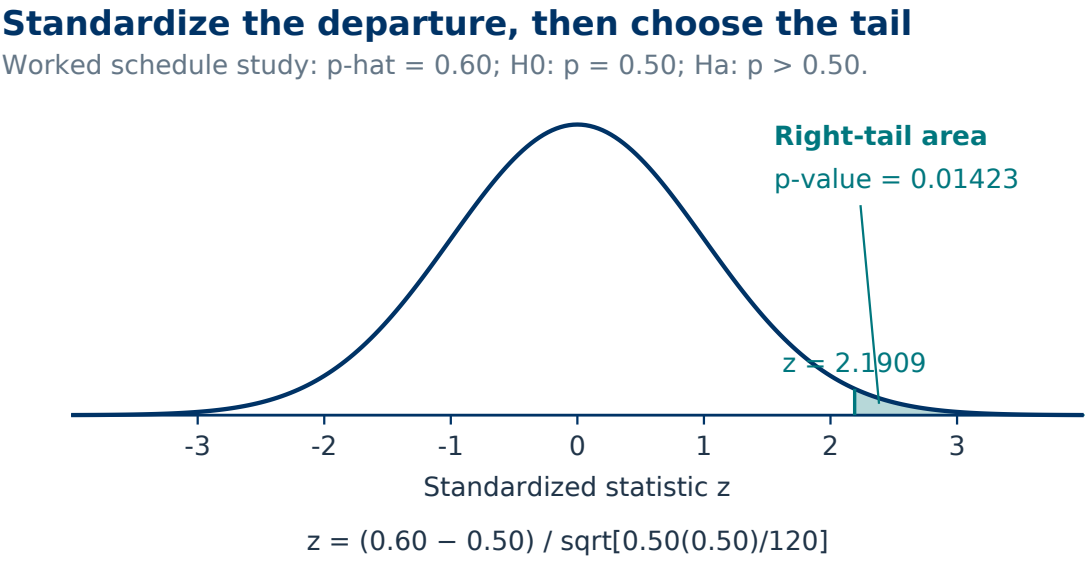
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.5, Try It 9.18, printed p. 478 (PDF p. 490). Original counts, benchmark, direction, and significance level retained. Sampling and response assumptions are explicit; full AP reasoning supplied.

LOCAL APPLICATION

Is support above one-half?

Illustrative study: an SRS of 120 of a club's 6,000 members finds 72 favoring a proposed schedule. Before collecting data, the club selected Ha: p > 0.50 and alpha = 0.05. Conduct a one-proportion z test.

The alternative chooses the shaded probability



The sample proportion is about 2.19 null standard deviations above 0.50. The greater-than alternative uses the right-tail area, approximately 0.01423.

1. Classify. Each sampled club member answered yes or no. This is a sample survey of one population.

2. Choose. Let p be the proportion of all 6,000 members who favor the schedule. Use a one-proportion z test with H_0: p = 0.50 and H_a: p > 0.50 at $\alpha = 0.05$. A high sample proportion favors H_a.

3. Check. The survey used an SRS of 120, only 2% of the roster. Under H_0 the expected counts are 120(0.50) = 60 for each response, both at least 10.

4. Calculate. The sample proportion is 72/120 = 0.60. Its null standard deviation is sqrt(0.50 × 0.50 / 120) = 0.04564. The statistic is z = (0.60 – 0.50)/0.04564 = 2.1909. The right-tail p-value is P(Z ≥ 2.1909) = 0.0142.

5. Conclude. The p-value, 0.0142, is below 0.05. Reject H_0. These data give convincing evidence that more than half of the club's 6,000 members favor the schedule.

Conclusion. If support were evenly split, the normal approximation puts the chance of a sample proportion at least as high as 0.60 at about 1.4%.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.5, Try It 9.18, printed p. 478 (PDF p. 490). Original counts, benchmark, direction, and significance level retained. Sampling and response assumptions are explicit; full AP reasoning supplied.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.5, Example 9.18, printed pp. 477-478 (PDF pp. 489-490). Original count, sample size, benchmark, and context retained; random sampling, population size, and alpha are specified.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S07.03 | Errors, power, and consequences

CED 3.8

- Describe Type I and Type II errors in context.
- Interpret power at a specified alternative value.
- Calculate error probabilities from power or simulation and explain design tradeoffs.

CONCEPT AND INTERPRETATION

A test decision can be wrong

A hypothesis test makes a decision from incomplete data. A Type I error is rejecting H0 when H0 is true. A Type II error is failing to reject H0 when the relevant alternative is true. These are conditional events, not labels for arithmetic mistakes. To describe either error in context, state both the decision and the underlying truth. For a test asking whether a process's defect proportion exceeds 0.02, a Type I error declares excess defects when the true proportion is 0.02; a Type II error fails to detect excess defects when the true proportion is greater than 0.02.

Alpha controls false positives under the null

The significance level alpha is the planned probability of a Type I error under the null model for a correctly calibrated test. A nominal 0.05 test is designed to reject about 5% of the time when H0 is true. Approximate and discrete procedures may have actual rates that differ somewhat from the nominal value. Alpha is chosen before data analysis in light of the consequences of false positives. It is not the probability that a particular rejection is wrong. A realized p-value is evidence from the observed data; alpha is the decision rule to which that evidence is compared.

Power depends on a specified alternative

Power is the probability of correctly rejecting a false H0. If beta is the probability of a Type II error at a specified true parameter value, then power is $1-\beta$. A claim of 80% power is incomplete unless it identifies the alternative effect, sample size, significance level, and procedure. A test may have low power to detect p=0.51 against p0=0.50 but much higher power to detect p=0.70. When interpreting power, begin with the assumed true alternative value, then state the probability that the procedure will produce evidence for the alternative.

Separate decision from truth

Decision	H0 is true	Specified alternative is true
Reject H0	Type I error; probability alpha	Correct detection; probability power
Fail to reject H0	Correct nonrejection; probability 1 − alpha	Type II error; probability beta
Planning change	Smaller alpha reduces false positives.	Larger n generally increases power, other factors fixed.

Power and beta must be tied to a specified alternative value and testing procedure. The realized p-value is not an entry in this probability table.

Change the design to improve detection

Holding other features fixed, larger samples usually reduce standard error and increase power. A true effect farther from the null is easier to detect. Increasing alpha enlarges the rejection region, increases power, and also raises the Type I error rate. Reducing alpha decreases false positives but generally increases Type II errors at a fixed alternative and sample size. Better measurement or a design that reduces relevant variability can improve power. These are comparisons holding other factors fixed; changing several features at once can create competing effects. Choose the direction of the alternative from the research question, not as a later device for increasing power.

Use consequences and simulations carefully

A simulation can estimate error rates by repeatedly applying the entire testing rule to samples generated under known truths. Under H0, the rejection fraction estimates Type I error probability. Under a specified alternative, the rejection fraction estimates power and the nonrejection fraction estimates beta. These simulations do not use the same generating proportion. Serious consequences from a false positive may justify a smaller alpha; serious consequences from a missed effect may justify larger sample size. Statistical evidence alone does not determine which consequence matters more. A transparent decision states the tradeoff, the assumptions, and the practical magnitude worth detecting.

Describe both the decision and the truth

OpenStax uses the safety of climbing equipment to separate two errors. A false alarm occurs when a safety check flags equipment that is actually safe. A missed danger occurs when the check fails to flag equipment that is actually unsafe. A statement such as “the equipment is unsafe” alone cannot identify an error, because it omits either the decision or the truth. In a statistical test, make the same distinction using the population parameter. Then state a possible practical consequence. The consequence helps a decision maker choose a significance level and sample size, but it does not change which logical event is called Type I or Type II.

KEY IDEA

Type I error conditions on a true null; power and Type II error condition on a specified false-null situation.

WATCH FOR

A large p-value may reflect low power. It is not evidence that the population effect is exactly zero.

OPENSTAX WORKED EXAMPLE

A climbing-equipment safety decision

OpenStax considers the null claim that a climber's equipment is safe. A safety procedure can reject that claim or fail to reject it. Describe Type I and Type II errors and a possible consequence of each. Explain why a nonrejection is not a guarantee of safety.

1. A Type I error rejects a true null claim: the procedure flags the equipment as unsafe even though it is safe.

2. A possible consequence is an unnecessary inspection or replacement, or cancellation of a planned climb.

3. A Type II error fails to reject a false null claim: the procedure fails to flag equipment that is actually unsafe.

4. A possible consequence is that the climber uses unsafe equipment. In this setting, that missed danger could have more serious consequences than an unnecessary inspection.

5. Failing to reject means that the procedure did not detect evidence against the safety claim. It does not prove safety. Reducing missed dangers may require better measurements or more testing, rather than merely declaring a smaller significance level.

Conclusion. Error descriptions must include the decision and the actual condition. Their practical costs are contextual judgments, while their Type I or Type II classification follows from the stated null claim.

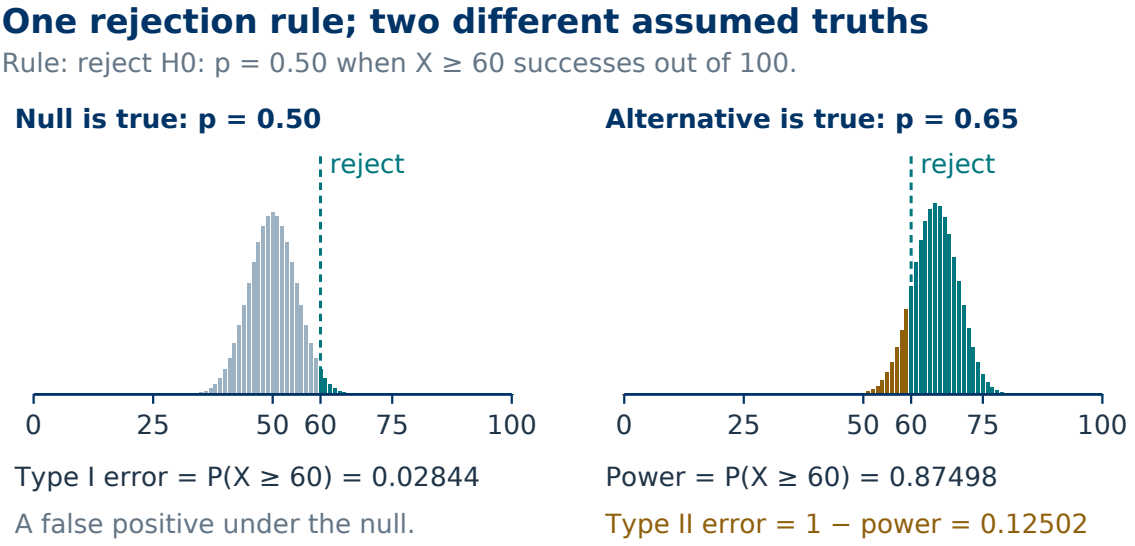
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.2, Example 9.5, printed p. 464 (PDF p. 476). Original climbing-equipment scenario retained; conclusions use fail-to-reject language and distinguish the logical error from its possible consequence.

LOCAL APPLICATION

A rule for detecting a biased binary process

Consider an illustrative process with independent binary trials. Test H0: $p = 0.50$ against $p > 0.50$ using 100 trials. A proposed rule rejects when at least 60 successes occur. Evaluate its actual Type I error probability and power at $p = 0.65$.

Power changes the assumed truth, not the rejection rule



Both panels use the exact binomial model and the same 60-success threshold. The null panel measures false positives. The p = 0.65 panel measures successful detection and missed effects.

1. Under H0, the success count X has a binomial distribution with n=100 and p=0.50. A Type I error occurs exactly when X>=60 even though p=0.50.

2. Calculate P (X>=60) using the upper binomial tail, equivalently 1-P (X<=59). The result is 0.028444. This is the actual false-positive probability for this particular count threshold.

3. For power at p=0.65, keep the same sample size and rejection rule but generate X from a binomial model with success probability 0.65. Compute P (X>=60)=0.874977.

4. Interpret power: If the process truly succeeds 65% of the time, this 100-trial rule detects a departure above 50% in about 87.5% of repeated studies.

5. The Type II error probability at p=0.65 is 1-0.874977=0.125023. It represents failing to detect the higher success proportion when the true value is 0.65.

Conclusion. The rule's actual false-positive rate is about 2.84%, and its power at 0.65 is about 87.5%. Neither number is the probability that a particular test conclusion is correct.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.2, Example 9.5, printed p. 464 (PDF p. 476). Original climbing-equipment scenario retained; conclusions use fail-to-reject language and distinguish the logical error from its possible consequence.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 9.2, Example 9.8, printed p. 466 (PDF p. 478). Source treatment benchmark and error comparison retained. Equality null, explicit population parameter, and careful nonrejection wording supplied.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S07.04 | Test a difference in population proportions

CED 3.12, 3.13

- Write ordered hypotheses for two population proportions.
- Use a pooled proportion for null expected counts and standard error.
- Conduct and interpret a two-proportion z test with appropriate scope.

CONCEPT AND INTERPRETATION

State the populations and equality claim

A two-sample z test for proportions compares two independent groups with a binary outcome. Define p_1 and p_2 in context and specify the order of subtraction. The null is $H_0 : p_1 - p_2 = 0$, equivalent to equal population proportions. A one-sided alternative claims the ordered difference is greater than or less than zero; a two-sided alternative claims it is not zero. Use a compatible outcome definition in both groups. The observed difference $\hat{p}_1 - \hat{p}_2$ is the evidence. It is not itself a population hypothesis, and unequal raw success counts do not establish unequal proportions.

Why the test pools the samples

Under H0, both groups have the same unknown population proportion. Estimate that common value by pooling successes: $\hat{p}_c = (x_1 + x_2)/(n_1 + n_2)$. This is a sample-size-weighted proportion, not generally the simple average of the two sample proportions. Pooling is justified by the equality assumption used to construct the null distribution. A confidence interval estimates an unrestricted difference and therefore does not impose this equality. The contrast between pooled test and unpooled interval standard errors follows from their different questions; it is not an arbitrary calculator convention.

Check null expected counts and design

The data should come from independent random samples or a randomized experiment with independent groups. When sampling without replacement, check the 10% condition for each sampled population. A population 10% check is not needed simply because participants were randomly assigned to treatments. Under the equal-proportions null, check $n_1\hat{p}_c$, $n_1(1-\hat{p}_c)$, $n_2\hat{p}_c$, and $n_2(1-\hat{p}_c)$; each must be at least 10. These are pooled expected counts. Do not substitute observed successes and failures as if this were an interval check. Paired yes/no responses do not satisfy independent-group assumptions.

A test and an interval use different variance estimates

Question	Proportion(s) in the SE	Count check
Test H0: p1 = p2	Pooled: (x1 + x2)/(n1 + n2)	Four expected counts under the pooled null
Estimate p1 − p2 with an interval	Separate observed p-hat 1 and p-hat 2	Four observed success/failure counts
Reminder test example	Pooled proportion = 160/320 = 0.50	80, 80, 80, 80
Reminder interval example	Separate proportions = 0.60 and 0.40	96, 64, 64, 96

The difference follows from the inferential question. An interval does not impose equality; the equality test does.

Compute the test statistic and probability

The null standard error is $SE_0 = \sqrt{\hat{p}_c(1-\hat{p}_c)(1/n_1 + 1/n_2)}$. Calculate $z = ((\hat{p}_1 - \hat{p}_2) - 0)/SE_0$. Under H0 and the conditions, z has an approximately standard normal distribution. Use a right, left, or two-sided normal probability according to the prespecified alternative. Record group labels, successes, sample sizes, and direction when using technology. Reversing the group order changes the sign of z and the corresponding one-sided direction, but does not change a two-sided p-value. An unexplained test-menu result is not a complete statistical argument.

Interpret evidence within the design

Interpret the p-value while assuming the two population proportions are equal. Compare it with alpha and state whether there is convincing evidence for the specified population difference. Random samples support generalization to their sampling populations. Random treatment assignment supports a causal treatment comparison for the study participants, assuming the experiment was properly implemented. Broad generalization additionally requires attention to recruitment. Do not conclude that programs caused a difference from observational comparisons. Failure to reject equality does not prove the proportions match. Statistical significance should be accompanied by an estimated difference and, when useful, an interval describing plausible practical sizes.

Connect pooling to an expected-count table

OpenStax explains that the equality null gives the two groups one common, unknown success proportion. The pooled estimate uses all successes divided by all observations. Multiplying this pooled proportion by each group's size gives the expected successes in that group if the null model were true. For 15 cracked valves among 100 type-A valves and 6 among 100 type-B valves, the pooled estimate is $21/200 = 0.105$. Both groups then have 10.5 expected cracked valves and 89.5 expected uncracked valves. The observed count of six does not prevent this particular test from meeting the AP null expected-count condition. The same data would fail the ordinary confidence-interval observed-count condition for group B. The two checks answer different questions.

KEY IDEA

The equal-proportions null justifies a common pooled estimate for the test, but the conclusion must still respect the sampling or experimental design.

WATCH FOR

Pool counts, not percentages: (x1+x2)/(n1+n2). Check the four pooled expected counts for this test.

OPENSTAX WORKED EXAMPLE

Compare reported seat-belt nonuse

An OpenStax textbook survey records 183 of 2,231 men and 156 of 2,169 women reporting that they do not wear seat belts. Treat these as independent random samples from large U.S. adult populations. Is there convincing evidence that the population proportion of women who do not wear seat belts is lower than the corresponding proportion of men? Use alpha = 0.01.

1. Let pW and pM be the population proportions of women and men, respectively, who report not wearing seat belts. Test H_0: pW − pM = 0 against H_a: pW − pM < 0 using a two-proportion z test.

2. Independent random sampling is assumed. The sampled populations are sufficiently large for each sample to be below 10%. The pooled estimate is (156+183)/(2169+2231) = 339/4400 = 0.0770455.

3. The pooled expected counts are approximately 167.1116 women reporting nonuse and 2,001.8884 reporting use; for men they are 171.8884 and 2,059.1116. All four exceed 10.

4. The observed difference is $156/2169 - 183/2231 = -0.0101035$. The pooled standard error is $\sqrt{0.0770455(0.9229545)(1/2169 + 1/2231)} = 0.0080410$. Thus $z = -1.25649$.

5. Use the left tail. The p-value is approximately 0.10447, greater than 0.01. Fail to reject H_0. The data do not provide convincing evidence that seat-belt nonuse is less common among women than among men in the populations represented by this textbook survey.

Conclusion. The sample difference is about one percentage point, but it is not statistically significant at the selected level. The test does not establish equality, and this historical textbook dataset is not presented as a current U.S. rate.

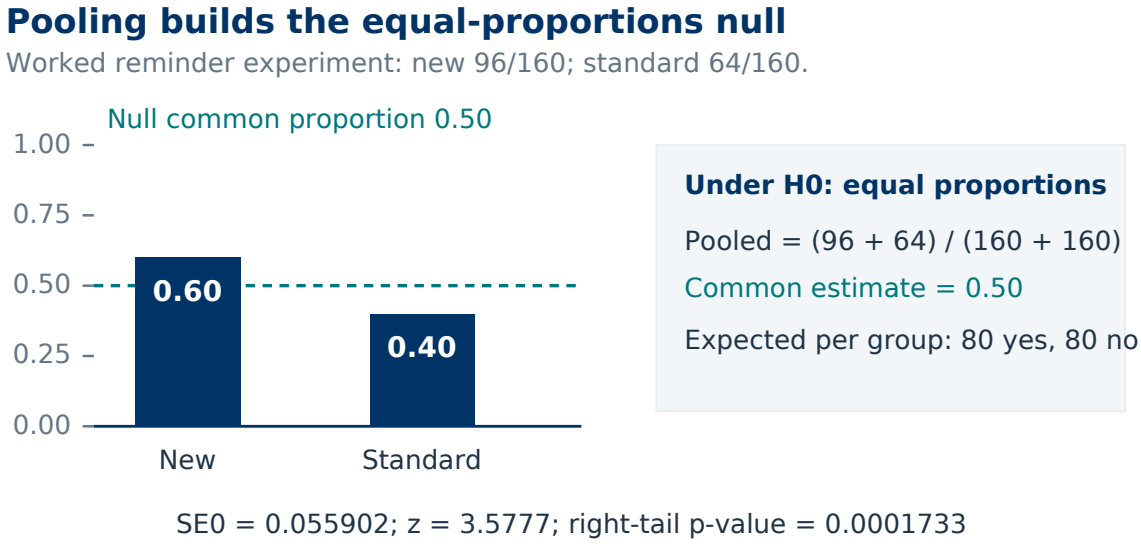
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, Example 10.9, printed pp. 525-526 (PDF pp. 537-538). Original sample counts, direction, and significance level retained. AP pooled expected-count checks and contextual qualifications are supplied.

LOCAL APPLICATION

Does a new reminder improve response?

Illustrative randomized experiment: 320 volunteers are assigned at random, 160 to a new reminder and 160 to the standard reminder. 96 new-reminder participants and 64 standard-reminder participants respond within 48 hours. Test for an improvement at alpha 0.05.

Observed difference versus the null distribution



The observed proportions remain 0.60 and 0.40. Pooling supplies a common proportion only for constructing the equal-proportions null model; it does not erase the observed difference.

1. Let pN and pS represent response probabilities under the new and standard reminders for the study population represented by these volunteers. State H0: pN-pS=0 and Ha: pN-pS>0. Use a two-proportion z test.

2. Random assignment provides a treatment comparison. Assume that different participants respond independently and do not influence one another's responses. The 10% sampling condition is not required merely for this randomized experiment. The pooled estimate is (96+64)/(160+160)=0.50; pooled expected successes and failures are 80 in each group, all at least 10. Here the simple average of the two sample proportions happens to agree because the groups are the same size.

3. The observed proportions are 0.60 and 0.40, so the estimated improvement is 0.20, or 20 percentage points. Under H_0, SE_0 = sqrt (0.50(0.50)(1/160+1/160))=0.055902.

4. The statistic is z=0.20/0.055902=3.577709. The improvement alternative uses the right tail: p-value=0.00017331.

5. Because this p-value is below 0.05, reject H0. There is convincing evidence that the new reminder increases 48-hour response for participants under these experimental conditions. Random assignment supports a causal treatment comparison.

Conclusion. The 20-point sample improvement is substantial descriptive context. Generalizing to all future recipients requires caution because volunteers were recruited rather than randomly sampled from that broader population.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, Example 10.9, printed pp. 525-526 (PDF pp. 537-538). Original sample counts, direction, and significance level retained. AP pooled expected-count checks and contextual qualifications are supplied.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 10.3, Try It 10.8, printed p. 524 (PDF p. 536). Counts, pressure threshold, and test question retained. Independent random samples, population sizes, and AP pooled expected-count checks are explicit.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S07.05 | Compare categorical distributions with chi-square

CED 3.14, 3.15

- Identify a chi-square test for homogeneity from the study design.
- Compute expected counts, cell contributions, and degrees of freedom.
- Conduct and interpret a homogeneity test with complete conditions.

07



Asking about transit preferences

CONCEPT AND INTERPRETATION

Compare distributions, not just one success rate

A chi-square test for homogeneity compares the distribution of one categorical variable across two or more populations or treatment groups. Each individual contributes to one mutually exclusive response category. Independent random samples might ask residents of two districts for a preferred transit option; a randomized experiment might compare response-category distributions under several treatments. The null states that the population or treatment distributions are the same across groups. The alternative states that at least one distribution differs. It does not require every group or every category to differ. The design and question identify the method, not just the table dimensions.

Expected counts express the null model

Organize observed counts in an r-by-c table, with groups as rows and response categories as columns. Under the same-distribution null, the pooled column proportions estimate the shared response distribution. The expected count in a cell is $E = (\text{row total})(\text{column total})/(\text{grand total})$. The expected table preserves the observed row and column totals. Expected counts can be nonintegers because they are model expectations, not actual individuals. Use counts rather than percentages in the test calculation. A large absolute discrepancy is interpreted relative to its expected count, so a discrepancy of five matters differently when E is ten versus one hundred.

Check design, sampling, and expected counts

For homogeneity, data should come from independent random samples or a randomized experiment. When random samples are drawn without replacement, verify that each sample is no more than 10% of its respective population. This population sampling check is not required simply for randomized treatment assignment. For this course's supplied Fall 2026 CED, all expected counts should be greater than 5. Check expected, not observed, counts. Small expected cells make the usual chi-square approximation unreliable. Do not combine categories after seeing which combination gives significance; any defensible regrouping must preserve the investigative meaning and be handled transparently.

Measure the aggregate discrepancy

The test statistic is $\chi^2 = \sum(O - E)^2/E$, summing across every cell. Each contribution is nonnegative, and larger values indicate more discrepancy from the null model. The approximate null distribution is chi-square with $df = (r - 1)(c - 1)$. Chi-square density curves are right-skewed, with the skew decreasing as degrees of freedom increase; the statistic itself can equal zero if observed and expected counts match exactly. The p-value is the right-tail area beyond the observed statistic. It is not a two-tailed normal probability, and the signs of individual O-E differences disappear when squared.

Conclude about the distribution as a whole

Compare the right-tail p-value with alpha and answer whether the data provide convincing evidence that the categorical distributions differ across the specified populations or treatments. Interpret the p-value assuming the null distributions are the same. A significant omnibus test identifies an overall inconsistency with homogeneity; it does not by itself establish a separate significance claim for every cell or every pair of groups. Inspecting large contributions helps describe where discrepancies occur, but formal follow-up testing requires additional care. Randomized treatment assignment can support a causal comparison; independent observational samples generally support population comparisons, not causal claims.

Build expected counts from a common distribution

OpenStax's housing example samples 250 men and 300 women at a college. If the population living-arrangement distributions are the same, both samples should follow a common set of category proportions, allowing for random variation. Pooling the samples gives 163 dormitory residents among 550 students. The null therefore predicts a dormitory proportion of $163/550$ in each group. Multiplying by the group sizes gives $250(163/550) = 74.09$ expected men and $300(163/550) = 88.91$ expected women in dormitories. Their expected counts differ because the sample sizes differ, even though the expected proportions agree. Apply the same logic to every category before measuring the total discrepancy.

KEY IDEA

Homogeneity asks whether one categorical response has the same distribution across populations or treatments.

07

WATCH FOR

Check expected counts, use every cell in the sum, and take a right-tail chi-square probability with (rows -1)(columns -1) degrees of freedom.

OPENSTAX WORKED EXAMPLE

Compare college living arrangements

OpenStax gives independent random samples of 250 men and 300 women college students. The category order is dormitory, apartment, with parents, other. The observed counts are men: 72, 84, 49, 45; women: 91, 86, 88, 35. Assume the two student populations contain at least 2,500 and 3,000 students, respectively. At alpha = 0.05, do the data provide convincing evidence that the population distributions of living arrangements differ?

1. Use a chi-square test for homogeneity. H_0: the distribution of living arrangements is the same in the two college-student populations. H_a: the population distributions differ.

2. The samples are independent and randomly selected. Both sampling fractions are at most 10%. The column totals are 163, 170, 137, and 80; the grand total is 550.

3. Use E = (row total)(column total)/550. The expected row for men is approximately (74.0909, 77.2727, 62.2727, 36.3636). The expected row for women is (88.9091, 92.7273, 74.7273, 43.6364). Every expected count exceeds 5.

4. For example, the men-with-parents contribution is $(49-62.2727)^2/62.2727 \approx 2.8289$. Sum all eight cell contributions to obtain $\chi^2 = 10.1287$. The degrees of freedom are $(2-1)(4-1) = 3$.

5. The right-tail p-value is $P(\chi^2_3 \geq 10.1287) \approx 0.01750$. Since $0.01750 < 0.05$, reject H_0. There is convincing evidence that the distributions of living arrangements differ between the two populations of college students.

Conclusion. The test detects an overall distributional difference. It does not establish a separate statistically significant difference in every category or show that gender causes living arrangements.

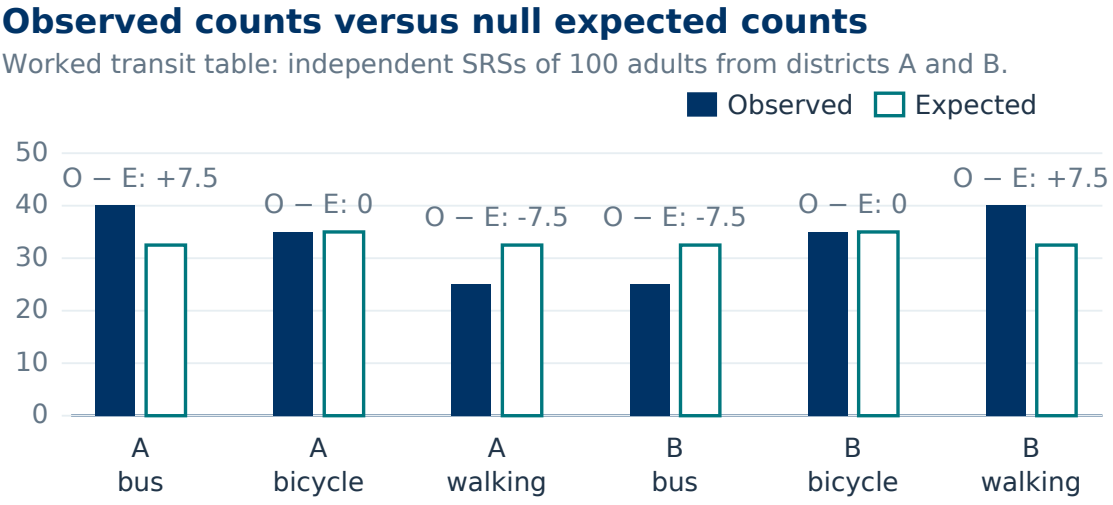
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 11.4, Example 11.8, Table 11.19, printed pp. 579-580 (PDF pp. 591-592). Source counts, categories, and inference retained. AP design/expected-count checks and general degrees-of-freedom formula are made explicit.

LOCAL APPLICATION

Transit preferences in two districts

Illustrative independent SRSs select 100 adults from each of two districts with at least 5,000 adults. Preferred options are bus, bicycle, or walking. District A counts are 40,35,25; district B counts are 25,35,40. Test homogeneity at alpha 0.05.

Expected counts describe the same-distribution null



Signed differences show direction; squared contributions measure discrepancy.

The null predicts 32.5, 35, and 32.5 in either district. The bicycle cells match expectation exactly. Bus and walking cells depart in opposite directions while contributing equally to chi-square.

Every cell contributes to the statistic

Cell	Observed O	Expected E	$(O - E)^2/E$
A: bus	40	32.5	1.730769
A: bicycle	35	35	0.000000
A: walking	25	32.5	1.730769
B: bus	25	32.5	1.730769
B: bicycle	35	35	0.000000
B: walking	40	32.5	1.730769
Total statistic			6.923077

Degrees of freedom are (2 − 1)(3 − 1) = 2. The right-tail p-value is 0.031381; the overall test does not separately test each response category.

1. H0: the distribution of preferred transit option is the same for all adults in districts A and B. Ha: the distributions differ. Use a chi-square test for homogeneity because separate populations were independently sampled for one categorical response.

2. Both samples are random and independent, and 100 is below 10% of each population. Row totals are 100,100; column totals 65,70,65; grand total 200. Expected counts in each row are 32.5,35,32.5, all greater than 5.

3. For district A, the contributions are (40-32.5)^2/32.5=1.730769, zero for the 35 cell, and (25-32.5)^2/32.5=1.730769. District B contributes the same three values.

4. Adding the six cell contributions gives $\chi^2 = 6.923077$. The degrees of freedom are $(2 - 1)(3 - 1) = 2$. The p-value is the right-tail probability beyond 6.923077 under a chi-square distribution with 2 degrees of freedom: approximately 0.031381.

5. Since 0.031381<0.05, reject H0. There is convincing evidence that the population distribution of preferred transit option differs between the two districts. The largest discrepancies concern bus and walking preferences.

Conclusion. At the 5% level, the survey gives evidence that preferred transit options are distributed differently in the two districts ($p \approx 0.031$). This overall test does not establish a difference for each option or a causal effect of district residence.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 11.4, Example 11.8, Table 11.19, printed pp. 579-580 (PDF pp. 591-592). Source counts, categories, and inference retained. AP design/expected-count checks and general degrees-of-freedom formula are made explicit.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 11.4, Try It 11.8, Table 11.20, printed p. 580 (PDF p. 592). Original counts and categories retained. Observational units are clarified as households with one primary vehicle, and AP conditions are explicit.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

S07.06 | Investigate association with chi-square

CED 3.14, 3.15

- Distinguish independence from homogeneity using sampling design.
- Conduct a chi-square independence test and interpret the p-value.
- Select an appropriate categorical inference procedure and describe its limits.

CONCEPT AND INTERPRETATION

Two categorical variables in one population

A chi-square test for independence asks whether two categorical variables are associated in one population. A single random sample is drawn, and each individual is classified on both variables. For example, a district may sample adults and record both commute mode and whether they have a flexible work schedule. H0 states that the two variables are independent in that population; Ha states that they are associated. Independence means that the distribution of one variable is the same across levels of the other, or equivalently that relevant conditional probabilities equal the corresponding marginal probabilities.

Design distinguishes the two chi-square tests

Homogeneity uses separate independent samples from populations, or random assignment to treatments, to compare the distribution of one categorical response. Independence uses one random sample from one population and measures two categorical variables. The table calculations, degrees of freedom, and right-tail probability can be identical even though the study designs and hypothesis wording differ. Do not decide the test name from whether the table has two rows, or from which variable is displayed first. Describe how individuals entered the study and what was measured. That design determines the inferential question and the scope of the conclusion.

Check the model before computing

For independence, state the random sample from the target population. When sampling without replacement, check $n \le 0.10N$ for the entire sample, not separately for observed subgroups formed after sampling. Individuals should contribute independent observations and one cell each. Compute each expected count using row total times column total divided by grand total. Under the supplied Fall 2026 CED condition, every expected count should be greater than 5. Repeated measurements on the same people, overlapping categories, or percentage entries treated as counts can violate the ordinary table setup even when software returns a result.

Calculate and interpret the association test

Calculate $\chi^2 = \sum (O-E)^2/E$ and $df = (r-1)(c-1)$. The p-value is $P(\chi^2_{df} \ge \chi^2_{obs})$ under the independence model. Interpret it as the probability, assuming the population variables are independent, of a table discrepancy at least as large as observed according to this statistic. Compare with the predetermined alpha. If rejecting, conclude there is convincing evidence of association between the named variables in the named population. If not rejecting, state insufficient evidence of association. Neither decision shows that one categorical variable causes the other.

Select a method and communicate limits

Use a one-proportion z test to compare a single population proportion with a fixed benchmark. Use a two-proportion z test for a directional or two-sided comparison of two independent binary proportions. Use a chi-square test when the investigative question concerns categorical distributions or association in a table, especially with more than two response categories. An uncorrected two-by-two chi-square test matches the two-sided pooled z test in p-value, but not a one-sided z test. Always inspect the design first. Statistical association can reflect confounding, selection effects, or a causal relationship; the test alone cannot distinguish those explanations.

Derive an expected count from independence

OpenStax connects contingency tables to the probability rule $P(A \cap B) = P(A)P(B)$ for independent events. If a sample of 755 drivers contains 70 with a speeding violation and 305 who report phone use while driving, the null model estimates the joint proportion as $(70/755)(305/755)$. Multiplying by 755 gives an expected overlap of $70(305)/755 \approx 28.28$ drivers. The overlap observed in a real table may differ. That difference is the evidence the test evaluates. Marginal totals alone determine an expected count, but without the actual cell counts they do not determine the test statistic or prove association.

07

KEY IDEA

The same table arithmetic can answer different questions: one sample and two variables suggests independence; independent samples and one response suggests homogeneity.

WATCH FOR

A significant independence test establishes evidence of association, not causation. Apply the 10% check to the full single sample.

OPENSTAX WORKED EXAMPLE

Volunteer type and weekly hours

OpenStax gives 839 adult volunteers classified by volunteer type and weekly hours. For hours 1-3, 4-6, and 7-9, the observed rows are community-college students: (111, 96, 48); four-year-college students: (96, 133, 61); nonstudents: (91, 150, 53). For this classroom analysis, assume one SRS of 839 from a volunteer population of at least 8,390, with each person recorded once. At alpha = 0.05, is there convincing evidence of an association between volunteer type and hours volunteered?

1. Use a chi-square test for independence. H_0: volunteer type and weekly-hours category are independent in the volunteer population. H_a: the variables are associated. One sample is classified on two categorical variables.

2. The stated SRS and population-size assumption justify random sampling and the 10% condition. Row totals are 255, 290, 294; column totals are 298, 379, 162; total n = 839.

3. The expected count for community-college students volunteering 1-3 hours is $255(298)/839 \approx 90.5721$. The expected rows are approximately (90.5721, 115.1907, 49.2372), (103.0036, 131.0012, 55.9952), and (104.4243, 132.8081, 56.7676). All exceed 5.

4. Sum $(O-E)^2/E$ over all nine cells to obtain $\chi^2 \approx 12.9909$. The degrees of freedom are $(3-1)(3-1) = 4$; the right-tail p-value is approximately 0.01132.

5. Since $0.01132 < 0.05$, reject H_0. There is convincing evidence of association between volunteer type and weekly-hours category in the population represented by the sample.

Conclusion. The source table supports this population inference only under a defensible random-sampling design. Because volunteer type was not randomly assigned, the test does not show that being a student causes a change in volunteering hours.

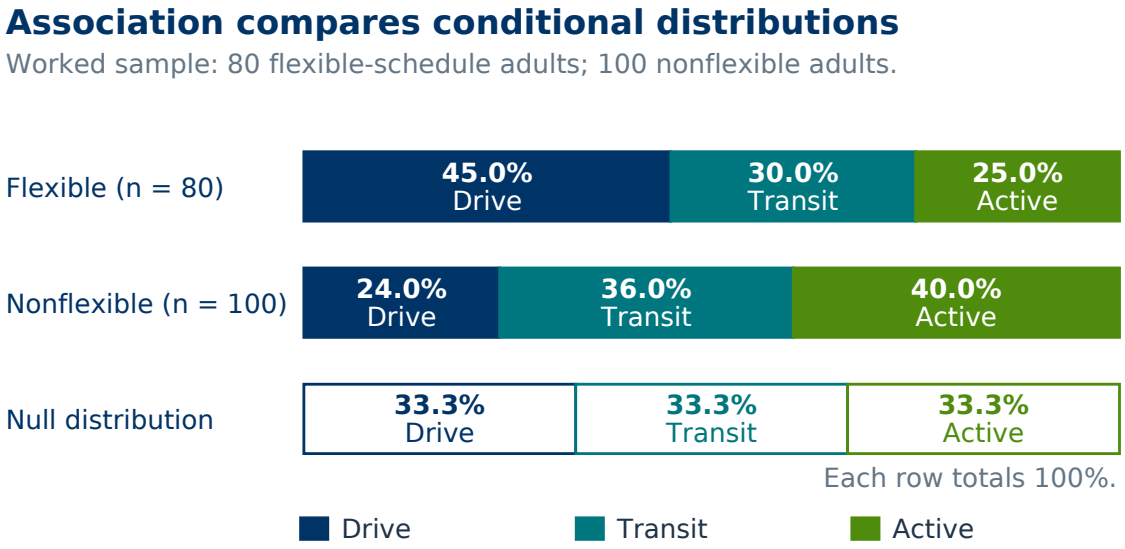
Source: OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 11.3, Example 11.6, Tables 11.15-11.16, printed pp. 575-577 (PDF pp. 587-589). Original nine counts and categories retained. Random selection, population size, and alpha are explicit classroom assumptions because the source does not supply enough design detail.

LOCAL APPLICATION

Work schedule and commute choice

Illustrative study: an SRS of 180 adults from a district of 20,000 records flexible schedule (yes/no) and main commute mode (drive/transit/active). Flexible counts are 36,24,20; nonflexible counts are 24,36,40. Test independence at alpha 0.05.

The null expects the same commute distribution in both groups



Observed percentages use the appropriate row denominators. The common null proportions are one-third in each category because the column totals are all 60. Association does not establish that schedule changes cause commuting changes.

Expected counts preserve different row sizes

Schedule / mode	Observed O	Expected E	O − E
Flexible / drive	36	26.6667	+9.3333
Flexible / transit	24	26.6667	-2.6667
Flexible / active	20	26.6667	-6.6667
Nonflexible / drive	24	33.3333	-9.3333
Nonflexible / transit	36	33.3333	+2.6667
Nonflexible / active	40	33.3333	+6.6667

All expected counts exceed 5. Summing $(O - E)^2/E$ gives 9.36 with 2 degrees of freedom and right-tail p-value 0.009279.

1. H0: flexible-schedule status and commute mode are independent among all district adults. Ha: these variables are associated. A chi-square independence test fits one random sample with two categorical variables.

2. The SRS supplies randomization; 180<=10% of 20,000=2,000. Row totals 80 and 100, column totals 60,60,60, grand total 180. Expected flexible counts are 26.6667 in each cell and nonflexible counts 33.3333 in each cell, all greater than 5.

3. Compute every contribution: flexible row 3.2667,0.2667,1.6667; nonflexible row 2.6133,0.2133,1.3333. Adding them gives chi-square=9.36.

4. Degrees of freedom are (2-1)(3-1)=2. The right-tail chi-square probability is 0.009279. Under independence, a discrepancy at least this large would occur about 0.93% of the time.

5. Since 0.009279<0.05, reject H0. There is convincing evidence of association between schedule flexibility and commute mode among district adults. The sample shows 45% driving among flexible-schedule adults versus 24% among nonflexible adults.

Conclusion. The random sample supports a district-level association statement. Because schedule flexibility was not randomly assigned, the analysis does not show that changing a schedule causes a commute-mode change.

Sources and adaptations

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 11.3, Example 11.6, Tables 11.15-11.16, printed pp. 575-577 (PDF pp. 587-589). Original nine counts and categories retained. Random selection, population size, and alpha are explicit classroom assumptions because the source does not supply enough design detail.

OpenStax, Introductory Statistics 2e, supplied 2026 printing, section 11.3, Try It 11.5, printed p. 575 (PDF p. 587). Original counts and expected-count task retained. A second part distinguishes expected counts from evidence of association.

Definitions and reasoning are adapted from the listed OpenStax sections. This AP edition uses p-hat notation, an equality null, at least 10 observed counts for proportion intervals and at least 10 null expected counts for proportion tests. It retains the supplied Fall 2026 CED condition of expected chi-square counts greater than 5. Worked assumptions are explicit; textbook cases are not reports of current local data.

Unit synthesis

1. An SRS of club members is used to test whether more than half favor a new schedule. The test of H_0: $p = 0.50$ against H_a: $p > 0.50$ gives p-value 0.014, where p is the population support proportion. Interpret this p-value in context.

2. An SRS of 120 of a club's 6,000 members finds that 72 favor a proposed schedule. At the 0.05 significance level, do the data provide convincing evidence that a majority of all club members favor the schedule? Justify your answer with an appropriate test.

3. At a specified alternative, power is 0.91. Find beta and interpret it.

4. Explain when a proportion standard error is pooled and when it is not.

5. A two-by-four table is used for homogeneity. Find df, identify the tail, and state the expected-count requirement used in this course.

6. Explain the difference between homogeneity and independence and one limitation shared by observational versions of both.

Unit 08 | Confidence Intervals for Means

Quantitative responses lead to questions about population means and mean differences. This unit develops t-based confidence intervals from the role of estimated variability, then separates one-sample, paired, and independent-group designs. Students practice checking conditions with evidence, interpreting confidence in context, and judging claims by the entire interval. Investigations show why sample size and pairing affect precision. The local worked examples and investigations use illustrative datasets; credited source examples retain their stated provenance. Technology performs arithmetic, while students remain responsible for the parameter, design, assumptions, and scope of the conclusion. The unit closes by connecting inference choices to the question that motivated data collection.

CED topics: 4.2, 4.3, 4.7, 4.8

IN THIS UNIT

Lesson	Focus
S08.01	Estimating a population mean
S08.02	Confidence, precision, and defensible claims
S08.03	Paired observations: analyze the differences
S08.04	Estimating a difference between independent means

Two trail routes: what defines a pair?

Study design	Observation used in the comparison
Each participant walks both routes	One within-person difference in minutes
Different participants walk each route	One time in minutes per participant

Only the first design creates a Route A and a Route B observation for each participant.

LOCAL QUESTION

How could you estimate mean travel time along a fixed trail segment? Decide whether comparing two routes would call for paired or independent observations.

S08.01 | Estimating a population mean

CED 4.2

- Explain why estimating population variability calls for a t-distribution.
- Verify conditions for a one-sample t-interval.
- Calculate and interpret standard error, critical value, and margin of error.

CONCEPT AND INTERPRETATION

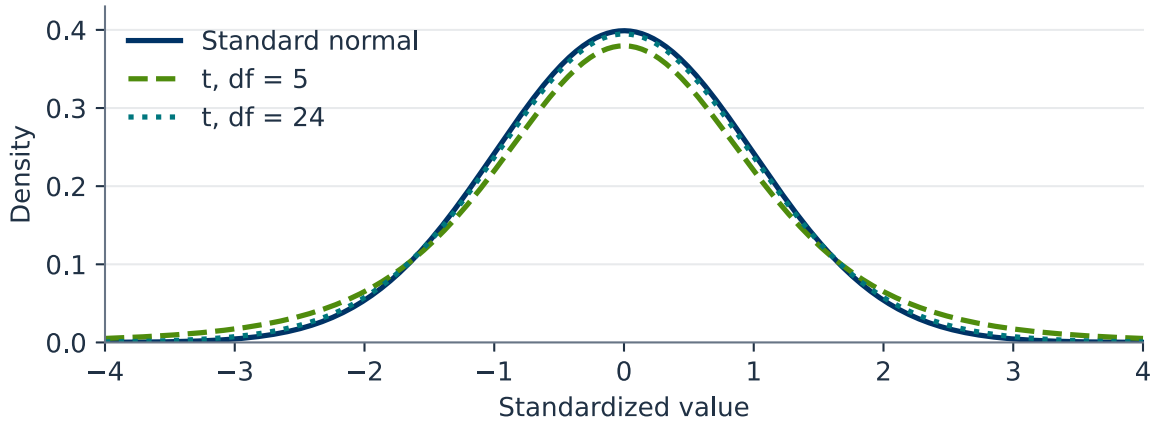
From observations to a population mean

A mean answers a question about a quantitative variable. Let μ be the population mean, $\bar{x}$ the sample mean, σ the population standard deviation, and s the sample standard deviation. The sample mean estimates the population mean. The sample standard deviation describes variation among individual observations. The standard error $SE = s/\sqrt{n}$ estimates the sampling variability of the mean. These quantities have the same measurement units but describe different distributions. Define the population and response variable before selecting a procedure.

Why the t-distribution appears

OpenStax begins with a practical difficulty: the population standard deviation is usually unknown. Substituting the sample standard deviation s for σ makes the denominator vary from sample to sample. For a simple random sample from a normal population, the standardized statistic $t = (\bar{x} - \mu)/(s/\sqrt{n})$ has a Student t distribution with n−1 degrees of freedom. The t distributions are symmetric about zero and have heavier tails than the standard normal distribution. Their exact shape depends on the degrees of freedom. As n increases, s generally estimates σ more precisely and the t distribution approaches the standard normal distribution. Use t whenever s estimates σ, including when n exceeds 30. The heavier tails explain why a small-sample t interval uses a larger multiplier than a z interval at the same confidence level.

Why t has heavier tails



All curves are centered at zero. The t curves account for estimating population variability; as df increases, t approaches the standard normal.

Why there are n−1 degrees of freedom

The sample standard deviation is computed from the deviations $x_i - \bar{x}$. These deviations sum to zero. Once n−1 of them are known, the last is determined by that zero total. Thus only n−1 deviations can vary freely. For 15 observations, use df=14; do not use the sample size as the degrees of freedom. This explanation, developed in OpenStax Section 8.2, connects the t model to the same sample-variability estimate used in the interval.

Conditions provide the justification

The data should come from a random sample or randomized experiment. When sampling without replacement, verify $n \leq 0.10N$ so dependence from the sampling fraction is negligible. For the sample-data condition, an approximately normal population supports the procedure; alternatively, a sample size of at least 30 supports a normal approximation for the sample mean. With fewer than 30 observations, inspect the sample for strong skewness and outliers. Large sample size does not fix selection bias, dependence, or poor measurement. Record actual evidence for each condition rather than writing three names.

Constructing the interval

A one-sample t-interval is $\bar{x} \pm t^* s/\sqrt{n}$. The critical value t^* encloses the chosen central confidence level in a t-distribution with $n-1$ degrees of freedom. For 95% confidence, use cumulative probability 0.975 because the remaining 5% is divided between two tails. The margin of error is $t^* SE$ and the total interval width is twice that margin. Retain unrounded intermediate values. When technology reports an interval, record the procedure, inputs, and degrees of freedom so another reader can reproduce it.

What the interval estimates

An interval for μ estimates one population mean. It does not estimate the middle 95% of individual values, nor does it predict the next observation. Write the conclusion with the confidence level, numerical endpoints, units, and defined population mean. The calculation describes sampling uncertainty under the stated model. A confidence interval cannot account automatically for nonresponse, an unrepresentative sampling frame, or a faulty measuring device. Those problems can shift the estimate even when the calculated interval is very narrow.

KEY IDEA

A t-interval estimates a population mean while accounting for estimation of population variability.

WATCH FOR

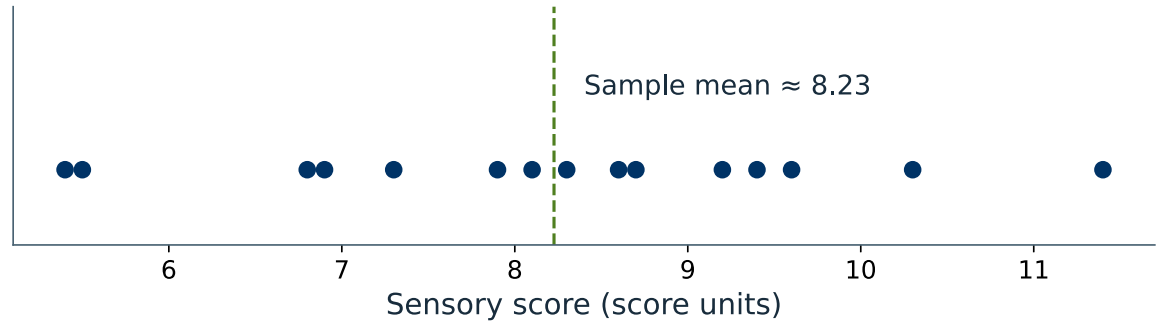
Do not substitute the sample standard deviation for the standard error, or use z merely because n is large.

OPENSTAX WORKED EXAMPLE

Estimating a mean from 15 sensory scores

OpenStax gives these sensory scores for 15 subjects: 8.6, 9.4, 7.9, 6.8, 8.3, 7.3, 9.2, 9.6, 8.7, 11.4, 10.3, 5.4, 8.1, 5.5, 6.9. For this exercise, assume the subjects are an independent random sample from a normal population containing at least 150 subjects. Construct and interpret a 95% confidence interval for the population mean sensory score.

The 15 source observations



One point represents one subject's sensory score. The dashed line marks the sample mean, not the unknown population mean.

1. Define μ as the population mean sensory score for the population represented by the sample. The response is quantitative, and σ is unknown, so use a one-sample t interval.

2. The exercise specifies random selection and a normal population. The sample is no more than 10% of the population. These assumptions justify the calculation; the scores alone do not establish how the subjects were selected.

3. From the raw scores, n=15, $\bar{x} = 8.2267$, and s=1.6722. Then $SE = s/\sqrt{15} = 0.43177$ score unit, and df=14.

4. For 95% confidence, leave 0.025 in each tail. The critical value is $t^* = 2.1448$. The margin of error is 2.1448(0.43177)=0.92605 score unit.

5. The interval is 8.22667 ± 0.92605, or (7.30061,9.15272). We are 95% confident that the population mean sensory score is between 7.30 and 9.15 score units.

Conclusion. This interval estimates the mean sensory score for the stated population. Without a treatment comparison, it does not show that a treatment reduced pain.

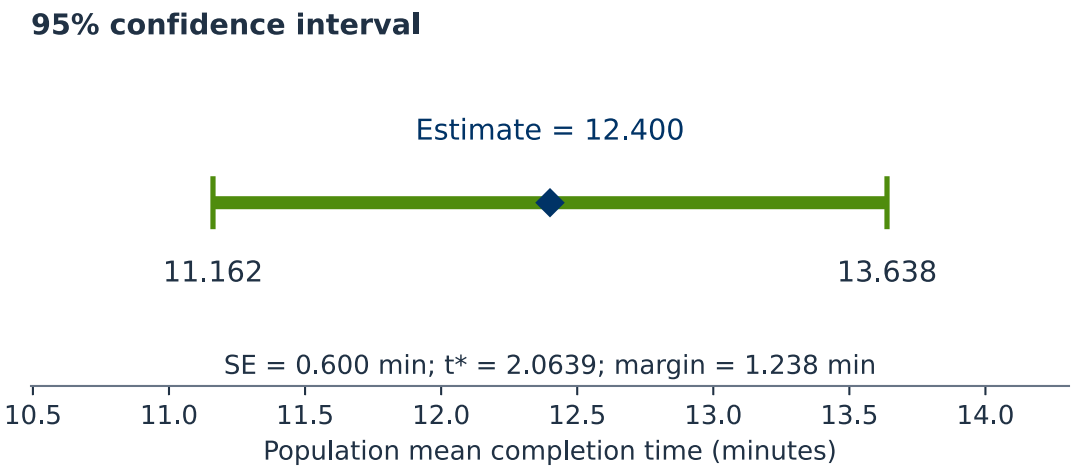
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 8.2, Example 8.8, pp.418-419 (PDF 430-431). Data retained; the explanation and any added assumptions are identified in the lesson source notes.
https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

How long is a routine print job?

Illustrative data: an SRS of 25 jobs from 600 comparable school print jobs has mean completion time 12.4 minutes and sample standard deviation 3.0 minutes. A plot is roughly symmetric with no outliers. Estimate the population mean using 95% confidence.

Anatomy of the print-job interval



Worked data: n = 25, mean = 12.4 minutes, s = 3.0 minutes, df = 24. The green interval estimates a population mean, not a range of individual job times.

1. Classify. Completion time in minutes is quantitative. The school took a simple random sample of jobs without assigning a treatment.

2. Choose. Let μ be the mean completion time of all 600 comparable print jobs. Use a one-sample t interval at 95% confidence because the population standard deviation is unknown.

3. Check. The 25 sampled jobs make up less than 10% of the 600 jobs. The stated plot of times shows no strong skewness or outliers in this small random sample.

4. Calculate. With df = 24, $t^* = 2.0639$ and $SE = 3/\sqrt{25} = 0.60$ minute. The interval is $12.4 \pm 2.0639(0.60)$, or (11.162, 13.638) minutes.

5. Conclude. We are 95% confident that the mean completion time of all 600 comparable print jobs is between 11.16 and 13.64 minutes.

Conclusion. The estimated margin of error is about 1.24 minutes for the batch mean. Individual jobs can take much more or less time.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 8.2, pp.416-419 (PDF pp.428-431), Example 8.8 and Try It 8.8. The t-distribution development and sensory-score example follow the source. Random sampling and the 10% condition are made explicit for AP use. The source does not establish that acupuncture or hypnotherapy caused a change; these examples estimate means only. https://openstax.org/details/books/introductory-statistics-2e

S08.02 | Confidence, precision, and defensible claims

CED 4.3

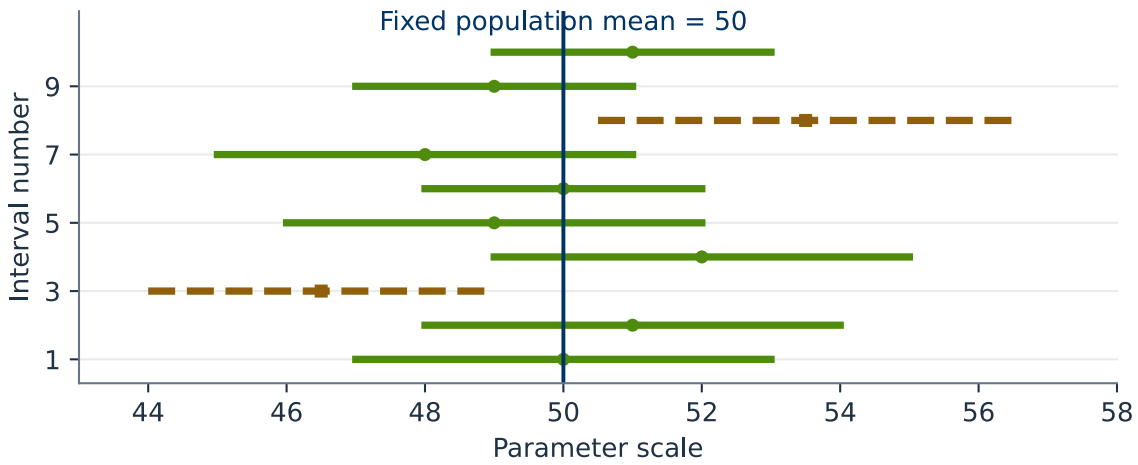
- Interpret confidence level as long-run procedure performance.
- Evaluate a numerical mean claim using an interval.
- Explain effects of confidence and sample size on precision.

CONCEPT AND INTERPRETATION

An interval and its confidence level answer different questions

A computed interval has fixed endpoints. The population mean is also fixed, although unknown. The interval either contains that mean or it does not. The confidence level describes the long-run success rate of the method: if we repeatedly sampled in the same way and constructed the same kind of interval, approximately the stated percentage would capture the population mean. A statement such as “95% of sample means are in this interval” changes the subject and is not the meaning of 95% confidence.

The parameter stays fixed; intervals vary



These ten fictional teaching intervals reproduce the investigation: eight capture 50 and two miss. The displayed 80% is not an empirical estimate of a 95% procedure’s coverage.

Writing an interpretation that can be assessed

Identify the population, quantitative response, and unit. Then report that you are C% confident that the interval from the lower to upper endpoint contains the population mean. Avoid language claiming that a random individual falls in the interval with probability C%. Avoid claiming that C% of population observations lie between the endpoints. The point estimate is useful, but the interval also shows which population values remain plausible under the sampling model. Precision in wording is part of the statistical result, rather than a decoration added after calculation.

Using an interval to assess a claim

Compare the proposed population mean to the entire interval. A value outside a two-sided confidence interval is inconsistent with that interval procedure at the corresponding level. A value inside is compatible with the data; it has not been proved true. For an ordinary one-sample t procedure, a two-sided 95% interval and the corresponding two-sided test at significance 0.05 give matching reject/fail-to-reject decisions, subject to rounding. The interval also indicates direction and plausible size of the discrepancy, which a yes/no decision alone cannot show.

Confidence has a precision cost

Holding the observed data fixed, increasing confidence increases the critical value t^*, the margin of error, and the total width. Decreasing confidence narrows the interval but lowers long-run coverage. Holding variability and confidence approximately fixed, increasing n reduces $SE = s/\sqrt{n}$. Four times the sample size makes this standard error half as large. The margin of error is approximately halved as well; it is not exactly proportional because t critical values change with degrees of freedom and a new sample can have a different standard deviation.

Precision is useful only with a sound design

A narrow interval from a biased sample may miss the population mean consistently. Before increasing sample size, examine the sampling frame, measurement definition, nonresponse, and possible dependence. A practical claim also needs a scale. A difference of one tenth of a minute may be statistically detectable yet unimportant for a school schedule. Conversely, a wide interval spanning consequential positive and negative effects leaves an important decision unresolved. State what is compatible with the data, what is excluded, and what the study design permits you to generalize.

KEY IDEA

Confidence describes the procedure; an interval describes plausible values for the population mean.

WATCH FOR

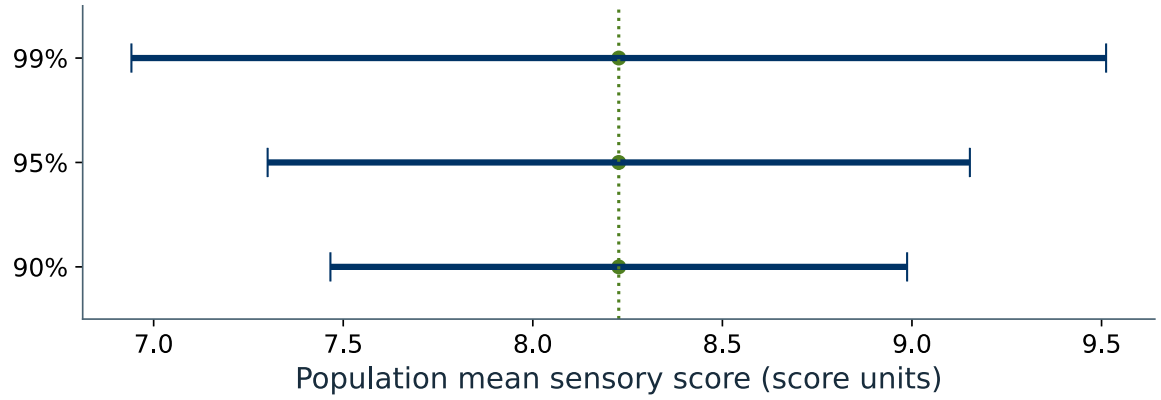
An interval containing a claimed value does not establish that the claim is true.

OPENSTAX WORKED EXAMPLE

The same scores at three confidence levels

Use the 15 sensory scores from the preceding source example. Their mean is 8.22667 and s=1.6722. Construct 90%,95%, and 99% intervals for the population mean. Explain which parts of the calculation change.

Same sample, different confidence levels



Each interval uses the same 15 scores and the same point estimate. Higher confidence widens the interval; the dots mark the sample mean.

1. The point estimate 8.22667, sample size 15, standard error 0.431770, and df 14 remain fixed because all three calculations use the same observations.

2. The central confidence areas 0.90,0.95,0.99 require upper-tail areas 0.05,0.025,0.005. The corresponding critical values are 1.76131,2.14479,2.97684.

3. The intervals are 90%: (7.46619,8.98715);95%: (7.30061,9.15272);99%: (6.94136,9.51198), in sensory-score units.

4. Increasing confidence extends each endpoint farther from the same estimate. It does not change the population mean or the observed scores.

5. The 95% confidence level means that about 95% of intervals made by this method from repeated samples of 15 under the same conditions would contain the fixed population mean. It does not mean that 95% of individual scores lie between 7.30 and 9.15.

Conclusion. Greater confidence costs precision when the data are held fixed. Report the confidence level together with the population mean, endpoints, and units.

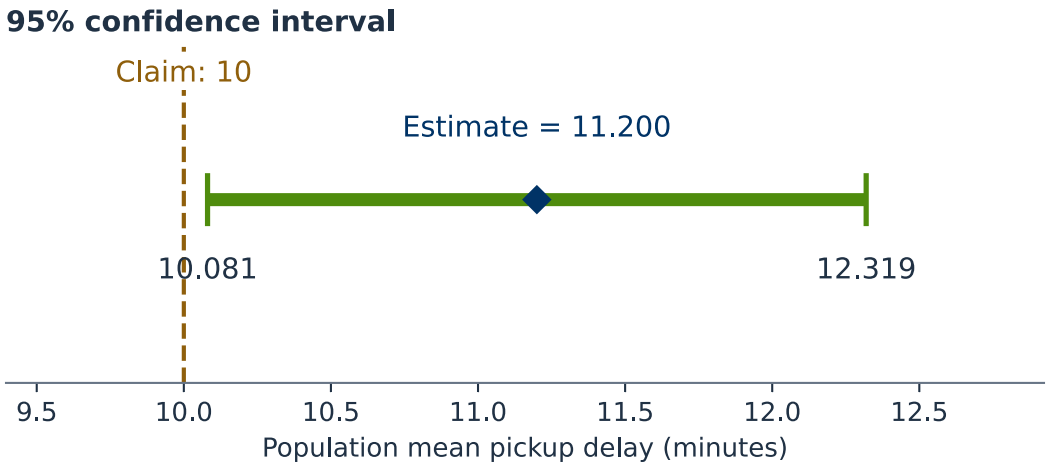
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 8.1-8.2; extension of Example 8.8, pp.418-419 (PDF 430-431). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Does a pickup service average 10 minutes?

Illustrative study: an SRS of 40 pickups from 2,000 comparable pickups gives mean delay 11.2 minutes and s = 3.5 minutes. Consider a 95% interval and the claim that the population mean is 10 minutes.

Assessing the 10-minute claim



Worked data: n = 40, mean = 11.2 minutes, s = 3.5 minutes. The complete 95% interval lies above the claimed mean of 10 minutes.

08

1. Let μ be the population mean delay for the 2,000 comparable pickups. Use a one-sample t-interval because delay is quantitative and σ is unknown.

2. Random selection is stated. The sampling fraction is 40/2000=0.02, below 10%. With n=40, the sample-data condition is supported.

3. Using df=39, $t^* = 2.0227$; SE = 0.5534; interval (10.081, 12.319), in minutes.

4. The entire interval is above 10 minutes. Thus 10 is not a plausible population mean at this confidence level under the stated procedure. The interval suggests an average delay larger than 10.

5. In repeated samples of 40 drawn in the same way, about 95% of such intervals would capture μ. This does not mean 95% of the individual pickups have delays inside this interval.

Conclusion. The data provide evidence against a mean delay of 10 minutes. The size of the plausible excess, not just its sign, should inform any operational response.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 8.1-8.2, pp.409-419 (PDF 421-431), Example 8.8. Confidence and width are developed from the source interval construction. The worked comparison extends Example 8.8 to 90% and 99% using its unchanged data; AP wording separates a fixed interval from repeated-sampling coverage.
https://openstax.org/details/books/introductory-statistics-2e

S08.03 | Paired observations: analyze the differences

CED 4.2, 4.3

- Identify paired data from the study design.
- Define a population mean difference with a subtraction order.
- Check and calculate a one-sample t-interval on differences.

CONCEPT AND INTERPRETATION

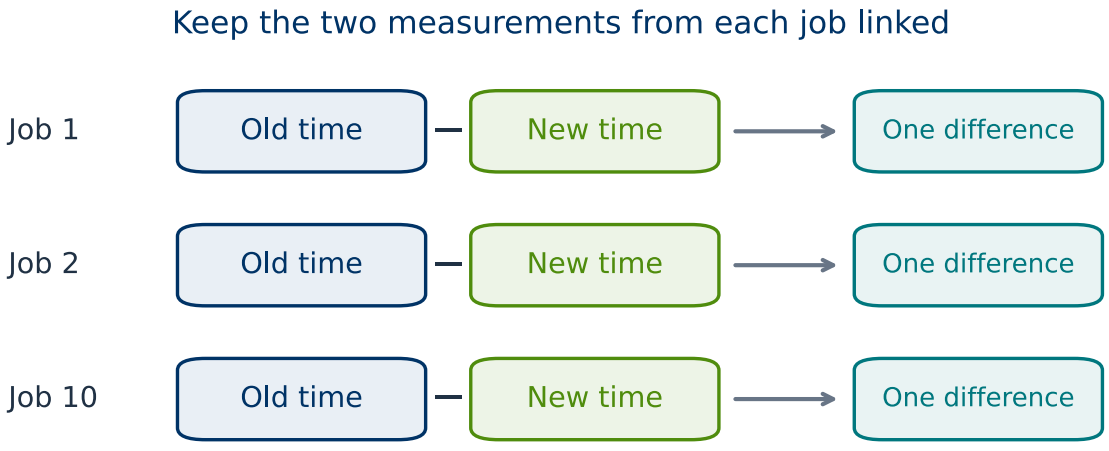
Pairing is a feature of the design

Two measurements are paired when a meaningful link joins them before analysis. They may come from the same person under two conditions, the same site at two times, or two individuals deliberately matched on a relevant characteristic. Two groups are not paired merely because they have the same number of observations. The connection must carry information. Pairing can remove variation between people or objects from the comparison, making within-pair differences more informative than two unrelated sets of measurements.

Create one quantitative variable

Choose a subtraction order and keep it throughout the analysis. If $d = \text{old} - \text{new}$, a positive difference indicates a smaller new measurement. Let μ_d denote the population mean of these differences. Calculate one difference for each pair, then compute $\bar{d}$ and s_d from that list. The sample size n is the number of pairs, not the number of individual measurements. A sample of 20 people measured twice gives 20 differences and 19 degrees of freedom, rather than 40 observations and 39 degrees of freedom.

Paired data become one sample of differences



One job contributes one old-minus-new difference. Ten jobs give ten differences and nine degrees of freedom; the measurements are not treated as twenty independent observations.

Conditions concern the differences

The design should involve random selection of pairs or random assignment of treatments or treatment order in an appropriate experiment. When pairs are sampled without replacement, compare the number of pairs with the population of eligible pairs for the 10% condition. With fewer than 30 pairs, inspect the distribution of differences for strong skewness and outliers. The separate distributions of the first and second measurements do not answer that question. Even two skewed measurement distributions can have well-behaved differences if the pairing removes a shared source of variation.

Constructing and interpreting the interval

Treat the difference list as one sample. The interval is $\bar{d} \pm t^* s_d/\sqrt{n}$, with $df = n - 1$. Its endpoints estimate the population mean difference in the stated order. If the entire interval is positive, the first measurement tends to exceed the second on average. If the interval contains zero, the data do not establish a nonzero population mean difference at that confidence level. Reverse the subtraction order and both the estimate and interval change sign: an interval (a, b) becomes (−b,−a).

Pairing does not guarantee a causal answer

A before-and-after study can show change without identifying its cause. Practice, time, weather, or another concurrent change may explain the difference. Randomized treatment order can help control order effects when each subject experiences both treatments. A randomized comparison supports a stronger causal interpretation than an uncontrolled pre/post observation. Generalization depends separately on selection of subjects. State these limits even when the paired t calculation is correct. A small p-value or narrow interval cannot supply design features that the study never had.

From two columns to one sample

OpenStax states the central rule for matched pairs: the differences are the data. Suppose the same student throws with each hand. Subtract that student's weaker-hand distance from that student's dominant-hand distance before computing any summary. Repeat for every student. The result is one list of differences, with one observation per student. Compute the sample standard deviation of this list directly. Subtracting the two original standard deviations does not give the standard deviation of differences, because the paired measurements can be associated.

KEY IDEA

Paired inference is one-sample inference on a defined list of differences.

WATCH FOR

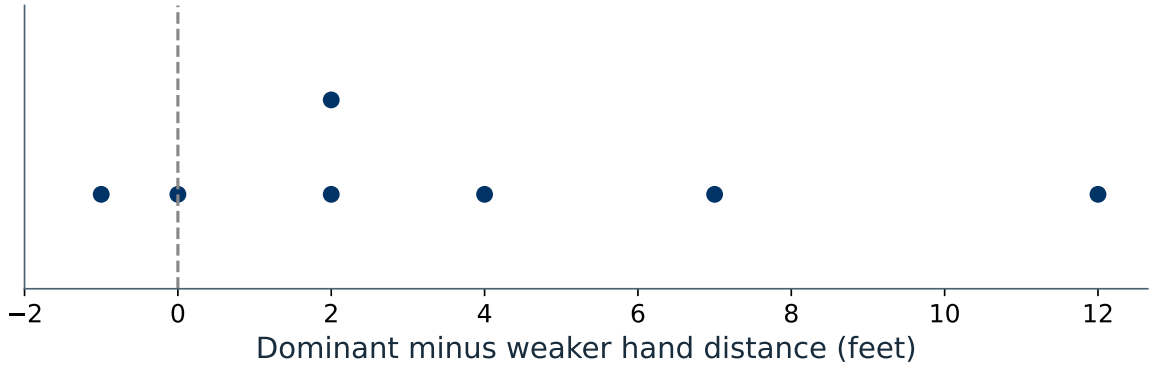
Check the shape of the differences; do not count two measurements per pair as two independent observations.

OPENSTAX WORKED EXAMPLE

Estimating the mean difference between two hands

OpenStax reports dominant-hand and weaker-hand shot-put distances for seven students. Dominant distances are 30,26,34,17,19,26,20 feet; corresponding weaker-hand distances are 28,14,27,18,17,26,16 feet. For this interval exercise, assume an SRS of seven from at least 70 comparable students and a normal population of dominant-minus-weaker differences. Construct a 95% confidence interval for the population mean difference.

Seven paired differences



One dot represents one student's dominant-minus-weaker shot-put distance. The analysis uses these seven differences, not fourteen independent throws.

1. Each student supplies two linked distances. Define d=dominant distance−weaker distance in feet. The seven differences are 2,12,7,−1,2,0,4.

2. Let μ_d be the population mean dominant-minus-weaker shot-put distance. Use a one-sample t interval on the seven differences. Random selection, the 10% check, and the stated normal difference distribution support the method.

3. The mean difference is 3.71429 feet and the sample standard deviation of differences is 4.49868 feet. Thus SE=4.49868/√7=1.70034 feet and df=6.

4. The 95% critical value is 2.44691. The interval is 3.71429 ±2.44691(1.70034), or(−0.44630,7.87487) feet.

5. We are 95% confident that the population mean dominant-minus-weaker distance is between −0.45 and 7.87 feet. Because zero is included, the data do not establish a nonzero population mean difference at the corresponding two-sided 5% level.

Conclusion. An average advantage in the sample does not guarantee convincing evidence of a population advantage. The interval includes zero and values on both sides.

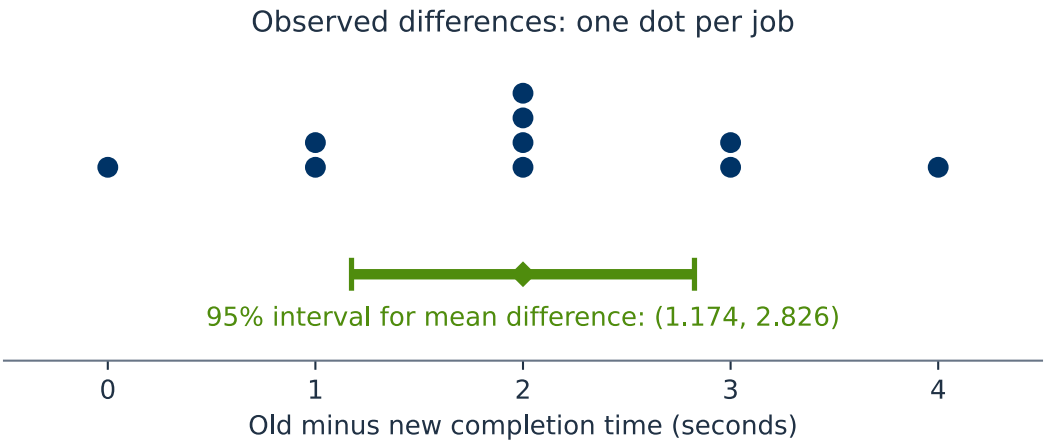
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.4, Example 10.13, pp.531-532 (PDF 543-544); confidence-interval extension. Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Two formatting methods on the same jobs

Illustrative crossover study: from 500 comparable print jobs, select an SRS of ten. Run each job using both formatting methods, randomizing which runs first. The old-minus-new time differences in seconds are 1, 3, 2, 4, 0, 2, 3, 1, 2, 2. Estimate the mean reduction for the 500 jobs with 95% confidence.

Individual differences and their mean interval



Worked differences: 1, 3, 2, 4, 0, 2, 3, 1, 2, 2 seconds. The lower green segment estimates the population mean difference; it does not contain all individual differences.

1. Let μ_d be the mean old-minus-new completion time for the 500 jobs. Each sampled job supplies one difference, so use a one-sample t-interval for the mean difference.

2. The SRS supports inference to the 500 jobs; $10 \le 0.10(500)$, so treating sampled jobs as approximately independent is reasonable. Randomizing method order within each job helps balance any advantage of running first.

3. The difference list is roughly symmetric with no strong outlier. n=10, mean difference=2.000 seconds, and sample standard deviation=1.1547 seconds.

4. Use df=9. $t^* = 2.2622$; SE = 0.3651; interval (1.174, 2.826), in seconds.

5. Both endpoints are positive. The interval estimates average time saved across the 500 jobs; an individual job need not save this much time.

Conclusion. We are 95% confident that the new formatting method saves between 1.17 and 2.83 seconds per job on average across the 500 jobs.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.4, pp.527-532 (PDF 539-544), Example 10.13, and Section 8.2. The differences-as-data explanation follows 10.4. The source shot-put data are retained and extended to an interval. The population normality assumption is retained; random sampling and population size are explicitly added for the classroom interval exercise. https://openstax.org/details/books/introductory-statistics-2e

S08.04 | Estimating a difference between independent means

CED 4.7, 4.8

- Distinguish independent groups from matched pairs.
- Construct an unpooled two-sample t-interval.
- Interpret the sign and size of a difference in population means.

CONCEPT AND INTERPRETATION

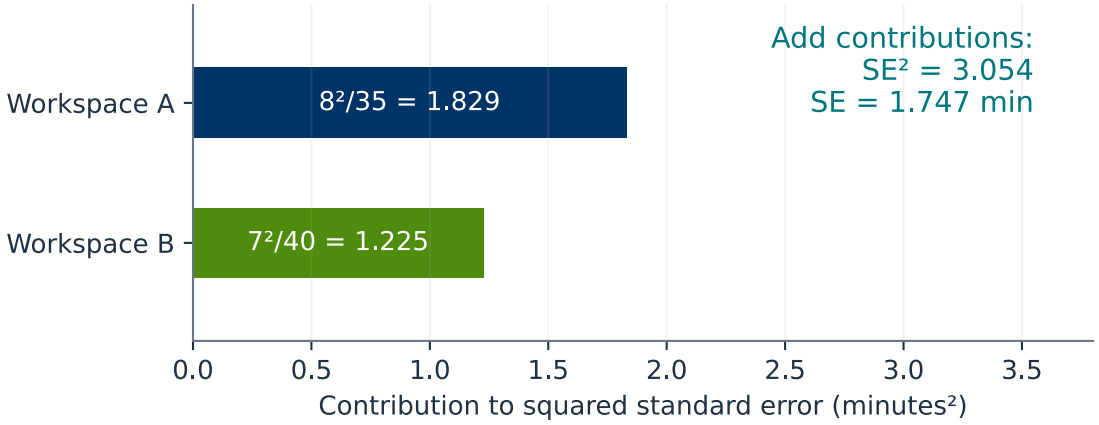
Two independent groups answer a different question

When observations in one group are not meaningfully linked to observations in the other, compare two independent population means. Define μ_1 and μ_2 with the populations, response, and units. The parameter is $\mu_1 - \mu_2$, and its point estimate is $\bar{x}_1 - \bar{x}_2$. Keep the subtraction order visible throughout the work. The sample sizes may differ; equal sample sizes do not imply pairing. If the same subjects supply both responses, return to a paired analysis instead.

Variability comes from both samples

For independent groups, estimate the standard error with $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$. The estimated variances add because both sample means contribute uncertainty. Do not subtract standard deviations even though the parameter is a difference. Do not pool the sample variances by default. The unpooled procedure allows the two population variances to differ. Technology calculates an appropriate Welch degrees of freedom, which need not be an integer. Report it as part of the output, rather than replacing it automatically with the total sample size minus two. The Welch degrees of freedom lie between the smaller of $n_1 - 1$ and $n_2 - 1$ and the upper bound $n_1 + n_2 - 2$.

Both independent groups contribute uncertainty



Worked sample variances are divided by their own sample sizes, then added before taking the square root. A difference of means does not use a difference of standard deviations.

Conditions must address both groups

The data should come from two independent random samples or a randomized experiment with independent groups. For sampling without replacement, check the 10% condition against each sampled population. A randomized treatment comparison does not require a population sampling-fraction check solely because participants were assigned to groups. Check the distributional condition for each group separately: a sufficiently large sample supports the sampling-mean approximation; for a group with fewer than 30 observations, use information that its population is approximately normal or inspect its data for strong skewness and outliers. Both groups must be justified. A large second sample cannot repair a very small, strongly skewed first sample.

Construct and read the difference interval

The two-sample t-interval is $(\bar{x}_1 - \bar{x}_2) \pm t^* SE$. Use the critical value corresponding to the desired confidence level and the technology-reported degrees of freedom. An entirely positive interval supports a larger first population mean; an entirely negative interval supports a smaller first population mean. An interval containing zero does not establish a difference at that confidence level. Explain the magnitude in response units, rather than describing only a plus or minus sign. Reversing the group order negates and reverses both endpoints.

Avoid two common shortcuts

Assess the difference using an interval for the difference itself. Visual overlap or nonoverlap of two separate confidence intervals is not an exact substitute for the appropriate difference procedure. Also distinguish a mean difference from differences among individual people or objects. An average advantage does not say that every member of one group outperforms every member of the other. Finally, match the scope of the conclusion to the design: random sampling supports generalization, whereas randomized treatment assignment supports a causal comparison under the experimental conditions.

KEY IDEA

For independent means, uncertainty from both groups enters an unpooled standard error.

WATCH FOR

A difference in means does not use a difference of standard deviations.

OPENSTAX WORKED EXAMPLE

Estimating a battery-life difference

Adapt the OpenStax A/B summaries to two battery brands. Independent SRSs from populations of at least 500 batteries per brand give A:n=25, mean=5.0 hours, s=1.0 hour; B:n=16, mean=4.7 hours, s=1.2 hours. Assume both population lifetime distributions are approximately normal. Construct and interpret a 95% confidence interval for $\mu_A - \mu_B$.

1. Define μ_A and μ_B as the population mean lifetimes for the two brands. The samples are independent, so use an unpooled two-sample t interval, with A−B kept throughout.

2. Both groups are random samples. The sampling fractions are no more than 25/500 and 16/500, both below 10%. The stated normal populations support the small-sample t approximation.

3. The estimated difference is 5.0−4.7=0.3 hour. The estimated variance contributions are $1.0^2/25=0.04$ and $1.2^2/16=0.09$, so SE=$\sqrt{0.13}$=0.360555 hour.

4. Technology gives Welch df≈27.8571 and $t^*\approx 2.04888$. The interval is 0.3±0.738734, or(−0.438734,1.038734) hours.

5. We are 95% confident that brand A's population mean lifetime is between 0.44 hour shorter and 1.04 hours longer than brand B's. Zero is included, so this interval does not establish a difference in population means.

Conclusion. The two samples suggest an A-minus-B difference of 0.3 hour, but the uncertainty includes both an advantage and a disadvantage for A.

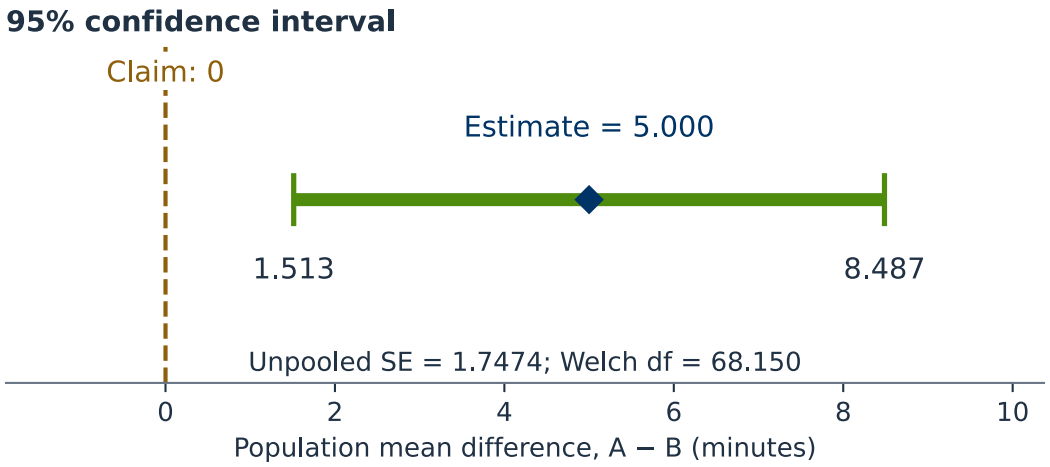
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.1, Try It 10.1, p.514 (PDF 526); context and interval task adapted. Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Comparing two classroom workspaces

Illustrative independent SRSs of comparable room-use periods are taken from two large collections, each exceeding 1,000 periods. Workspace A: n=35, mean=42 minutes, s=8. Workspace B: n=40, mean=37 minutes, s=7. Estimate $\mu_A - \mu_B$ with 95% confidence.

The independent-means interval



Worked samples: A has n = 35, mean 42, s = 8; B has n = 40, mean 37, s = 7. The interval is entirely positive in the A-minus-B order.

08

1. Define μ_A and μ_B as mean task completion times for the two populations of comparable room-use periods. The groups are independent; use an unpooled two-sample t-interval.

2. Both samples are random. Each is at most 10% of its population. Both n values are at least 30, so the sample-data condition is supported.

3. The point estimate is 42−37=5 minutes. SE=$\sqrt{(8^2/35+7^2/40)}$=1.7474 minutes. Welch degrees of freedom is approximately 68.150.

4. Using the central 95% t critical value, SE = 1.7474; df = 68.150; interval (1.513, 8.487), in minutes.

5. The endpoints are positive, supporting a larger population mean completion time in A. Because this is a comparison of random samples rather than random assignment of workspace, other differences between settings could explain the association.

Conclusion. We are 95% confident that mean completion time in A exceeds that in B by about 1.51 to 8.49 minutes for the stated populations.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.1, pp.512-516 (PDF 524-528), Try It 10.1. The unpooled standard error and independent-group distinction follow 10.1. The source A/B summary values are retained, recontextualized as battery lifetimes in hours, and used for interval estimation. Independent SRSs, population sizes, and normality are explicit. The source's small-n and normal-approximation shortcuts are not adopted; AP condition checks and the t method govern.
https://openstax.org/details/books/introductory-statistics-2e

Unit synthesis

1. Explain the roles of s and $s/\sqrt{n}$.

2. An SRS of 25 has mean 30 and s = 10. Show the standard error and find a 95% mean interval when the plot is suitable.

3. Twenty subjects each receive both conditions. What is the analysis unit?

4. An interval for $\mu_A - \mu_B$ is $(-4,1)$. What claim is unsupported?

5. How does quadrupling n affect the margin of error approximately?

6. Why can a narrow interval from convenience data be misleading?

Unit 09 | Hypothesis Tests for Means

This unit uses quantitative data to evaluate claims about population means. Students apply one-sample, paired, and independent two-sample t tests, with emphasis on selecting the method from the study design. Every test begins with a defined parameter and a question-based alternative, then checks conditions before interpreting a p-value. The final lesson compares inference families across the course and connects matching t tests with confidence intervals. Students distinguish statistical evidence from practical importance and separate mean change from causal explanation. Original illustrative investigations develop these ideas before students complete independent contextual responses and mixed homework.

CED topics: 4.4, 4.5, 4.9, 4.10

IN THIS UNIT

Lesson	Focus
S09.01	Testing a claim about one population mean
S09.02	Testing a population mean difference from paired data
S09.03	Testing a difference between independent means
S09.04	Choosing inference and linking tests with intervals

Two measurements on each pear

Pear ID	Before storage (g)	After storage (g)
01	180	172
02	156	150
03	202	193

Illustrative records. Subtract in the same direction within each pear.

LOCAL QUESTION

Would measuring each pear before and after storage create paired data? Define the difference and the population before choosing a test.

S09.01 | Testing a claim about one population mean

CED 4.4, 4.5

- State hypotheses about a defined population mean.
- Calculate and interpret a one-sample t test and p-value.
- Make a contextual decision with an explicit significance comparison.

CONCEPT AND INTERPRETATION

Start with the claim and parameter

A significance test asks whether data are sufficiently inconsistent with a specified population claim. Let μ be the population mean of the quantitative response. State $H_0 : \mu = \mu_0$, where μ_0 is the claimed value. Choose $H_a : \mu > \mu_0$, $H_a : \mu < \mu_0$, or $H_a : \mu \neq \mu_0$ according to the investigative question before examining the sample result. Hypotheses concern a population parameter, not the observed sample mean. The sample mean is already known after data collection; it is evidence used to assess the claim.

Check whether this procedure fits the design

Use a one-sample t test when the response is quantitative, one population mean is targeted, and the population standard deviation is unknown. Random sampling or randomized experimentation must support the data collection. For sampling without replacement, check $n \leq 0.10N$. For the sample-data condition, use an approximately normal population, a sufficiently large sample, or, for a small sample, a distribution without strong skewness and outliers. Write the actual evidence. A large n is not a substitute for randomization or independence, and the 10% condition does not demonstrate a normal sampling distribution.

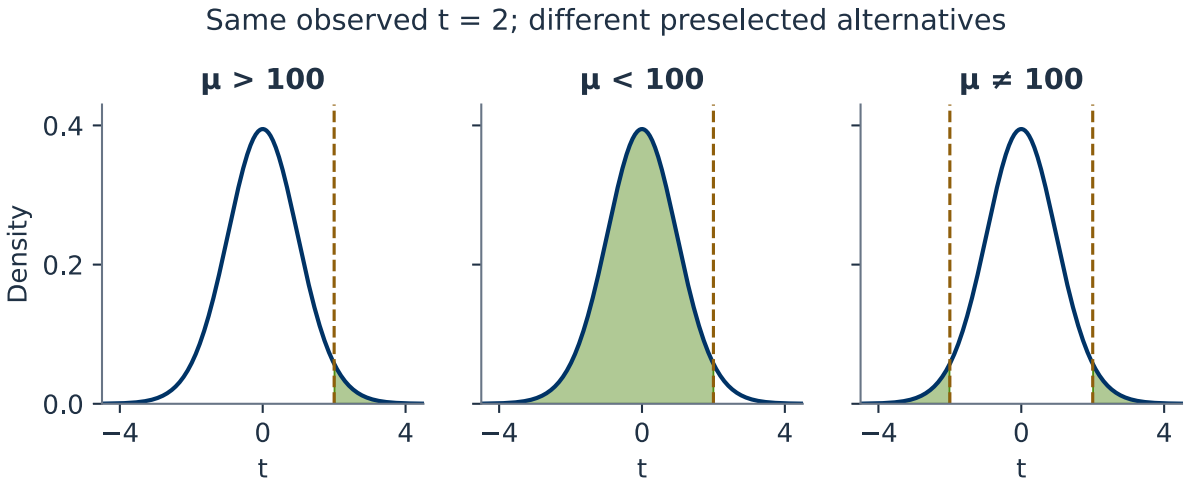
Measure disagreement in standard-error units

The test statistic is $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, with $df = n - 1$. The numerator is observed minus hypothesized mean. The denominator estimates how much sample means vary under repeated sampling. A positive statistic means the sample mean exceeds the null mean; a negative statistic means it falls below. The magnitude measures the discrepancy in estimated standard errors. Preserve the sign until selecting the appropriate tail. Technology should receive the sample information, null value, and alternative direction, rather than an unexplained request to test a list.

Interpret the p-value conditionally

The p-value is the probability, assuming the null hypothesis and procedure conditions, of a test statistic at least as extreme in the direction of the alternative as the observed statistic. A greater-than alternative uses the upper tail, a less-than alternative uses the lower tail, and a two-sided alternative uses both tails beyond the observed magnitude. A p-value is not the probability that the null hypothesis is true. It also is not the probability that the result was caused by chance. State the null assumption and the repeated-sample outcome in context.

The alternative chooses the tail



All panels use t with df = 24 and observed t = 2. A different alternative changes which results count as equally or more extreme.

09

Decide and answer the question

Compare the p-value explicitly with the predetermined significance level α. If p-value is at most α, reject the null; otherwise fail to reject it. Then state whether there is convincing evidence for the alternative claim about the defined population mean. Failure to reject is insufficient evidence, not proof of equality. A statistically detectable departure may still be too small to matter practically. A numerical test cannot repair biased recruitment or justify causation from an observational comparison. Communicate those limits together with the result.

KEY IDEA

A mean test compares an observed mean with a null value in estimated standard-error units.

WATCH FOR

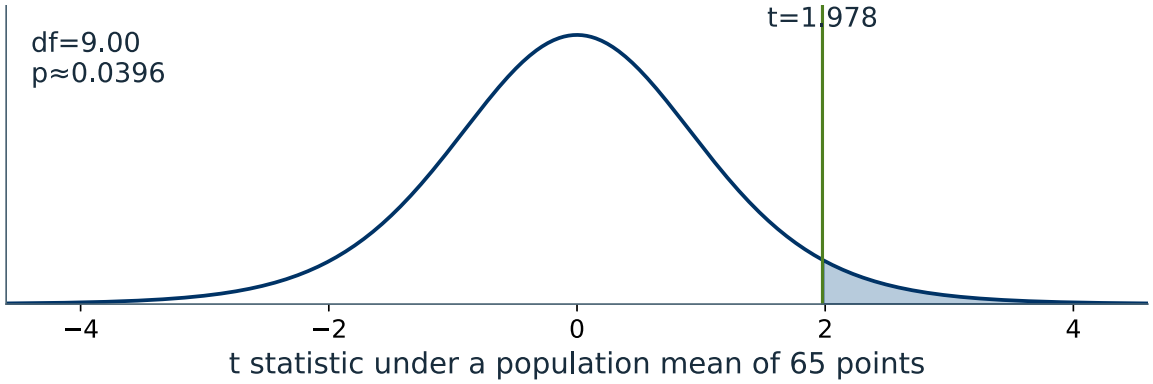
Choose the alternative from the question before examining which direction the data favor.

OPENSTAX WORKED EXAMPLE

Is the mean first-test score greater than 65?

In OpenStax Example 9.16, ten statistics test scores are 65,65,70,67,66,63,63,68,72,71 points. For this exercise, assume these are an SRS from at least 100 students taking the same test and that the population score distribution is approximately normal. At α=0.05, do the data provide convincing evidence that the population mean score exceeds 65 points?

Locate the observed statistic in the null model



The shaded area is the p-value in the direction specified by the alternative. The distribution assumes the null hypothesis and the stated conditions.

1. Let μ be the mean score of all students in the stated test population. The hypotheses are H_0:μ=65 points and H_a:μ>65 points. The greater-than alternative is chosen from the instructor's question before using the sample results.

2. Use a one-sample t test. The stated SRS, sample no larger than 10% of the population, and normal population support this small-sample method. The population standard deviation is unknown.

3. The scores give $\bar{x}=67$ points, s=3.19722 points, and SE=3.19722/√10=1.01105 points. Then t=(67−65)/1.01105=1.97814, with df=9.

4. The upper-tail probability is $P(T_9 \geq 1.97814)=0.039648$. If the population mean were 65 points and the conditions held, a t statistic at least this large would occur in about 3.96% of repeated samples.

5. Since 0.039648<0.05, reject H_0. There is convincing evidence that the population mean first-test score exceeds 65 points for the stated population.

Conclusion. The conclusion concerns a population mean. It does not claim that every student scores above 65 or that the null hypothesis has a 3.96% probability of being true.

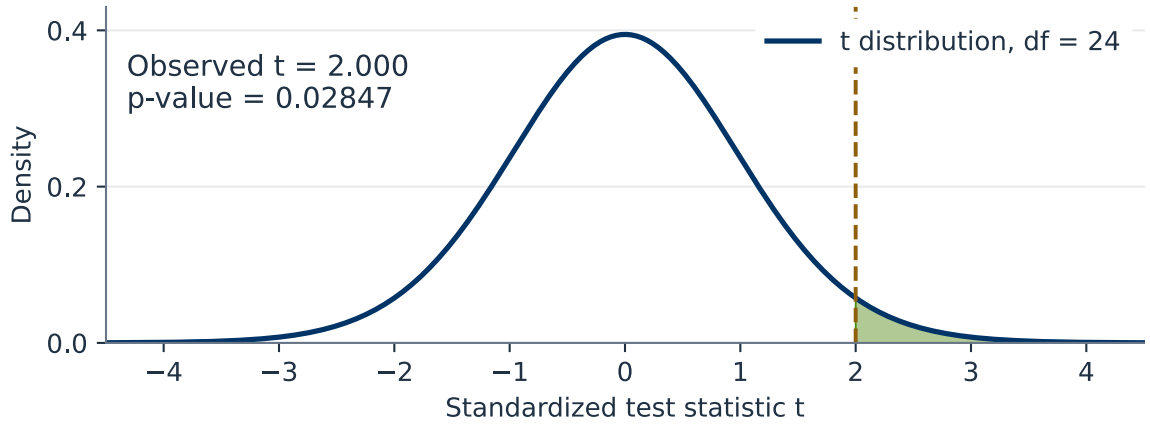
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 9.5, Example 9.16, pp.475-476 (PDF 487-488). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Do packets average more than the labeled mass?

Illustrative quality-control study: an SRS of 25 packets from a batch of 2,000 has mean mass 104 grams and s=10 grams. A plot is roughly symmetric with no outliers. Test whether the population mean exceeds 100 grams at significance 0.05.

Upper-tail evidence for packet mass



Worked data: n = 25, sample mean 104 g, s = 10 g; H_0: $\mu = 100$ g and H_a: $\mu > 100$ g. The excess is 4 g and SE = 2 g, so observed t = 2.

09

1. Classify. Packet mass is quantitative. The team drew an SRS from one production batch.

2. Choose. Let μ be the mean mass of the 2,000 packets in the batch. Use a one-sample t test of H_0: $\mu = 100$ g against H_a: $\mu > 100$ g at $\alpha = 0.05$. A high sample mean favors H_a.

3. Check. The SRS contains 25 of 2,000 packets, well below 10%. The mass plot has no strong skewness or outliers, supporting use of the t model for this small sample.

4. Calculate. $SE = 10/\sqrt{25} = 2$ g. The statistic is $t = (104 - 100)/2 = 2.00$ with 24 degrees of freedom. The null model gives an upper-tail p-value of 0.02847.

5. Conclude. Since 0.02847 is below 0.05, reject H_0. The sample gives convincing evidence that the batch's mean packet mass exceeds the 100 g label.

Conclusion. The evidence supports a population mean above the label value at the 5% significance level. The estimated excess of 4 grams should also be evaluated for practical importance.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 9.1,9.3-9.5, pp.457-476 (PDF 469-488), Example 9.16. The complete test uses the source's ten test scores and 65-point null mean. The sampling frame is specified for AP conditions, and the p-value is interpreted through the standardized t statistic, avoiding an unconditional probability claim about the sample mean with an estimated denominator. https://openstax.org/details/books/introductory-statistics-2e

S09.02 | Testing a population mean difference from paired data

CED 4.4, 4.5

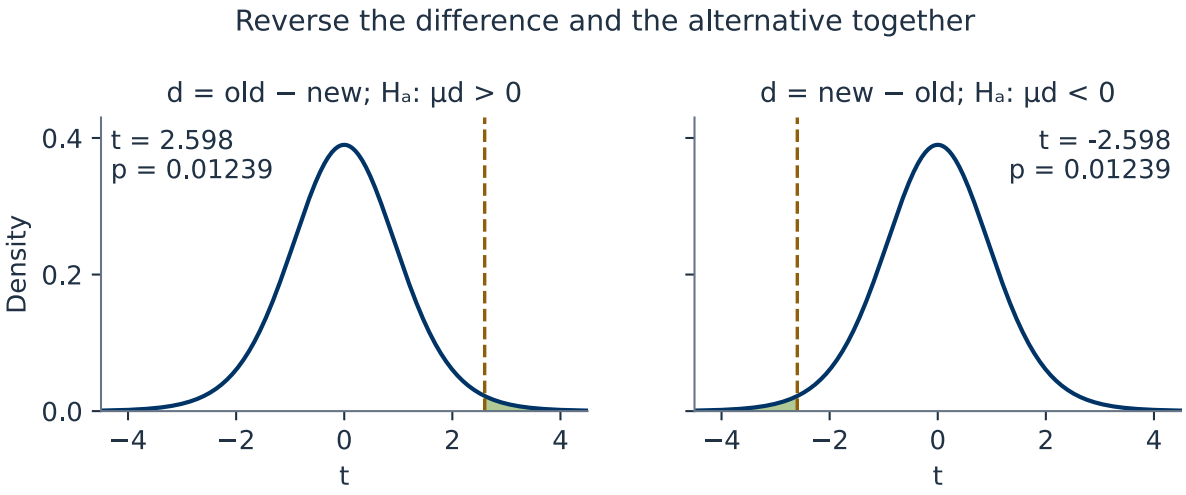
- Set hypotheses for a defined paired difference.
- Perform a paired one-sample t test.
- Separate evidence of change from evidence of causation.

CONCEPT AND INTERPRETATION

Define what one difference means

In paired data, a meaningful design link joins the two observations. Construct a single difference for each pair and state the order. Let $d = \text{first} - \text{second}$ and let μ_d denote the population mean of those differences. A test of no average change uses $H_0 : \mu_d = 0$. If a positive difference represents improvement, the alternative for average improvement is $H_a : \mu_d > 0$. Changing the subtraction order changes the signs in the hypotheses and statistic, but it must not change the substantive question.

The sign changes; the evidence does not



The same paired interface comparison gives mirrored t statistics and equal one-sided p-values when the subtraction order and alternative are changed consistently.

Pairing changes the analysis unit

The differences, not the original measurements treated separately, form the sample. With n pairs, use $t = \bar{d}/(s_d/\sqrt{n})$ for a zero null value and $df = n - 1$. The sample standard deviation s_d must be calculated from the difference list. It generally cannot be reconstructed from the two original standard deviations alone. Differences within a subject can be much less variable than measurements across subjects. That is a reason to pair thoughtfully before collecting data; arbitrarily matching rows afterward does not create the intended design.

Check the right distribution and independence

Use random sampling of pairs or a properly randomized experiment. When sampling without replacement, check the 10% condition using the number of pairs and the eligible population of pairs. When n is small, inspect the distribution of differences for strong skewness or outliers. The two original measurement distributions do not replace this check. Different pairs should be independent. For a crossover experiment, consider order, practice, fatigue, and carryover effects; randomized order can balance some effects but cannot guarantee that every carryover problem disappears.

Interpret evidence in the defined direction

The p-value is computed under a zero population mean difference. It measures the probability of a difference-based t statistic at least as extreme in the direction of the alternative. Compare it with the significance level and write the conclusion about μ_d in context. A rejection says that a nonzero population mean difference is supported; it does not say that every individual changes in the same direction. Include units and keep the sign convention visible. A useful report also estimates how large the average change may be.

A change is not automatically a treatment effect

A before/after sample without a randomized comparison may show a real average change while leaving its cause unresolved. Students can improve through ordinary practice, equipment can drift, and weather can change between measurements. A paired analysis controls stable differences between subjects, not all changes over time. Randomized treatment order in a suitable crossover design supports a stronger treatment comparison. Generalization still depends on recruitment. State a causal conclusion only when the design supports it, rather than allowing a low p-value to replace the design argument.

KEY IDEA

Test paired data through one defined difference per pair.

WATCH FOR

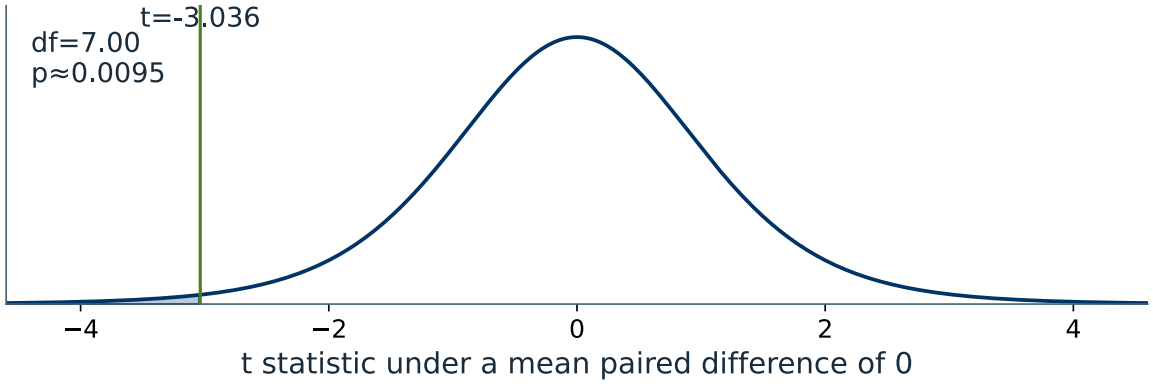
A statistically significant before/after change does not by itself identify the cause.

OPENSTAX WORKED EXAMPLE

A before-and-after test on eight subjects

OpenStax gives sensory scores before and after hypnotism for eight randomly selected subjects. Before:6.6,6.5,9.0,10.3,11.3,8.1,6.3,11.6. After:6.8,2.4,7.4,8.5,8.1,6.1,3.4,2.0. Lower scores indicate less pain. Assume independent subjects from a population of at least 80 and normally distributed after-minus-before differences. At α=0.05, is there convincing evidence that the population mean after-minus-before score is negative?

Locate the observed statistic in the null model



The shaded area is the p-value in the direction specified by the alternative. The distribution assumes the null hypothesis and the stated conditions.

1. Define d=after−before for each subject and μ_d as the population mean of these differences. H_0:μ_d=0; H_a:μ_d<0. A decrease is negative under this subtraction order.

2. The eight differences are 0.2,−4.1,−1.6,−1.8,−3.2,−2.0,−2.9,−9.6. Use a one-sample t test on this list, with n=8 and df=7, because measurements are paired within subjects.

3. Random selection and the 10% condition support treating subjects as independent. The source stipulates a normal difference population; the small sample alone cannot establish that condition. The difference −9.6 makes checking the assumption especially important in an actual study.

4. The mean difference is −3.125 score units, s_d=2.91143, and SE=1.02935. The statistic is t=−3.125/1.02935=−3.03591. The lower-tail p-value is 0.009478.

5. Since 0.009478<0.05, reject H_0. Under the stated assumptions, the data provide convincing evidence that the population mean sensory score is lower after than before. The design does not rule out other explanations for a before/after change.

Conclusion. This paired comparison supports a population mean decrease under its assumptions. It does not by itself establish that hypnotism caused the decrease.

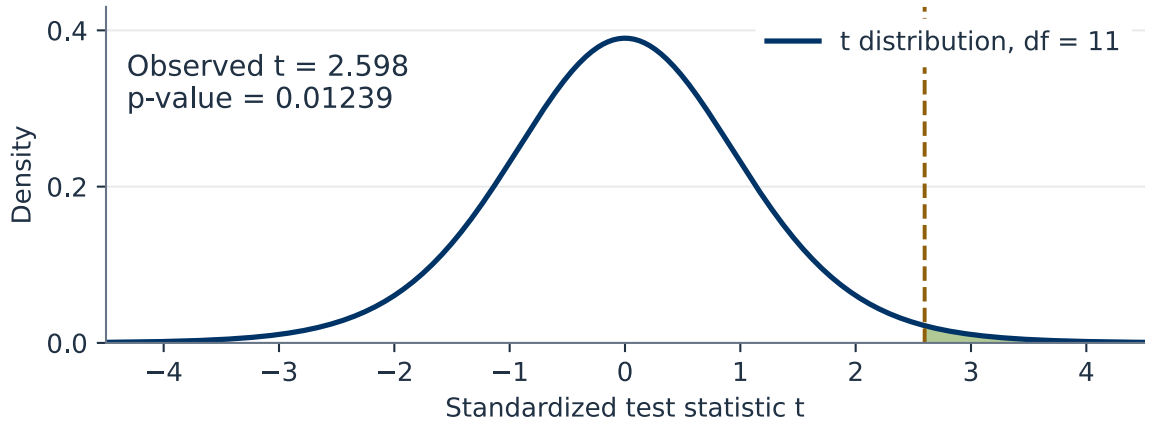
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.4, Example 10.11, pp.528-529 (PDF 540-541). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Does a revised interface reduce time?

Illustrative crossover experiment: 12 independent participants use two interfaces in randomized order. Let d=old−new completion time. The differences have mean 1.5 seconds and s_d=2.0 seconds; their plot has no strong skewness or outliers. Test whether the new interface lowers mean time at 0.05.

A test on twelve paired differences



Worked interface differences: n = 12, mean old − new = 1.5 seconds, s_d = 2.0 seconds. The one-sample paired t statistic uses 11 degrees of freedom.

09

1. Let μ_d be the mean old-minus-new completion time for comparable participants under these conditions. State H_0: $\mu_d=0$ and H_a: $\mu_d>0$.

2. The same participants use both interfaces, so analyze 12 differences with a one-sample paired t test. Order is randomized; participants are independent. The suitable difference plot supports the small-sample condition.

3. $SE=2/\sqrt{12}\approx0.57735$ second. $t=(1.5-0)/0.57735\approx2.59808$ and $df=11$.

4. The upper-tail p-value is 0.01239. Under no population mean difference, a difference-based t statistic at least 2.59808 has probability about 0.01239.

5. Because $0.01239<0.05$, reject H_0. There is convincing evidence of lower mean completion time with the new interface under the experimental conditions. Participant recruitment limits broader generalization.

Conclusion. The randomized crossover comparison supports an average time reduction for comparable participants. It does not establish that every participant benefits.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.4, pp.527-530 (PDF 539-542), Example 10.11. The eight paired sensory-score observations and normal-difference assumption are retained from the source. AP equality-form null hypotheses are used. The source's causal claim about hypnotism is narrowed because a before/after comparison without a randomized comparison group does not establish treatment causation. https://openstax.org/details/books/introductory-statistics-2e

S09.03 | Testing a difference between independent means

CED 4.9, 4.10

- State hypotheses for two defined population means.
- Use an unpooled two-sample t test with technology degrees of freedom.
- Make a contextual conclusion tied to design and practical size.

CONCEPT AND INTERPRETATION

State the two-population comparison

Define μ_1 and μ_2 as population means for the same quantitative response measured in two groups. State the null as $H_0 : \mu_1 - \mu_2 = 0$. Choose the alternative direction from the investigative question. "Different" gives a two-sided alternative, while a specific prediction of a larger or smaller first mean gives the corresponding one-sided alternative. The difference of sample means estimates the parameter but does not replace it in the hypotheses. Preserve the population definitions and group order in every sentence.

Require independent groups

The two-sample t procedure applies to independent samples or independent treatment groups in a randomized experiment. Each person or object belongs to one group for the comparison; there is no meaningful matched link across groups. If the same subjects receive both treatments, use paired differences instead. Random sampling supports population generalization and randomized treatment assignment supports a causal comparison. These are separate design features. A result from independent convenience groups can be numerically analyzed, but a formal population conclusion requires assumptions that such selection may not support.

Check conditions for both sources of data

For random samples without replacement, verify $n_1 \leq 0.10N_1$ and $n_2 \leq 0.10N_2$. The sampling-fraction check is unnecessary solely for random assignment in an experiment. Examine the sample-data condition for both groups. Both sample sizes at least 30 support normal approximations; alternatively, both population distributions may be approximately normal. If either sample is below 30, inspect both sample distributions for strong skewness and outliers. Do not add the sample sizes and declare a total above 30 sufficient. For example, one group of 5 and another of 60 still leaves a small group whose distribution needs attention.

Calculate unpooled uncertainty

For a null difference of zero, $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$. Both sample variances contribute to the standard error. Use an unpooled two-sample t procedure, often called Welch's procedure, with degrees of freedom supplied by technology. Do not select pooled variance merely because the null states that the means are equal; equal means and equal variances are different assumptions. The statistic's sign follows the subtraction order. Use the appropriate t tail or tails for the prespecified alternative.

Same mean difference, different standardized evidence

Study	n in each group	SD in each group	SE of difference	Observed t
I	16	8	2.828	1.414
II	64	8	1.414	2.828
III	16	4	1.414	2.828

All three investigation studies compare means 54 and 50. Studies II and III share t but have different degrees of freedom, so their p-values need not match.

Connect the numerical decision to the question

Interpret the p-value under the null hypothesis of equal population means. Compare it with the predetermined significance level and state reject or fail to reject. Then answer the original question in terms of the population means and response units. Include a confidence interval when the plausible magnitude matters. Statistical significance does not mean all individual responses differ, and it does not prove that an observed group label caused the difference. The strength of a causal or generalization claim comes from the design as well as the inference calculation.

KEY IDEA

A two-sample t test compares two population means using independent-group uncertainty.

WATCH FOR

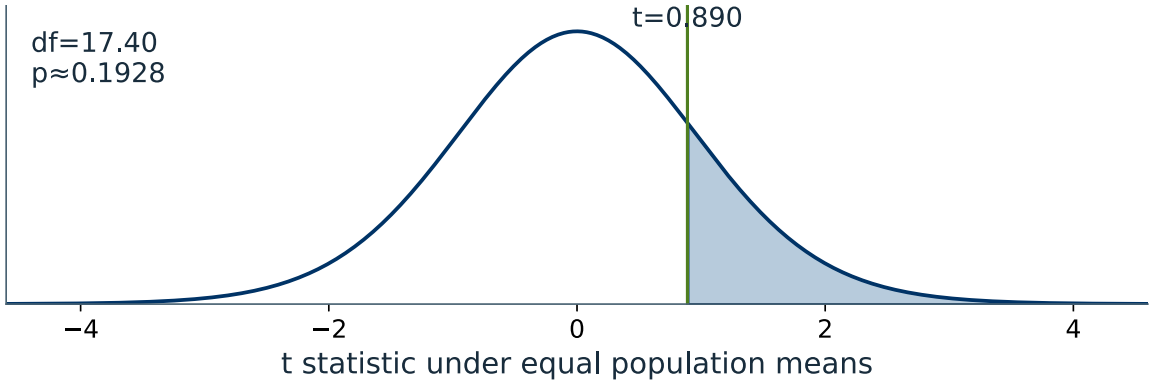
Equality of means under the null does not justify pooling variances.

OPENSTAX WORKED EXAMPLE

Do graduates of College A take more math classes?

OpenStax compares graduates of two colleges. An independent SRS of 11 College A graduates has mean 4 math classes and s=1.5 classes. An SRS of 9 College B graduates has mean 3.5 classes and s=1 class. Assume the source's approximately normal population models and at least 110 and 90 eligible graduates, respectively. At α=0.01, is there convincing evidence that College A's population mean is higher?

Locate the observed statistic in the null model



The shaded area is the p-value in the direction specified by the alternative. The distribution assumes the null hypothesis and the stated conditions.

1. Let μ_A and μ_B be the mean numbers of math classes completed by the two graduate populations. H_0:$\mu_A-\mu_B=0$; H_a:$\mu_A-\mu_B>0$.

2. Use an unpooled two-sample t test. The samples are independent and random. Both meet the 10% condition; the stated normal population models support inference with these small samples.

3. The estimated difference is 4−3.5=0.5 class. SE=$\sqrt{(1.5^2/11+1^2/9)}$=0.561833 class. Technology gives Welch df≈17.39784.

4. The test statistic is t=0.5/0.561833=0.889944. The upper-tail p-value is 0.192819.

5. Since 0.192819>0.01, fail to reject H_0. The sample does not provide convincing evidence that College A's population mean number of completed math classes exceeds College B's.

Conclusion. A sample mean difference of half a class is compatible with substantial sampling variation here. Failure to reject does not establish equal population means.

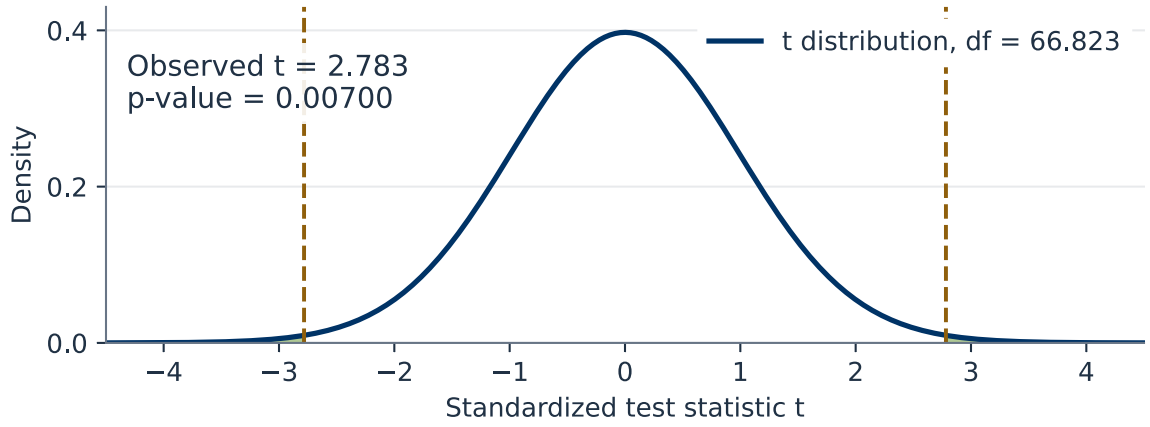
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.1, Example 10.2, pp.514-516 (PDF 526-528). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Comparing task instructions

In an illustrative experiment, 70 volunteers are randomly assigned to instructions A or B, with 35 per group. Completion times for A have mean 42 minutes and s = 8 minutes; for B, mean 37 minutes and s = 7 minutes. At $\alpha = 0.05$, is there evidence that the instructions affect mean completion time under these conditions?

Two-sided evidence for instruction groups



Worked randomized groups: n_A = n_B = 35; means 42 and 37 minutes; sample SDs 8 and 7 minutes. H_a: $\mu_A - \mu_B \neq 0$, so both equally extreme tails count.

1. Let μ_A and μ_B be mean completion times under the two instruction conditions for comparable participants. H_0: $\mu_A - \mu_B = 0$; H_a: $\mu_A - \mu_B \neq 0$.

2. Random assignment supports comparison of independent groups. Both group sizes are 35, so the sample-data condition is supported. A 10% population sampling condition is not required for the random-assignment comparison.

3. Use an unpooled two-sample t test. The estimated mean difference is 5 minutes, with $SE = \sqrt{(64/35 + 49/35)}$. SE = 1.7968; df = 66.823.

4. For the two-sided alternative, t = 2.7827; two-sided p-value = 0.00700.

5. The p-value is below 0.05, so reject H_0. There is convincing evidence of different mean completion times under the two instruction conditions. Random assignment supports attributing the difference to instructions for these experimental conditions, while volunteer recruitment limits broader generalization.

Conclusion. The volunteers using instructions A averaged 5 minutes longer than those using B. The randomized comparison gives evidence of an instructions effect under these conditions (two-sided p ≈ 0.007). The volunteers do not establish its size for all users.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 10.1, pp.512-516 (PDF 524-528), Example 10.2. The two colleges' sample sizes, means, standard deviations, and one-sided question are retained. Independent SRSs and populations of at least 110 and 90 graduates are added explicitly. Use unpooled t inference and a contextual failure-to-reject conclusion, not proof that the colleges are equal. https://openstax.org/details/books/introductory-statistics-2e

S09.04 | Choosing inference and linking tests with intervals

CED 4.2, 4.3, 4.4, 4.5, 4.7, 4.8, 4.9, 4.10, 1.1, 1.10

- Choose a method from response type, design, and inferential goal.
- Explain when interval and test decisions correspond.
- Write a valid investigative question and a defensible inference argument.

CONCEPT AND INTERPRETATION

The question comes before a formula

A valid investigative question identifies a population or study setting, a variable or variables, and a purpose that anticipates variability in data. Decide whether you seek an estimate or assess a claim. “How long does a randomly selected job take?” differs from “What is the mean completion time?” and “Does a new method reduce mean time?” The first may require prediction, the second estimation, and the third a comparative study and test. Do not select a procedure from a keyword without identifying the actual observational unit and response.

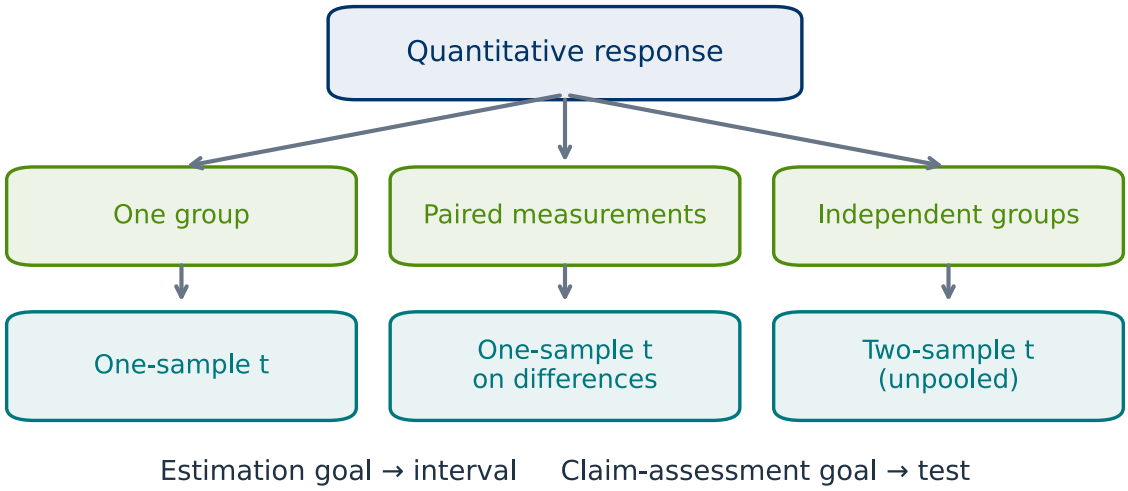
Separate categorical and quantitative responses

A binary categorical response leads to a population proportion. One population suggests one-proportion procedures; two independent groups suggest a difference in proportions. A quantitative response leads to a population mean. One group uses one-sample t procedures, paired observations use one-sample t procedures on differences, and independent groups use unpooled two-sample t procedures. A two-way count table comparing categorical distributions suggests chi-square homogeneity or independence, with the choice determined by how the data were collected. The numbers in a problem are not enough to classify the variable.

Identify pairing before calculating

Ask whether two observations are linked within the same individual or within a deliberately formed pair. If so, define the difference and analyze that quantitative response. If groups were independently sampled or assigned, use the independent-group comparison. Merely having two columns, equal sample sizes, or before/after labels in unrelated groups does not establish pairing. The design determines dependence. Choosing a paired procedure because its p-value is smaller is not an acceptable substitute for examining how the data were generated.

Let response and dependence determine the mean procedure



This map is for quantitative responses with unknown population standard deviations. Categorical responses lead to proportion or chi-square procedures instead.

09

Intervals and tests can answer complementary questions

A confidence interval estimates plausible parameter values and effect size. A test assesses a specified null value against a prespecified alternative. For the same ordinary t procedure, a two-sided test at significance α rejects exactly when the null value is outside the corresponding two-sided $100(1-\alpha)\%$ confidence interval, apart from rounding. This matching statement requires the same data, model, standard error, and two-sided structure. Do not pair a one-sided 5% test mechanically with a two-sided 95% interval. Proportion methods may use different standard errors for intervals and tests, so exact equivalence also needs care there.

Build an argument rather than a calculator transcript

A complete response defines the parameter, names the method, gives the required condition evidence, shows the calculation or relevant technology output, and reaches a contextual conclusion. For a test, include hypotheses and the comparison between p-value and significance. For an interval, interpret both the estimate and any claim it assesses. Then explain the scope of inference and important limitations. These pieces should answer the investigative question in order. A correct output screen without a parameter or design argument does not show that the procedure answers the intended question.

KEY IDEA

Choose the method from the question, response, and design before using technology.

WATCH FOR

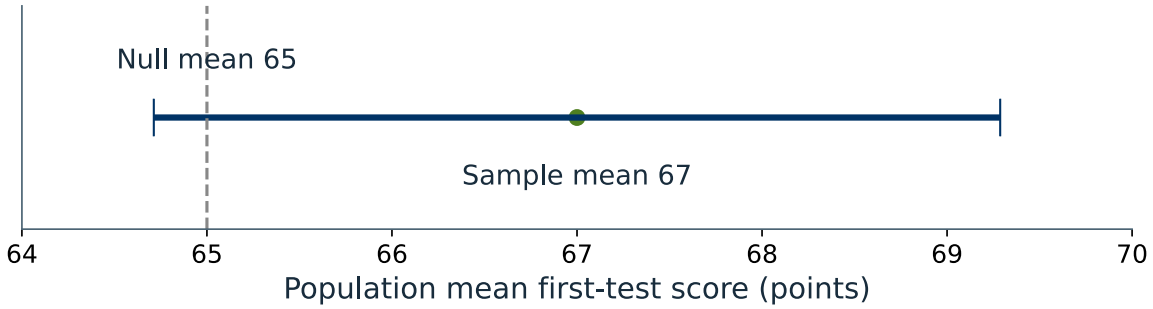
Matching test/interval decisions require matching procedures and compatible tail structure.

OPENSTAX WORKED EXAMPLE

A two-sided test and its matching interval

Return to the ten first-test scores in OpenStax Example 9.16. Under the same stated assumptions, ask whether the population mean differs from 65 points. Compare a two-sided 5% test with a 95% confidence interval.

The matching 95% interval contains 65



The interval (64.713,69.287) contains the null mean. The matching two-sided 5% test therefore fails to reject.

1. The parameter is μ, the population mean first-test score. The null is H_0:μ=65 and the alternative is H_a:μ≠65. Use the same one-sample t model and the same ten scores for both procedures.

2. The sample mean is 67, s=3.197221, SE=1.011050, and df=9. The test statistic is 1.978141 and the two-sided p-value is 0.079296.

3. The 95% critical value is 2.262157. The interval is 67±2.262157(1.011050), or(64.71285,69.28715) points.

4. The p-value exceeds 0.05, so the test fails to reject. The interval contains 65, so 65 is compatible with the 95% interval. These decisions agree because the data, t method, and two-sided levels match.

5. The original source's greater-than test at 0.05 rejects, with p≈0.03965. That does not contradict this two-sided result: the directional and two-sided questions use different tail areas.

Conclusion. The test addresses a specified claim, while the interval also displays plausible population means. Matching their levels requires attention to whether the test is one-sided or two-sided.

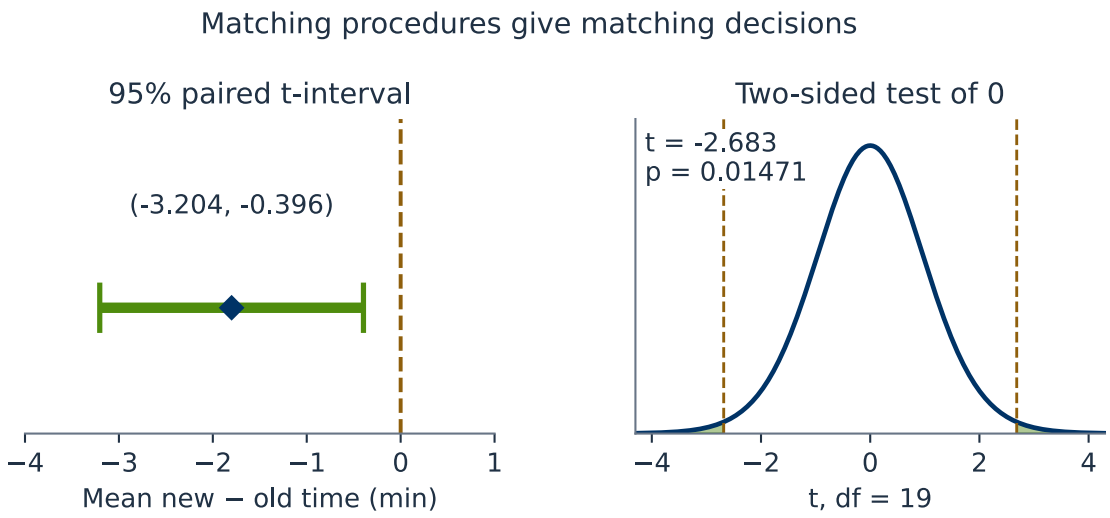
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Extension of Section 9.5, Example 9.16, pp.475-476 (PDF 487-488). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Choose the analysis before reading the result

A teacher wants to know whether a new checklist reduces time needed to correct formatting errors. Twenty volunteers complete two comparable tasks, one under each checklist. Independently randomize which checklist is applied to which task and randomize the order of tasks for each participant. The new-minus-old differences have mean −1.8 minutes, s_d=3 minutes, and no strong skewness or outliers. Use a two-sided 5% test and a 95% interval.

The checklist interval and test agree



Worked data: 20 paired new-minus-old differences have mean −1.8 minutes and s_d = 3 minutes. The 95% interval excludes zero; the matching two-sided 5% test rejects zero.

09

1. The response is quantitative and paired within participant. Define μ_d as the population mean new-minus-old correction time for comparable participants. The interval and test must use the 20 differences.

2. For the requested two-sided test, state H_0: $\mu_d=0$ and H_a: $\mu_d\neq0$. A directional claim could instead be prespecified, but this problem requests the two-sided comparison.

3. Randomized task-to-checklist allocation and order, independent participants, and a suitable difference plot support the paired t procedure. Use n=20 and df=19; no sampling-fraction argument is required merely for random assignment.

4. Calculation: t = -2.6833; df = 19; p-value = 0.01471. The 95% interval is t^* = 2.0930; SE = 0.6708; interval (-3.204, -0.396), in minutes.

5. The test rejects at 0.05 and the matching interval excludes 0. The negative interval suggests a mean reduction with the new checklist. Randomized order supports the treatment comparison, while volunteers limit generalization.

Conclusion. The 95% interval corresponds to an average reduction of about 0.40 to 3.20 minutes with the new checklist under the conditions tested. Both endpoints imply a shorter time, consistent with the two-sided test at the 5% level.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 9.5 Example 9.16, pp.475-476; Section 10.1 Example 10.2, pp.514-516; Section 10.4 Example 10.11, pp.528-529 (PDF 487-488,526-528,540-541). Source study designs are used to compare one-sample, independent-group, and paired inference. The matching t interval/test calculation is developed using the source first-test scores. No inference on regression slope is introduced. https://openstax.org/details/books/introductory-statistics-2e

Unit synthesis

1. State hypotheses for testing whether population mean time is below 20 minutes.

2. For n = 25, sample mean 104, and s = 10, test $\mu = 100$ against a two-sided alternative. Show the standard error.

3. Why does a paired test use differences?

4. What is wrong with saying p=0.07 means a 7% probability the null is true?

5. How does a matching 95% t-interval support a two-sided 5% test?

6. What design supports causation and what supports generalization?

Unit 10 | Two-Variable Data and Regression

Relationships between quantitative variables require both graphical and numerical reasoning. This unit begins with paired observations and scatterplots, then develops correlation, prediction, residuals, and least-squares regression. A recurring illustrative print-job dataset lets students connect different representations of the same relationship. Counterexamples show why a high correlation cannot establish linear form or causation. Students use technology to fit models while taking responsibility for variable roles, units, input range, and contextual interpretations. The final lesson explains the coefficient of determination through a mean-only prediction baseline. Residual standard deviation is identified as a supporting extension; inference for regression slopes is outside the revised required scope.

CED topics: 5.1, 5.2, 5.3, 5.4, 5.5

IN THIS UNIT

Lesson	Focus
S10.01	Seeing relationships in paired quantitative data
S10.02	Correlation: strength of a linear association
S10.03	Linear predictions and their limits
S10.04	Residuals and checking a linear model
S10.05	Least squares and explained variation



Local question

Elevation and walking time may be related. What would a fitted line describe, and why would a strong association alone not establish causation?

Table Rock summit landscape

S10.01 | Seeing relationships in paired quantitative data

CED 5.1

- Construct a scatterplot with explanatory and response variables.
- Describe direction, form, strength, and unusual features in context.
- Use a scatterplot to assess a claim without overclaiming causation.

CONCEPT AND INTERPRETATION

One point records one observational unit

A bivariate quantitative dataset contains two numerical measurements for each person, object, or event. Keep each ordered pair together. If one measurement is a job's page count and the other is its completion time, one plotted point represents one job. Sorting one column without moving the corresponding second column destroys the observed relationship. Begin by defining the observational unit and the measurement units for both variables. A scatterplot is not appropriate merely because categories have been assigned numerical codes; both variables must have meaningful quantitative scales.

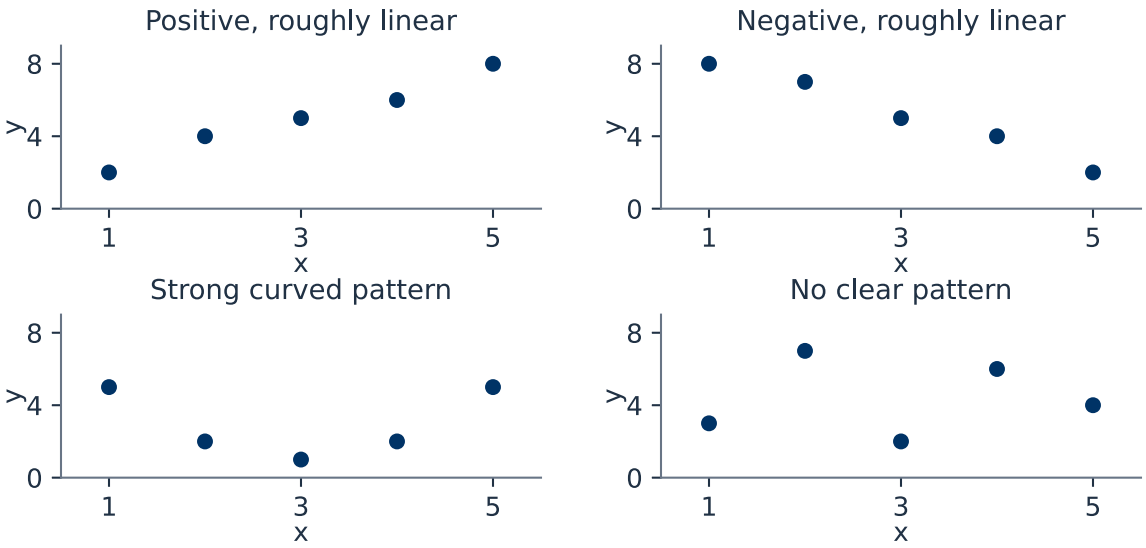
Choose axes to match the question

The explanatory variable x is the quantity used to explain or predict the response variable y. Place x on the horizontal axis and y on the vertical axis. Naming a variable explanatory does not establish that it causes changes in y. Sometimes neither variable is naturally explanatory; the choice of axes then follows the intended prediction question. Label both axes with names and units, choose scales that show the data honestly, and plot every pair. An uneven numerical scale can make a relationship appear stronger or weaker than it is.

Describe the overall pattern in context

A useful description addresses direction, form, strength, and unusual features. A positive association means that larger x values tend to accompany larger y values; a negative association means that larger x values tend to accompany smaller y values. Describe form as approximately linear or nonlinear and explain the visible pattern. Strength refers to how closely points follow that pattern. A strong curved relationship is still strong, but it is not well summarized as a strong linear association. Use the actual variables in the description so a reader knows what increases or decreases.

Describe the form as well as the direction



These are artificial pattern examples. Strength refers to closeness to the observed pattern; a curved relationship can be strong.

Unusual points and clusters can change the story

Look for groups of points, large gaps, observations far from the main pattern, and observations at unusual x values. A point unusual in y relative to its x value may signal a recording error or a meaningful exceptional case. A point far out in x can strongly affect a fitted line, especially if it does not follow the existing pattern. Investigate the context before deleting a point. Separate clusters may represent different populations or conditions, and an overall trend can conceal different within-group patterns.

Use a picture as evidence, with limits

A scatterplot can support a descriptive claim about the observed association, such as larger print jobs tending to take longer. It cannot by itself establish a population causal effect. The plot also does not show how observations were selected, whether important variables were omitted, or how precisely each value was measured. Link the graphical evidence back to the investigative question and study design. A pattern can motivate a linear model or a better experiment; it does not make those next steps automatic. Always inspect the plot before calculating a numerical summary.

KEY IDEA

A scatterplot preserves the two measurements belonging to each observational unit.

WATCH FOR

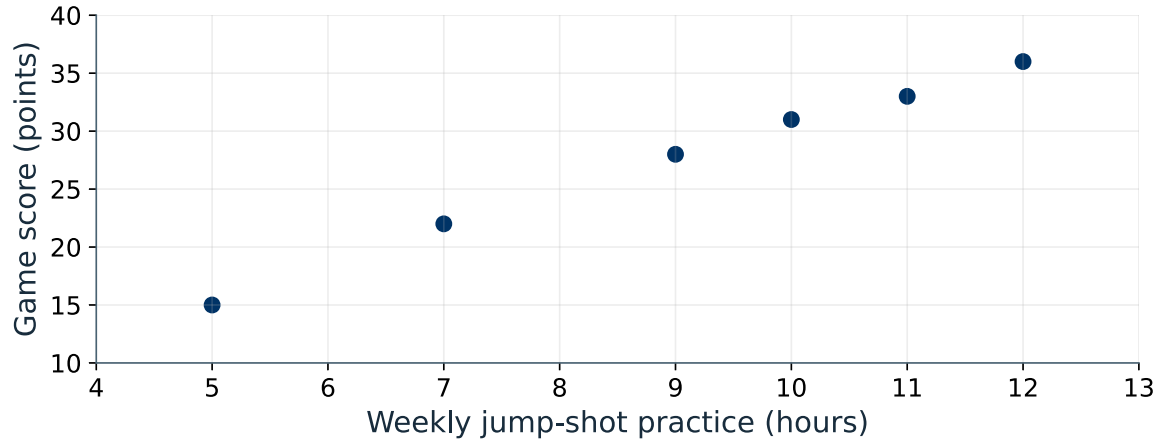
An explanatory variable is a prediction role, not a causal conclusion.

OPENSTAX WORKED EXAMPLE

Practice hours and game points

OpenStax describes Amelia, a high-school basketball player. Her recorded pairs of weekly jump-shot practice hours and game points are (5,15),(7,22),(9,28),(10,31),(11,33),(12,36). Construct a scatterplot and describe the association. Does the plot support her belief that more practice hours are associated with more game points?

Amelia's practice hours and game points



OpenStax Try It 12.5. Each point preserves one reported practice-hour/game-point pair. These are the source data, not observations from SMS students.

10

1. Each ordered pair links one week's jump-shot practice with the corresponding game score. Keep each pair together. Put practice hours on the horizontal axis and game points on the vertical axis.

2. Use axes that include 5-12 hours and 15-36 points. For example, the point(7,22) says that the week with 7 practice hours was paired with 22 points in the game.

3. The association is strong, positive, and approximately linear. As practice hours increase, game points tend to increase; no isolated point is apparent in these six observations.

4. The pattern supports an association in Amelia's recorded data. Because practice time was not randomly assigned, the plot alone does not establish how many additional points an extra hour of practice would cause.

Conclusion. The source data connect a statistical description to a familiar school activity while keeping the causal claim within what the study design supports.

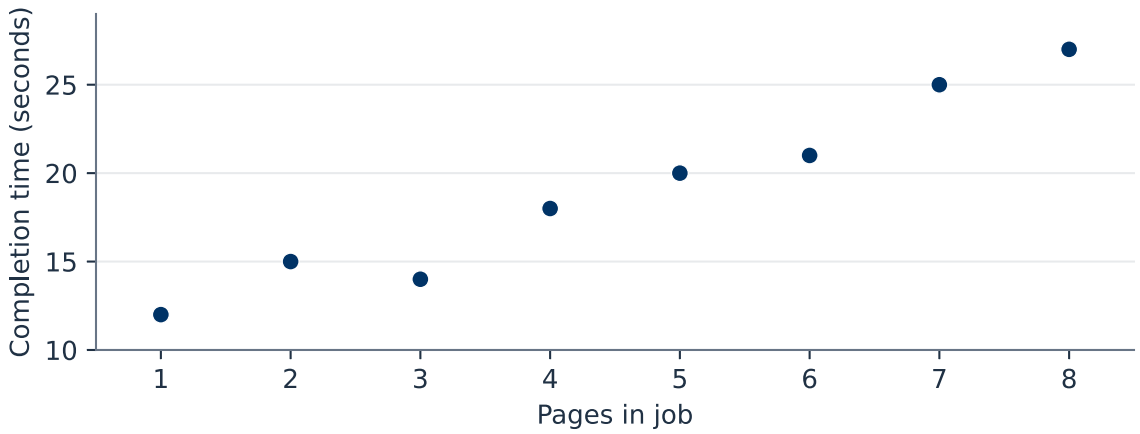
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 12.2, Try It 12.5, pp.621-622 (PDF 633-634). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Do larger jobs take longer?

Illustrative print-job data pair pages x with completion time y in seconds: (1,12),(2,15),(3,14),(4,18),(5,20),(6,21),(7,25),(8,27). Describe the association and assess the claim that larger jobs tend to take longer.

The paired print-job observations



Worked illustrative pairs: (1,12), (2,15), (3,14), (4,18), (5,20), (6,21), (7,25), (8,27). One point represents one print job.

1. The observational unit is one print job. Both page count and completion time are quantitative. Use pages as explanatory x and seconds as response y because the question asks whether pages helps predict time.

2. Construct a scatterplot with a horizontal axis spanning 1 to 8 pages and a vertical axis spanning approximately 10 to 30 seconds. Plot the eight given pairs without rearranging one column independently.

3. The association is positive: jobs with more pages tend to have longer completion times. The form is roughly linear; a straight trend describes the overall change reasonably well.

4. The points are close to that trend, so the association appears strong. The 3-page job at 14 seconds is somewhat below nearby values, but no isolated point or separate cluster is obvious.

5. The plot supports the descriptive claim for these illustrative jobs. It does not prove that page count is the sole cause; printer settings, file complexity, and warm-up state could also affect time.

Conclusion. The data show a strong, approximately linear positive association between page count and completion time, with no obvious separate cluster.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 12.2-12.3, pp.620-631 (PDF 632-643), Try It 12.5 and Example 12.6. Source data, scatterplot reasoning, least-squares development, and interpretations are retained/adapted. Slope and correlation significance tests from Section 12.4 are excluded from the revised AP core. Contextual interpretations use predicted change and do not imply causation. https://openstax.org/details/books/introductory-statistics-2e

S10.02 | Correlation: strength of a linear association

CED 5.2

- Interpret the sign and magnitude of correlation.
- Explain invariance to positive unit changes and limits of r.
- Distinguish correlation from causation and nonlinear association.

CONCEPT AND INTERPRETATION

What correlation summarizes

The correlation coefficient r summarizes the direction and strength of a linear association between two quantitative variables. It is unit-free and lies between −1 and 1, inclusive. A positive value indicates an upward linear association; a negative value indicates a downward one. Values close to either endpoint show points close to a straight-line pattern. A value near zero indicates little linear association. Interpret r with a scatterplot and the variable names, rather than translating it mechanically into a causal or percentage claim.

Magnitude measures linear closeness

A correlation of −0.90 represents a stronger linear association than 0.40, because 0.90 is closer to 1 in magnitude. The negative sign indicates direction, not weakness. Perfect correlations of 1 or −1 occur when all points lie on a nonvertical straight line, provided both variables vary. There is no universal cutoff that turns every relationship into “strong” or “weak”; context, sample size, and the plot matter. Correlation is not the slope. A steep line can have small r if points scatter widely, and a shallow line can have r close to 1.

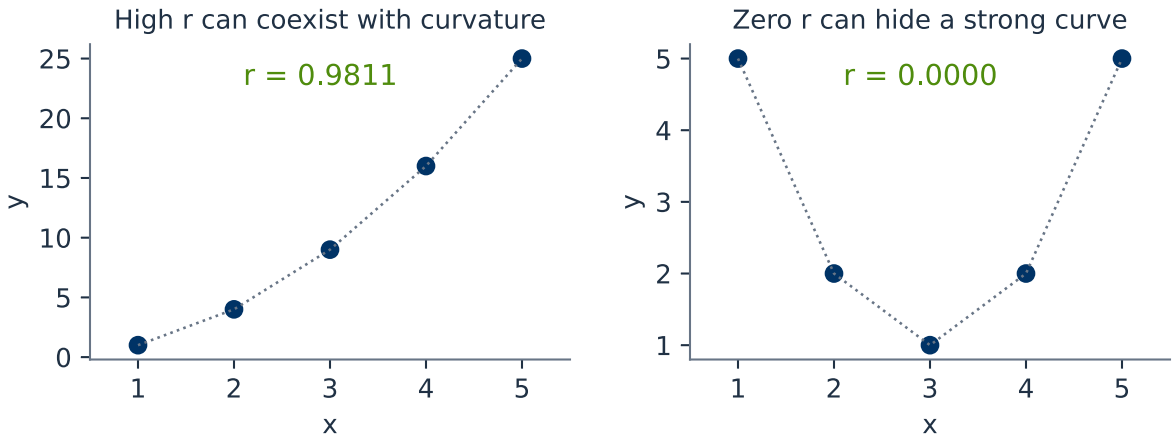
Units and axis roles

Changing centimeters to meters or seconds to minutes multiplies a variable by a positive constant and leaves r unchanged. Adding a constant also leaves it unchanged. Multiplying one variable by a negative constant reverses r’s sign but not its magnitude. Interchanging x and y leaves correlation unchanged because the standardized association is symmetric. Regression predictions are not symmetric in that way, because a model that predicts y from x has a different target than one predicting x from y. Keep correlation and prediction roles separate.

Why the scatterplot remains necessary

A strong curved association may have r near zero if positive and negative linear tendencies cancel. A correlation close to 1 or −1 can also occur for data that curve noticeably over a restricted range. One unusual observation can substantially change r, so investigate surprising points and compare the pattern with and without them when appropriate. A numerical summary does not reveal clusters, gaps, or the data-collection process. Always inspect the paired values graphically before deciding whether r is a useful description of their relationship.

Two reasons to inspect the scatterplot



Exact investigation data: left y = 1, 4, 9, 16, 25; right y = 5, 2, 1, 2, 5; both use x = 1, 2, 3, 4, 5. Correlation summarizes linear association only.

An association does not identify a cause

10

Correlation does not distinguish a direct causal effect from confounding, reverse causation, selection effects, or coincidence. For example, two variables may both respond to a third factor. To make a causal claim, consider whether an experiment assigned the explanatory condition randomly and whether the design controlled competing explanations. Random sampling may make the association representative of a population, but it does not randomize the explanatory variable. Describe what larger x values tend to accompany in y, and state separately what the design does or does not establish.

KEY IDEA

Correlation measures direction and strength of linear association, not slope or causation.

WATCH FOR

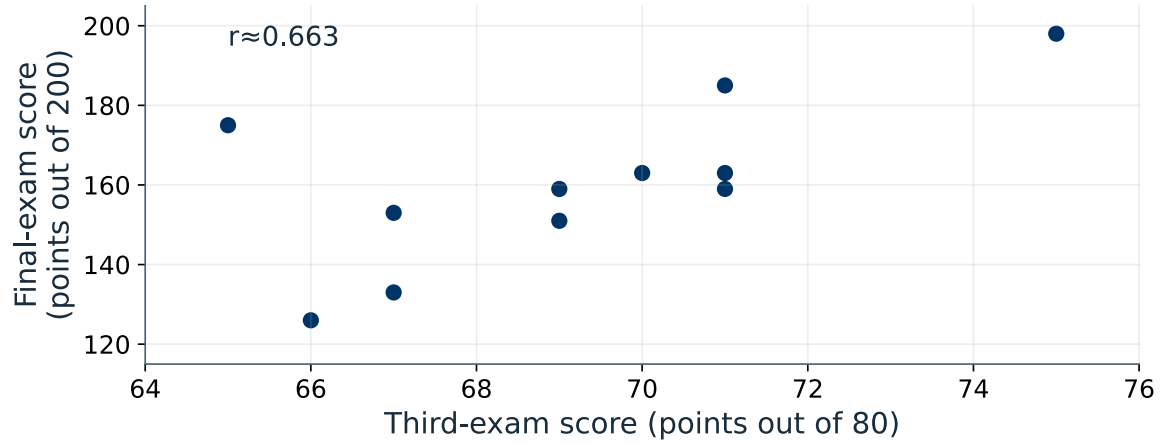
A correlation near zero can occur with a strong nonlinear relationship.

OPENSTAX WORKED EXAMPLE

Interpret a correlation between exam scores

The OpenStax third-exam/final-exam pairs are (65,175),(67,133),(71,185),(71,163),(66,126),(75,198),(67,153),(70,163),(71,159),(69,151),(69,159). The explanatory score x is out of 80 points and the response score y is out of 200 points. Technology gives $r \approx 0.66309$. Interpret this value and explain what happens to it if third-exam scores are converted to percentages.

The same third-exam score can accompany different finals



OpenStax Example 12.6, Table 12.3. The scatterplot retains all eleven students and shows why a numerical correlation is not a prediction-accuracy percentage.

1. The positive sign says that students with higher third-exam scores tend to have higher final-exam scores. The magnitude indicates a moderately strong positive linear association in these eleven students; the scatterplot still matters.

2. Converting the third-exam score to a percent multiplies each x by 100/80=1.25. This positive change of units leaves the standardized association, and hence r, unchanged.

3. The correlation is unitless. A value of 0.66309 does not mean that 66.309% of students earn a particular final score or that 66.309% of predictions are correct.

4. Random sampling can support generalization to the population sampled, but it does not make the earlier exam score a randomized treatment. The correlation alone is not a causal effect.

Conclusion. The correlation summarizes a linear association between the two scores. It remains the same after a positive linear change of either score's units.

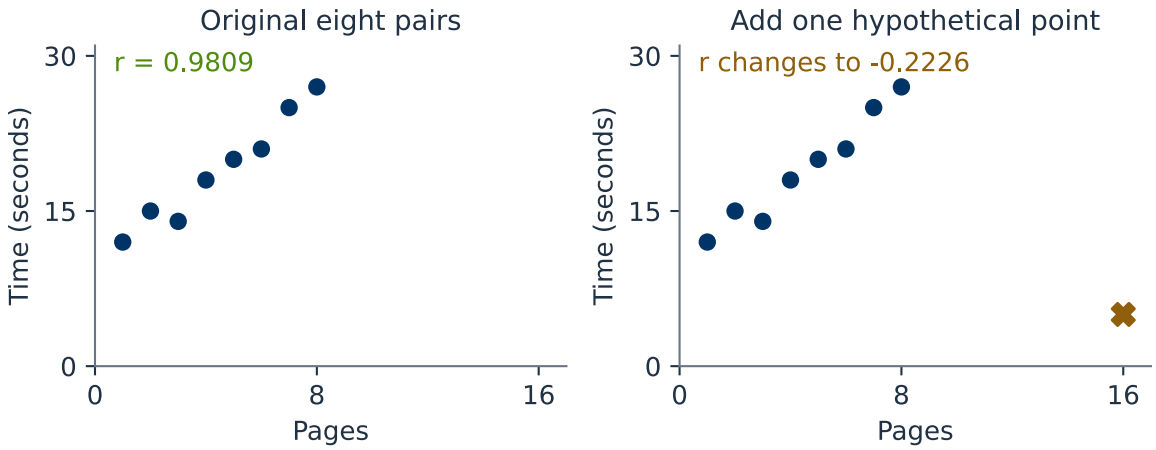
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 12.3, Example 12.6 and correlation discussion, pp.623-630 (PDF 635-642). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Interpreting a high correlation carefully

For the illustrative print jobs x=(1,2,3,4,5,6,7,8) pages and y=(12,15,14,18,20,21,25,27) seconds, technology reports r=0.98093. Interpret the statistic and explain two conclusions it does not support.

A single unusual observation can change r



Left: the exact worked print-job dataset. Right: a clearly hypothetical added pair (16 pages, 5 seconds), marked with an X, illustrates sensitivity; it is not an additional observed print job.

1. The variables are quantitative and the scatterplot is roughly linear, so correlation is an appropriate descriptive summary for these eight pairs.

2. The positive sign means larger page counts tend to accompany longer completion times. Its magnitude, close to 1, indicates a strong linear association in this dataset.

3. The value 0.98093 is unit-free. Expressing completion time in minutes would leave r unchanged, although it would change the numerical slope of a prediction line.

4. It is incorrect to say 98.093% of jobs are predicted correctly. Correlation is not an accuracy percentage or a proportion of observations. A later coefficient-of-determination calculation answers a different variation question.

5. It is also incorrect to say page count has been proved to cause completion time. These illustrative observations were not a randomized experiment; other job or printer features can matter.

Conclusion. Across the eight observed print jobs, larger page counts went with longer completion times in a strong positive linear pattern ($r \approx 0.981$). That correlation alone cannot establish a causal effect of adding pages.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 12.2-12.3, pp.620-631 (PDF 632-643), Try It 12.5 and Example 12.6. Source data, scatterplot reasoning, least-squares development, and interpretations are retained/adapted. Slope and correlation significance tests from Section 12.4 are excluded from the revised AP core. Contextual interpretations use predicted change and do not imply causation. https://openstax.org/details/books/introductory-statistics-2e

S10.03 | Linear predictions and their limits

CED 5.3, 5.5

- Use a fitted line to predict a response.
- Interpret slope and intercept with units and prediction language.
- Distinguish interpolation from extrapolation.

CONCEPT AND INTERPRETATION

A prediction equation has a defined direction

A linear model for predicting y from x is $\hat{y} = a + bx$. Here $\hat{y}$ is the predicted response, a is the intercept, and b is the slope. The actual observed response is y and generally differs from its prediction. State which variable is being predicted and which provides the input. A fitted line describes a trend in a set of paired data; it does not claim that every observation falls on the line. Use a linear model only when the relationship is reasonably described by a linear pattern.

Interpret slope as predicted response change

The slope b is the predicted change in y for a one-unit increase in x. Its units are response units per explanatory unit. If time in seconds is predicted from page count, a slope of 2.1 means predicted completion time increases by 2.1 seconds per additional page. Use “predicted” or “on average” to distinguish the fitted trend from an exact individual change. A negative slope means predicted y decreases as x increases. The slope is not the correlation, and changing measurement units changes its numerical value.

Units change coefficients, not the physical prediction

Response and input	Prediction equation
Seconds from pages x	$\hat{y}$ = 9.464286 + 2.119048x
Minutes from pages x	$\hat{m}$ = 0.157738 + 0.035317x
Seconds from tens of pages z	$\hat{y}$ = 9.464286 + 21.190476z

All equations describe the same fitted print-job model. Here x is pages and z is tens of pages, so x = 10z.

Interpret the intercept with contextual judgment

The intercept a is the predicted y value when x=0. This is a mathematical interpretation even when x=0 is outside the data range. A meaningful real-world interpretation requires zero to make sense and the prediction to be reasonable. A line fitted to adult heights, for example, should not be interpreted as reliably predicting a body measurement at height zero. A nonphysical or extrapolated intercept does not automatically invalidate every prediction within the observed range, but it does warn against giving the intercept more meaning than the data support.

Interpolation and extrapolation

Interpolation predicts within the range of x values used to fit the model. Extrapolation predicts beyond that range. Extrapolation is less reliable because the observed relationship may change outside the measured conditions. Even interpolation is uncertain: the model may have substantial scatter, hidden groups, or a weak design. When making a prediction, report the input value, the equation, substitution, the result with units, and whether it lies in range. Do not treat an algebraically available number as evidence that the model applies there.

Prediction does not reverse automatically

A regression equation predicting y from x minimizes errors in the y direction. Solving that same equation for x does not generally produce the least-squares regression for predicting x from y. The intended response determines which vertical discrepancies are minimized. Likewise, a model built in one setting should not automatically be transferred to another printer, student population, season, or measurement process. Describe the domain where the original data provide support. A useful prediction is an estimate accompanied by evidence about model adequacy and applicability, not merely a substituted number.

KEY IDEA

A regression prediction must retain variable roles, units, and the observed input range.

10

WATCH FOR

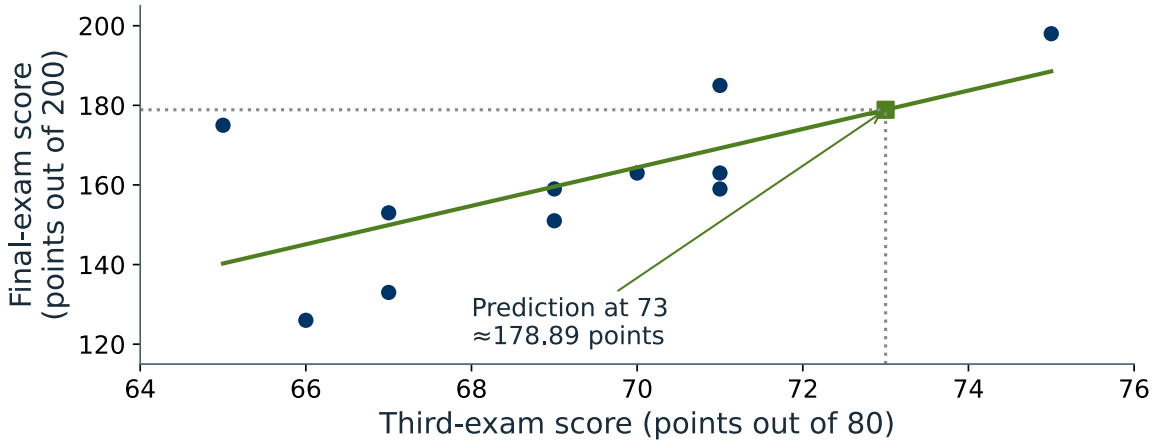
The intercept can have a mathematical meaning without a useful contextual interpretation.

OPENSTAX WORKED EXAMPLE

Predict a final-exam score

The OpenStax third-exam/final-exam pairs are $(65,175),(67,133),(71,185),(71,163),(66,126),(75,198),(67,153),(70,163),(71,159),(69,151),(69,159)$. The explanatory score x is out of 80 points and the response score y is out of 200 points. The fitted line is $\hat{y} = -173.51336 + 4.827394x$. Predict the final score for x=73, interpret the slope, and evaluate using the line at x=50.

A prediction within the observed range



The fitted line is shown only over the observed 65-75-point third-exam range. The square is the model's prediction at 73, not an additional student.

1. At x=73, the predicted final-exam score is −173.51336+4.827394(73)=178.8864 points out of 200. The value 73 is within the observed third-exam range 65-75.

2. The slope predicts an increase of about 4.83 final-exam points for each additional third-exam point. It describes the fitted association rather than a guaranteed change for an individual student.

3. The intercept predicts a final score of −173.51 for a third-exam score of 0. Zero is far outside 65-75 and a negative exam score is not meaningful here, so this intercept has no useful substantive interpretation.

4. For x=50, the equation gives 67.8563 final-exam points. This is extrapolation well below the observed range, so the data do not establish that the fitted trend continues there.

Conclusion. Use the line within its supported range and state predictions in final-exam points. A calculator can return an extrapolation without making it reliable.

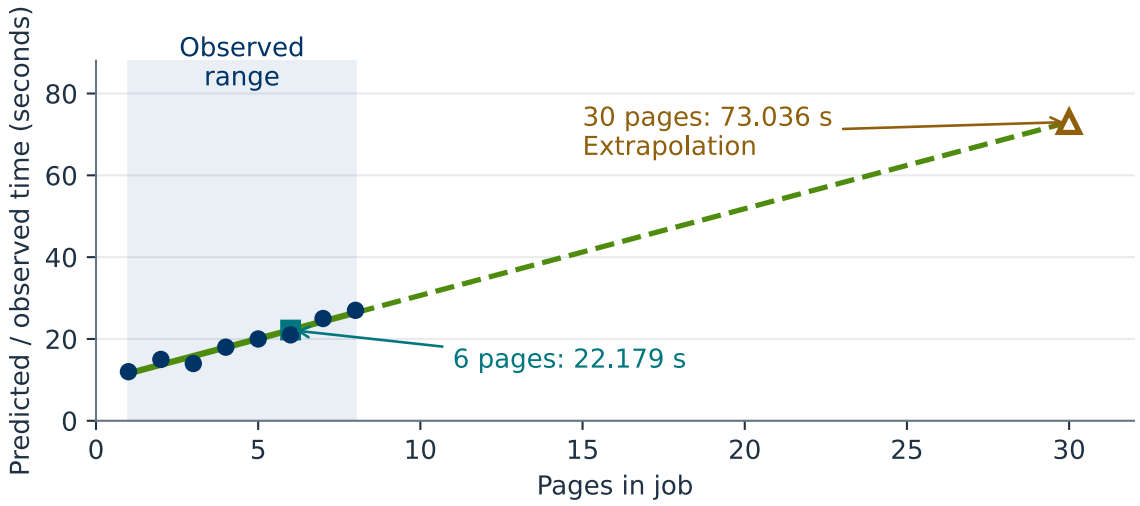
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 12.3, Example 12.6 and slope/prediction discussion, pp.623-629 (PDF 635-641). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

Predicting print-job completion time

For the illustrative jobs with observed page counts 1 through 8, a fitted model is predicted seconds=9.4643+2.11905(pages). Predict a 6-page job, interpret the coefficients, and evaluate a 30-page prediction.

Interpolation and extrapolation on the same line



Worked model: predicted seconds = 9.464286 + 2.119048 × pages. The shaded region is the observed 1-8-page range. The dashed continuation gives algebraic predictions without observed-range support.

1. Define x as page count and predicted y as completion time in seconds. The coefficient 2.11905 has units seconds per page; 9.4643 has units seconds.

2. For x=6, substitute into the model: predicted y=9.4643+2.11905(6)≈22.1786 seconds. Six pages lies within the observed 1- to 8-page range, so this is interpolation.

3. The slope means each additional page is associated with an increase of about 2.119 seconds in predicted completion time. It is a fitted average trend, not a guarantee for every job.

4. The intercept predicts 9.4643 seconds for a 0-page job. Zero pages was not observed, and a 0-page job may not be a meaningful operation; treat this as a model intercept rather than a measured warm-up time.

5. For 30 pages, the equation produces about 73.036 seconds. This is far outside the observed range, so the number is an extrapolation with weak empirical support. A different operating pattern may occur for large jobs.

Conclusion. The model predicts about 22.2 seconds for a 6-page print job, within the observed range of 1-8 pages. Its prediction of about 73.0 seconds for 30 pages depends on the untested assumption that the trend continues that far.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 12.2-12.3, pp.620-631 (PDF 632-643), Try It 12.5 and Example 12.6. Source data, scatterplot reasoning, least-squares development, and interpretations are retained/adapted. Slope and correlation significance tests from Section 12.4 are excluded from the revised AP core. Contextual interpretations use predicted change and do not imply causation. https://openstax.org/details/books/introductory-statistics-2e

S10.04 | Residuals and checking a linear model

CED 5.4

- Calculate and interpret a signed residual.
- Construct a residual plot.
- Use residual patterns to assess linear-model adequacy.

CONCEPT AND INTERPRETATION

A residual measures observed minus predicted

The residual is $e = y - \hat{y}$. It measures the vertical discrepancy between an observation and the fitted prediction at the same x value. A positive residual means the actual response is above the predicted response, so the model underpredicts. A negative residual means the actual response is below the prediction, so the model overpredicts. The residual has the response variable's units. Keep the subtraction order fixed. "The residual is 3" is incomplete without the units and what the sign says about the prediction.

Construct a new plot of model errors

A residual plot places x values, or fitted response values, on the horizontal axis and residuals on the vertical axis. Include a horizontal reference at zero. Each point still represents one observational unit, but the vertical coordinate is now its prediction discrepancy. Removing the overall fitted trend makes systematic departures easier to see. Compute residuals using unrounded regression coefficients when possible, then round for reporting. Rounding the line too early can introduce artificial discrepancies or make residual sums appear inconsistent.

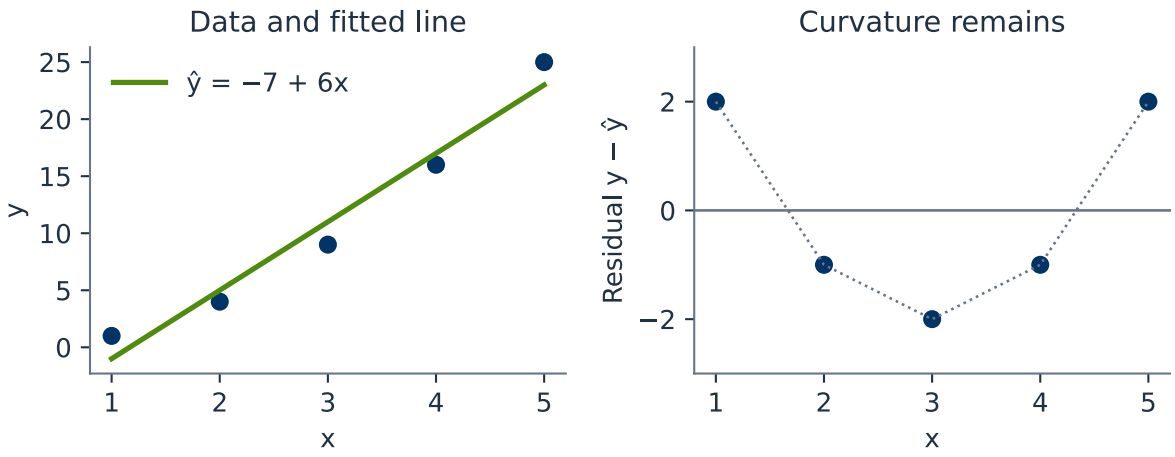
Random-looking scatter supports linear form

If residuals appear scattered around zero with no systematic curve, the linear form is reasonable for the observed range. This is evidence about form, not proof that the model is perfect or that its predictions are precise. The vertical scale matters: a random-looking residual pattern can still contain large errors in practical units. Describe the plot and connect it to the original model. A statement such as "the residual plot is linear" is misleading; an appropriate residual plot should not retain a systematic trend that the model failed to explain.

Patterns reveal limitations

A U-shaped or arch-shaped residual pattern suggests curvature remaining after the straight line is fitted. A linear model is then not the most appropriate description of the observed association. A widening spread suggests that prediction errors vary across the input range. Isolated large residuals identify observations poorly predicted by the line. These findings should prompt investigation of the data, context, and model. The required course focus is recognizing when a linear model is inadequate; transforming variables or fitting nonlinear models is not needed to state that conclusion.

Residuals expose what the line misses



Exact investigation data: y = 1, 4, 9, 16, 25 at x = 1, 2, 3, 4, 5. The residuals are 2, −1, −2, −1, 2. Their U-shape shows the linear form is inadequate.

Residuals do not answer every design question

10

A clean residual plot cannot reveal random sampling if none occurred, nor can it establish a causal effect. It checks a particular feature of the prediction model against the observed pairs. A large residual does not automatically justify deleting a valid observation; it may represent a meaningful exception. Also distinguish a large residual from an unusual x value. A point can be far out in x but close to the line, or near the center of x but poorly predicted. Context is needed to decide what those cases mean.

KEY IDEA

Residuals show what a fitted line leaves unexplained at each observed input.

WATCH FOR

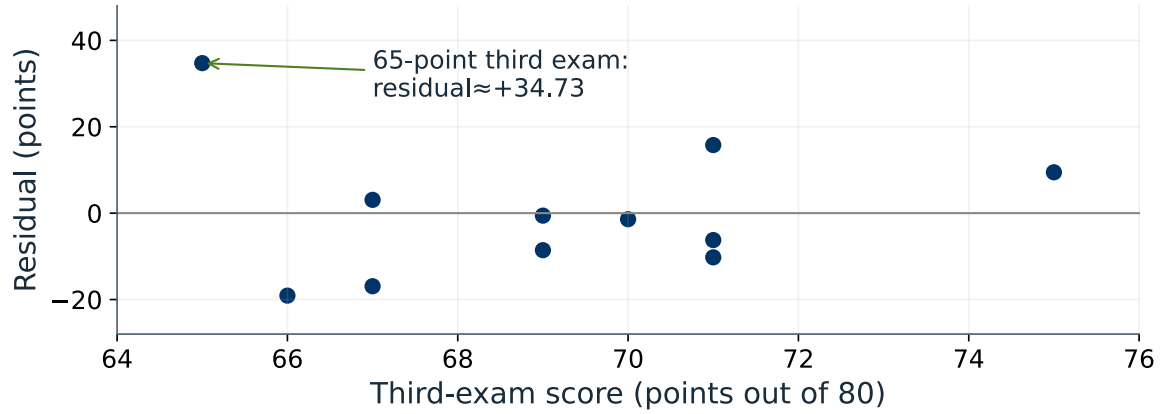
Positive residual means underprediction; negative residual means overprediction.

OPENSTAX WORKED EXAMPLE

What does the regression line miss?

The OpenStax third-exam/final-exam pairs are (65,175),(67,133),(71,185),(71,163),(66,126),(75,198),(67,153),(70,163),(71,159),(69,151),(69,159). The explanatory score x is out of 80 points and the response score y is out of 200 points. For the student with x=65 and y=175, calculate and interpret the residual using $\hat{y} = -173.51336 + 4.827394x$. Explain what a full residual plot would add.

Residuals for all eleven source observations



Each residual equals observed final-exam score minus the fitted prediction. The high positive residual at x=65 is the worked student's prediction discrepancy.

1. The line predicts −173.51336+4.827394(65)=140.26725 final-exam points.

2. The residual is observed−predicted=175−140.26725=34.73275 points. The positive sign places this observed point above the fitted line.

3. This student's final score was about 34.73 points higher than predicted from the third-exam score. Equivalently, the model underpredicted this student's final score by 34.73 points.

4. A residual plot repeats this calculation for every student, puts the third-exam score on the horizontal axis, and plots residual points around a horizontal zero line. Look for systematic curvature and unusually large residuals.

5. One large residual does not describe the entire model. A visible curve would indicate a systematic mismatch in linear form; large residual magnitudes indicate imprecise individual predictions even if a straight-line form is reasonable.

Conclusion. A residual is a signed prediction discrepancy measured in response units. It is not a new exam score and need not indicate a data-entry mistake.

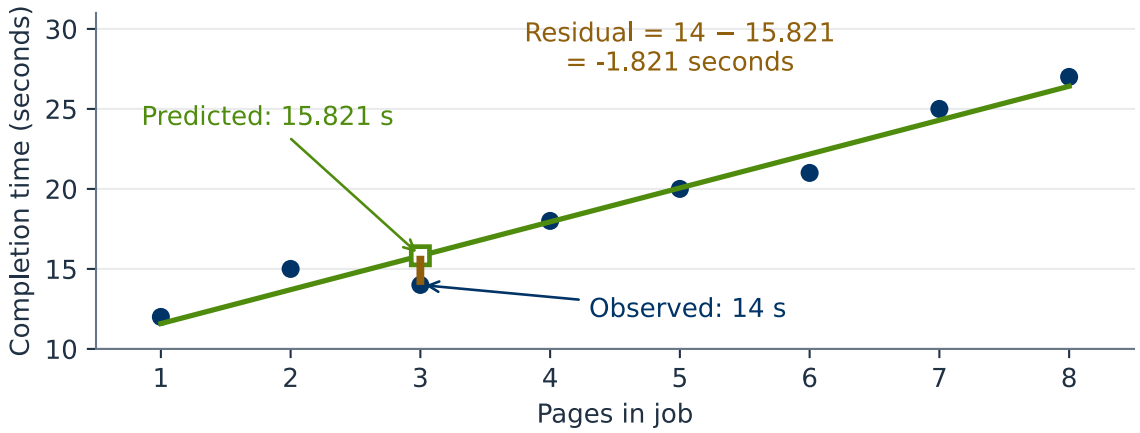
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 12.3, Example 12.6 and residual development, pp.623-627 (PDF 635-639). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

The 3-page job and its prediction

The illustrative print-job model is predicted seconds=9.4642857+2.1190476 x. A 3-page job actually takes 14 seconds. Calculate and interpret its residual, then explain how all eight residuals help assess the line.

A signed vertical prediction discrepancy



Worked 3-page job: observed time 14 seconds; predicted time 15.821 seconds. The gold segment represents a residual of −1.821 seconds, so the model overpredicts.

1. At x=3, the predicted completion time is 9.4642857+2.1190476(3)=15.8214285 seconds.

2. The residual is observed−predicted=14−15.8214285=−1.8214285 seconds. The sign is negative because the job finished faster than predicted.

3. In context, the line overpredicts this job's completion time by about 1.821 seconds. Equivalently, the observed time is 1.821 seconds below the prediction. Avoid calling the negative residual a negative completion time.

4. For the eight jobs, residuals rounded to 3 decimals are 0.417,1.298,−1.821,0.060,−0.060,−1.179,0.702,0.583. Plot these against page counts 1 through 8 and include a zero line.

5. There is no clear systematic curve in this small illustrative residual plot. A linear form is reasonable for these data, but the residuals show that individual times still differ from predictions.

Conclusion. The 3-page print job finished in 14 seconds, about 1.82 seconds sooner than predicted. Across all eight jobs the residual plot shows no clear curve. Individual times still vary around the fitted line.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 12.2-12.3, pp.620-631 (PDF 632-643), Try It 12.5 and Example 12.6. Source data, scatterplot reasoning, least-squares development, and interpretations are retained/adapted. Slope and correlation significance tests from Section 12.4 are excluded from the revised AP core. Contextual interpretations use predicted change and do not imply causation. https://openstax.org/details/books/introductory-statistics-2e

S10.05 | Least squares and explained variation

CED 5.5, 5.4, 5.2

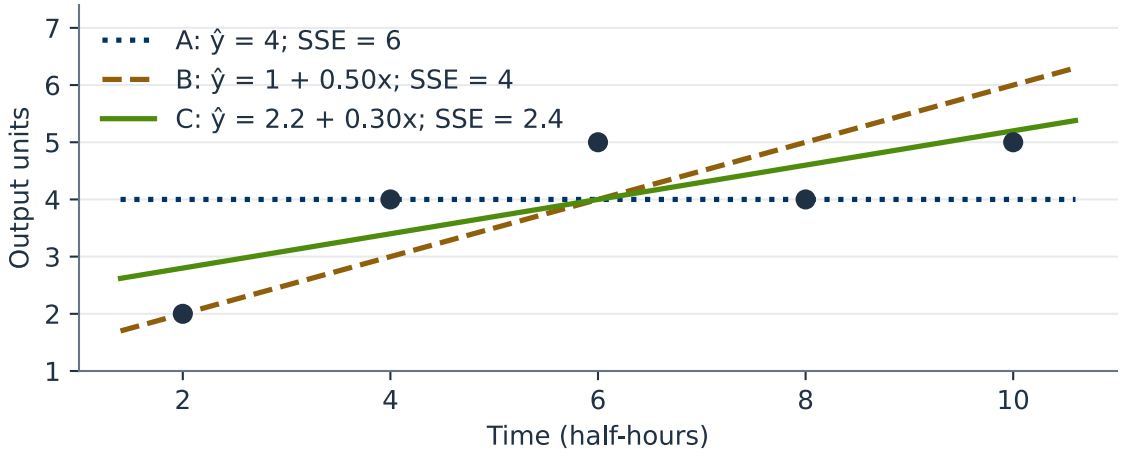
- Use technology to fit a least-squares regression line.
- Explain the squared-residual criterion and the mean-point property.
- Interpret the coefficient of determination in context.

CONCEPT AND INTERPRETATION

Why least squares uses squared residuals

A fitted line needs a criterion for deciding which prediction errors are preferable. Least squares chooses the line that minimizes the sum of squared residuals, $SSE = \sum(y - \hat{y})^2$. Squaring prevents positive and negative discrepancies from canceling and gives large errors extra weight. The best line under this criterion need not pass through the first and last observed points. Use technology to calculate its coefficients, then inspect the scatterplot and residuals to decide whether linear form is reasonable. Optimization of a line cannot make a curved relationship linear.

Least squares compares all residuals



Time is measured in half-hours. The five pairs are (2,2), (4,4), (6,5), (8,4), and (10,5). Candidate C has the smallest sum of squared residuals.

Read the fitted model as a statistical summary

Technology returns an intercept a and slope b for $\hat{y} = a + bx$, and often correlation r and its square. Confirm which variable was supplied as explanatory and which as response. A least-squares line with an intercept passes through $(\bar{x}, \bar{y})$. This property is a useful check on transcription and units. Interpret slope as predicted response change per unit of x and interpret the intercept only where zero has contextual meaning. A fitted coefficient is a sample summary, not a guaranteed law for every individual.

Compare the model with a mean-only prediction

Before using x, a simple prediction for every response is the overall mean $\bar{y}$. The total squared variation around that mean is $SST = \sum(y - \bar{y})^2$. The fitted line leaves SSE unexplained. For ordinary simple least-squares regression with an intercept, the fraction of total variation explained by the linear relationship is $r^2 = 1 - SSE/SST$, when SST is positive. This explains why a line is judged against a baseline, not simply by whether some predictions are close to observations.

Interpret the coefficient of determination precisely

The coefficient of determination r^2 is the proportion of observed variation in the response variable explained by its linear relationship with the explanatory variable. Name both variables. If $r^2 = 0.60$, then 60% of response variation about its mean is explained by the fitted linear relationship with x. It does not mean 60% of observations lie on the line, 60% of predictions are correct, or 60% of y is caused by x. Squaring removes direction: both positive and negative correlations of the same magnitude have the same r^2.

Fit quality and responsible use

A high r^2 can coexist with curvature or a very influential observation, so it does not replace graphical checks. A low r^2 indicates much variation remains unexplained by this linear model, even if the estimated slope has a meaningful direction. As a supporting extension, the residual standard deviation $s = \sqrt{SSE/(n-2)}$ summarizes residual size in response units when n>2; slope inference is not part of this course's revised required regression scope. Assess predictions using both the residual pattern and practical error scale, and retain the study-design limits on causal interpretation.

10

KEY IDEA

Least squares minimizes squared prediction errors; r^2 compares the line with predicting every response by its mean.

WATCH FOR

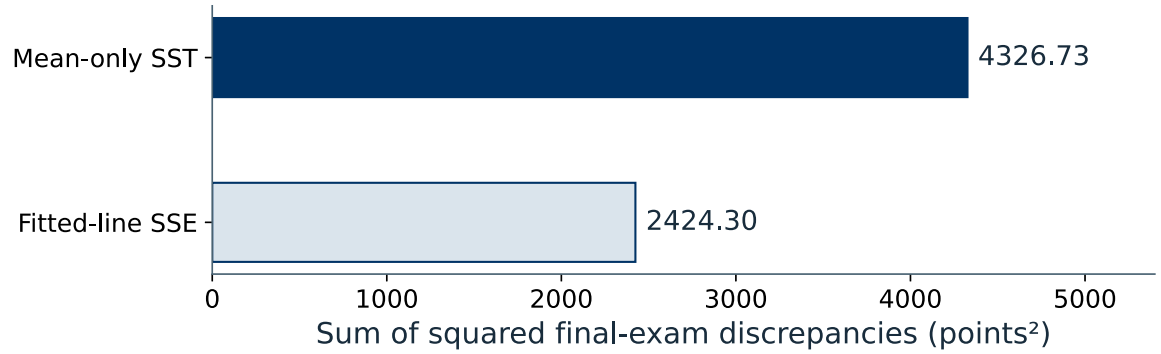
Explained variation is a statistical description, not a percentage of outcomes caused or predicted correctly.

OPENSTAX WORKED EXAMPLE

How much exam-score variation is explained?

The OpenStax third-exam/final-exam pairs are (65,175),(67,133),(71,185),(71,163),(66,126),(75,198),(67,153),(70,163),(71,159),(69,151),(69,159). The explanatory score x is out of 80 points and the response score y is out of 200 points. Technology gives $\hat{y} = -173.513363 + 4.827394x$ and r≈0.663094. Calculate and interpret the coefficient of determination and verify the mean-point property.

Compare the line with predicting the mean



The line reduces squared discrepancy from 4326.73 to 2424.30. The reduction is about 43.97% of the total observed squared variation.

1. The coefficient of determination is $r^2 \approx 0.439693$, so the linear relationship with third-exam score explains about 43.97% of the observed variation in final-exam score about its mean.

2. The sample means are $\bar{x} = 69.181818$ and $\bar{y} = 160.454545$. Substituting the first mean into the fitted line gives 160.454545, confirming that the least-squares line passes through the mean point.

3. For these observations, SST=4326.727273 squared points and SSE=2424.295100 squared points. Then $1 - SSE/SST \approx 0.439693$, the same explained fraction.

4. The remaining 56.03% is the observed variation left unexplained by this fitted linear relationship. It is not a count of students whose predictions are wrong, and it does not measure a causal share.

Conclusion. Least squares improves on predicting every final score by the sample mean. The explained percentage measures this reduction in squared variation.

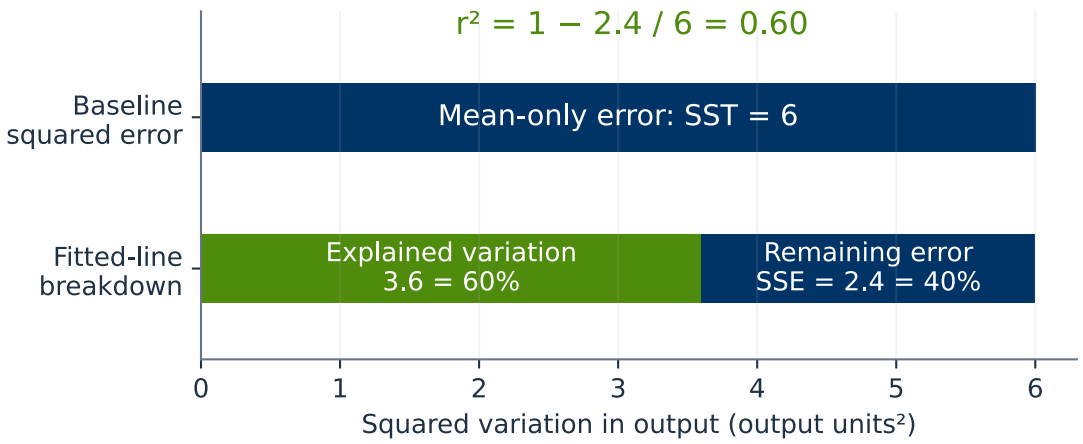
Source: Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Section 12.3, Example 12.6 and coefficient-of-determination discussion, pp.623-631 (PDF 635-643). Data retained; the explanation and any added assumptions are identified in the lesson source notes. https://openstax.org/details/books/introductory-statistics-2e

LOCAL APPLICATION

How much variation does the line explain?

For five jobs let x = (2, 4, 6, 8, 10) be the number of half-hours worked, and let y = (2, 4, 5, 4, 5) be output units. Technology gives predicted y = 2.2 + 0.30x and r ≈ 0.774597. Interpret the slope and the coefficient of determination.

Total variation split by the fitted line



Both bars represent the same total variation, SST = 6. Predicting the mean output of 4 gives squared error 6. The fitted line leaves squared error 2.4 and accounts for 3.6, or 60%, of response variation about its mean.

1. The sample means are 6 half-hours and 4 output units. At x = 6, the line predicts 2.2 + 0.30(6) = 4, so it passes through the mean point.

2. The predicted values are 2.8, 3.4, 4.0, 4.6, and 5.2. The residuals are −0.8, 0.6, 1.0, −0.6, and −0.2. Their squared sum is SSE = 2.4 in squared output units.

3. The mean-only predictions are all 4. The squared deviations from 4 are 4,0,1,0,1, so SST=6. The explained fraction is 1−2.4/6=0.60, which agrees with r^2.

4. The linear relationship with time explains 60% of the observed variation in output about its mean. This fraction is unitless.

5. The slope is 0.30 output units per additional half-hour. The explained fraction is 0.60. The residual plot and the observed range of 2 to 10 half-hours guide model use.

Conclusion. The line reduces squared prediction error from 6 to 2.4 relative to the mean-only baseline, explaining 60% of the observed response variation.

Sources and adaptations

Barbara Illowsky and Susan Dean, OpenStax, Introductory Statistics 2e, supplied 2026 printing. Sections 12.2-12.3, pp.620-631 (PDF 632-643), Try It 12.5 and Example 12.6. Source data, scatterplot reasoning, least-squares development, and interpretations are retained/adapted. Slope and correlation significance tests from Section 12.4 are excluded from the revised AP core. Contextual interpretations use predicted change and do not imply causation. https://openstax.org/details/books/introductory-statistics-2e

Unit synthesis

1. Describe the four features of a scatterplot that should be considered.

2. Can r=0 coexist with a strong relationship?

3. For predicted y=4+3 x, x=2, and observed y=12, find prediction and residual.

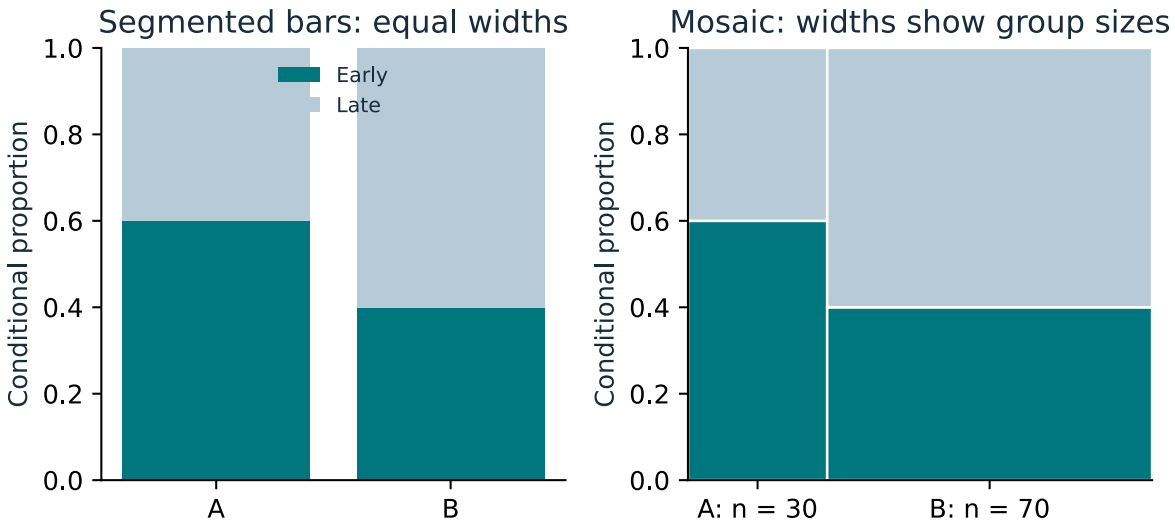
4. What does a U-shaped residual plot indicate?

5. A model has SSE=15 and SST=60. Find and interpret r^2.

6. What does a high r^2 not establish?

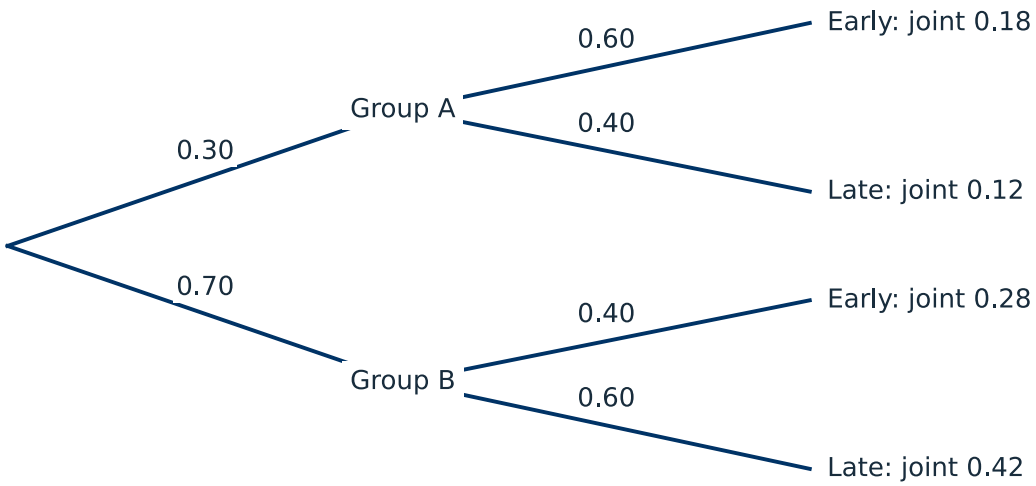
Visual reference

Conditional proportions and group size



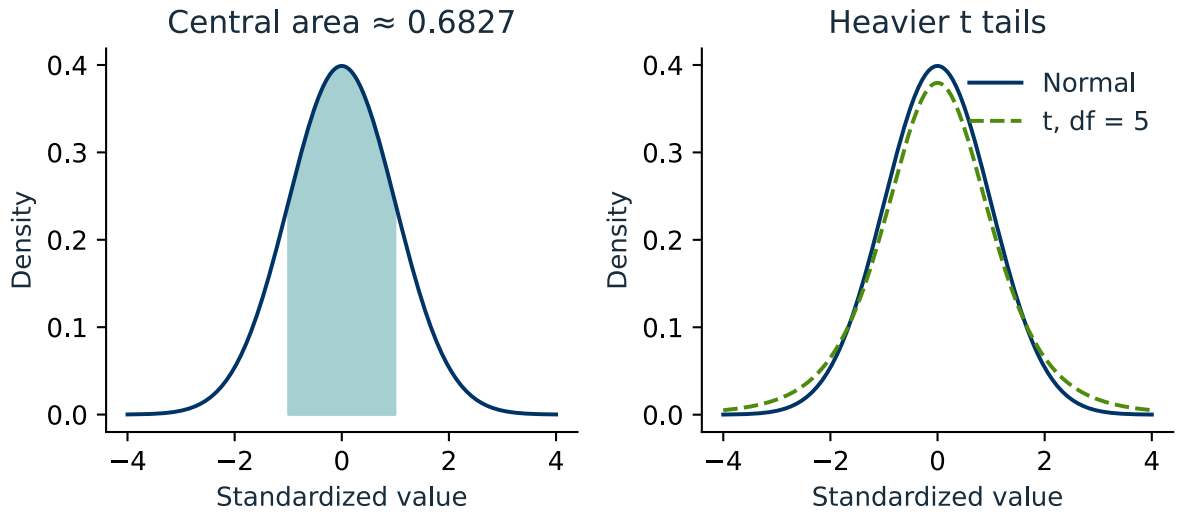
Illustrative table: group A has 18 early and 12 late responses; group B has 28 early and 42 late responses. Early response is 60% in A and 40% in B. Segmented bars compare conditional proportions using equal widths. Mosaic widths show that A supplies 30% of all observations and B supplies 70%; each rectangle's area represents its joint proportion.

A tree records changing conditions



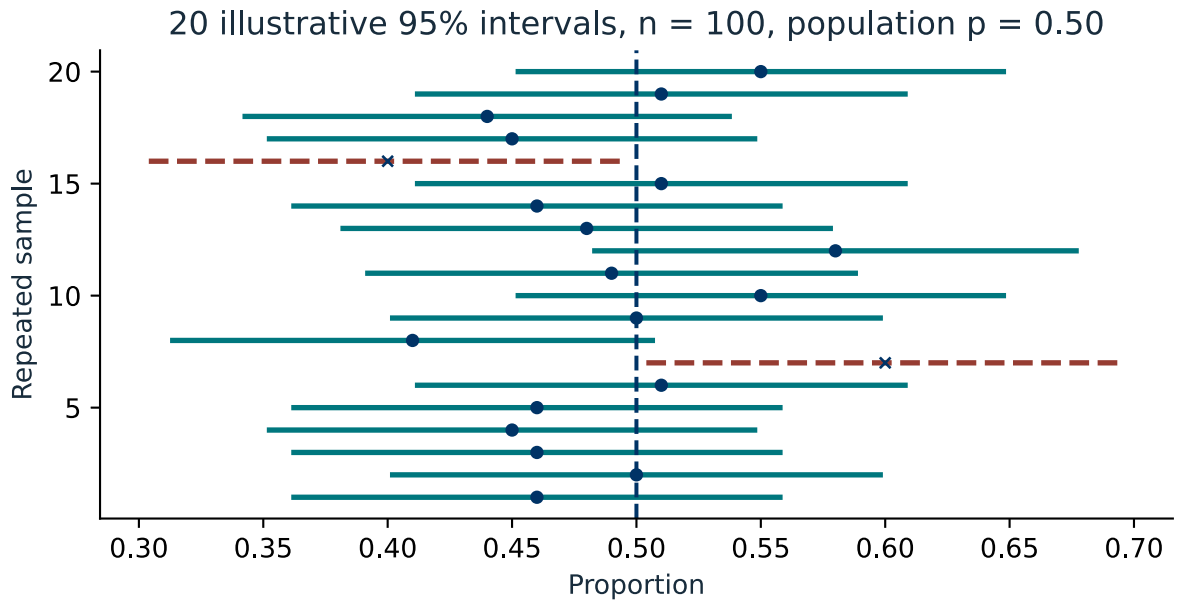
Use the groups in the display above. Let A mean membership in Group A and E mean an early response. P(A) = 0.30, P(E given A) = 0.60, and P(E given Group B) = 0.40. Multiply along a path for a joint probability. Add the early-response paths: P(E) = 0.18 + 0.28 = 0.46. Reversing the conditioning gives P(A given E) = 0.18/0.46 ≈ 0.3913. Different conditional early-response probabilities show that group and response are associated.

Normal and t models



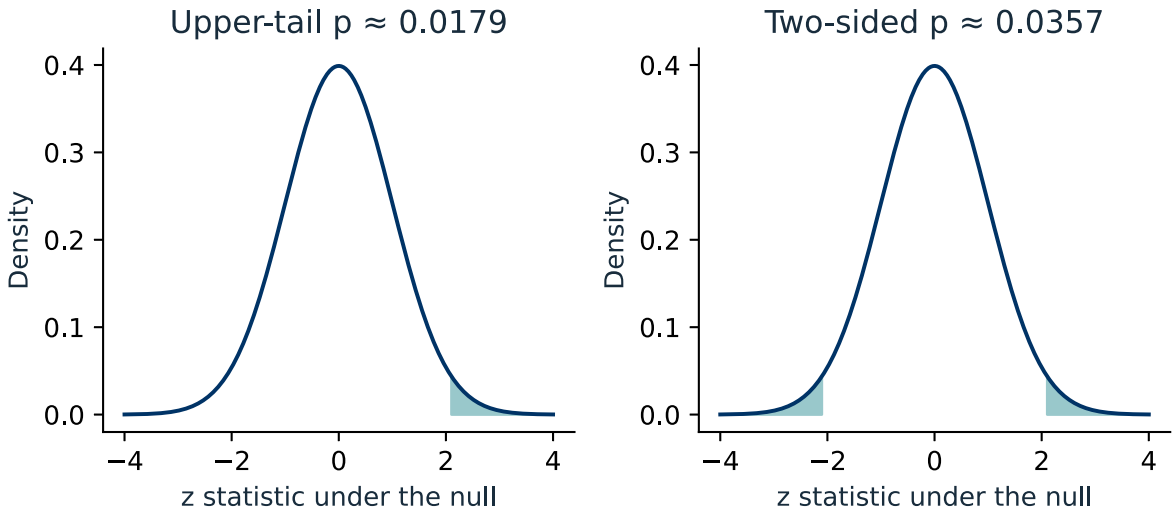
The shaded standard normal interval is within one standard deviation of the mean. A t distribution is symmetric about zero but has heavier tails; the degrees of freedom govern the difference from the normal model. Density height is not itself a probability.

Confidence is a property of a method



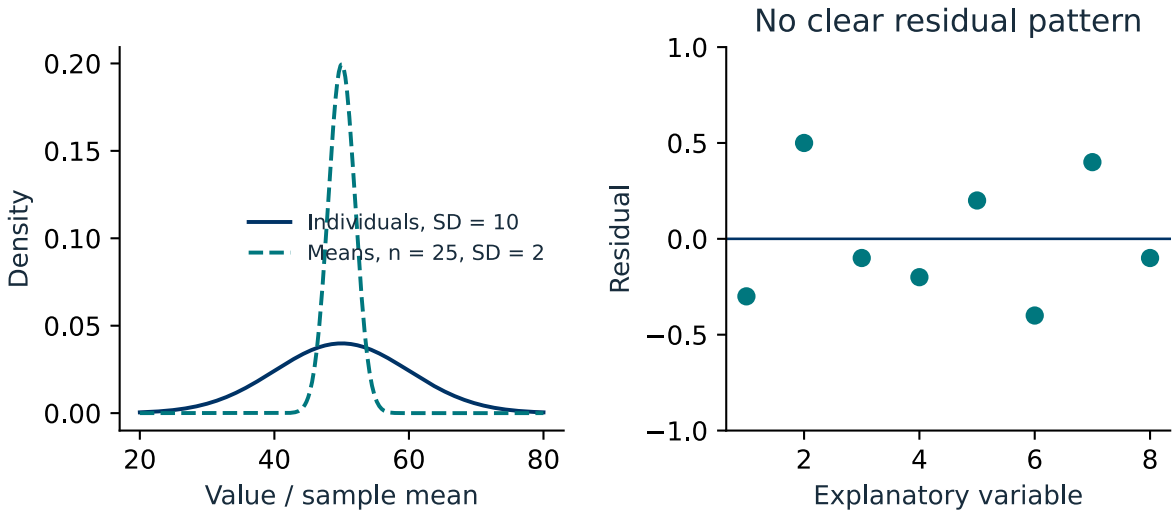
The vertical line marks the fixed population proportion, 0.50. Eighteen of these twenty simulated intervals cross it. The two dashed intervals, samples 7 and 16, miss it. The 95% level describes long-run performance of the procedure; the count can vary from batch to batch.

The alternative determines the tail



Both panels use z = 2.10 and the same null model. An upper-tail alternative counts results at least as large. A two-sided alternative counts results at least as far from zero in either direction. Choose the alternative from the question before examining the result.

Identify the distribution you are describing



Left: when individual values are normal with mean 50 and SD 10, means of independent samples of size 25 are normal with the same mean and SD 2. Right: a residual plot checks whether a straight line is a reasonable description; apparent random scatter supports linear form but does not establish causation.

Choosing an inference method

Use the SMS 5C routine. **Classify** the variables and study design. **Choose** a procedure and define the parameter. **Check** its conditions with evidence from the study. **Calculate** the interval or test result. **Conclude** in the context of the investigative question. The table helps with the first two steps.

Data and goal	Interval	Test
One binary response / one population	One-proportion z interval	One-proportion z test
Binary response / two independent groups	Two-proportion z interval, separate estimated SE	Two-proportion z test, pooled null SE for equal proportions
One quantitative sample	One-sample t interval	One-sample t test
Paired quantitative observations	One-sample t interval on paired differences	One-sample t test on paired differences
Two independent quantitative groups	Two-sample t interval	Two-sample t test
Two categorical variables in a count table	Not a chi-square interval procedure	Chi-square independence or homogeneity; design determines the name

CONDITIONS ARE REASONS

Random selection or assignment addresses the design. The 10% condition addresses dependence from sampling without replacement. Success/failure counts or the data-shape condition address the sampling model. Naming a condition without connecting it to the data is incomplete.

For proportions

A one-proportion interval uses observed success and failure counts, each at least 10. A one-proportion test checks expected counts under the null. Two-proportion intervals check observed successes and failures in each group; an equal-proportions test uses pooled expected counts in each group. A chi-square table uses every expected cell count, each greater than 5 under the supplied Fall 2026 CED.

For means

Use the t model when the population SD is unknown. For a small sample, assess strong skewness and outliers in the relevant data. For paired data, those data are the differences. For independent groups, check each group. A large sample reduces concern about ordinary nonnormality but does not repair bias, dependence, or a poorly defined target population.

Responding to AP questions

Read the command verb before choosing a calculation or writing a paragraph. A correct number does not answer a request for an interpretation. A list of facts does not answer a request to compare. Give the evidence asked for, and make the meaning of your answer clear.

Task word	What your response should supply
Calculate	A numerical result with enough supporting work to show the method. Include units where the quantity has units.
Compare	An explicit similarity or difference between the named groups or quantities. Use numerical evidence when available.
Describe	The relevant features of the data or display. For a distribution, address the features requested and name the measured variable.
Explain / Justify	A reason connected to the facts of this study. Show why the evidence supports or limits the claim.
Identify / Classify	The requested name, category, or feature. Add a reason when the prompt also requests one.
Interpret	The meaning of the number, interval, graph, or result for the people or objects in the problem.
Complete an inference procedure	Identify the method and parameter, verify the relevant conditions, calculate, and give a contextual conclusion.
Construct	The requested graph or table, with the labels, scale, and entries needed to understand it.

A conclusion that stands on its own

For an estimate, name the population quantity and give the estimate with its units. For a confidence interval, add the confidence level and interpret the interval as estimating a population parameter. For a test, state whether the evidence supports the alternative claim, in context. Respect the scope allowed by the sampling or assignment method.

Example: “We are 95% confident that the mean weekday wait for all riders on this route is between 6.2 and 8.1 minutes.” The sentence identifies the population, variable, parameter, interval, units, and confidence level. Its validity still depends on the study design and method conditions.

Adapted guidance based on College Board, AP Statistics Course and Exam Description, effective Fall 2026, task verbs, printed p.159 (PDF p.164). https://apcentral.collegeboard.org/media/pdf/ap-statistics-course-and-exam-description.pdf

Using technology and writing digitally

Technology performs arithmetic and model calculations. Your written response explains what the calculation means. Record the inputs and essential outputs so another person can reproduce your result without seeing a screenshot.

Task	Inputs to preserve	Evidence to report
One-variable statistics	List of values; frequency list if used	n, mean, sample SD, relevant quantiles; labeled graph
Normal probability / percentile	Mean, SD, interval bounds or left-tail area	Model statement, probability or cutoff with units
Binomial probability	n, success probability, exact count or cumulative bound	Defined random variable, event, probability
Interval	Counts or summary statistics; confidence level; paired/independent choice	Method, conditions, interval, contextual confidence interpretation
Significance test	Null value, tail, counts or summary statistics; pairing	Hypotheses, conditions, statistic, df when relevant, p-value, conclusion
Chi-square table	Observed count matrix	Expected counts, df, statistic, upper-tail p-value
Regression	Matched x and y lists; correct explanatory variable	Model, r or r^2, contextual slope, prediction/residual and model check

Typed notation

In a plain-text response, write p-hat, x-bar, mu, sigma, and p1 - p2 when a symbol editor is inconvenient. Define each symbol in words. For a simulation, specify the model, one trial, the statistic, the repetitions, and how the event proportion estimates the requested probability. For a test, distinguish the p-value from the population proportion p.

Precision and sense checks

Keep extra digits in intermediate calculations. Round final probabilities or p-values consistently, usually to four decimals unless the prompt requests another precision; use $p < 0.0001$ when appropriate rather than $p = 0$. Report interval endpoints with sensible units. Check that probabilities lie between 0 and 1 and that an interval or residual has the correct units and sign.

Graphing calculator and Chromebook routines

Use the course calculator or statistical software to enter data, select the correct model, and interpret output. Menu locations and software features may change, so these books specify required inputs and outputs rather than an unverified button sequence. Practice typing complete answers periodically; follow current College Board rules and the approved reference booklet for exam simulation.

Notation and vocabulary

Symbol / term	Meaning
Population / sample	The group of interest / the observed subset
Parameter / statistic	A population quantity / a quantity calculated from sample data
N / n	Population size / sample size
p / p-hat	Population proportion / sample proportion
mu / x-bar	Population mean / sample mean
sigma / s	Population SD / sample SD
SE	Standard deviation of a statistic across repeated samples, usually estimated from the data
mu-d / d-bar	Population mean paired difference / sample mean paired difference
Bias	A systematic tendency of a method to miss the target
Sampling variability	How a statistic changes across repeated samples
Confidence level	The long-run capture rate of an interval method under its conditions
p-value	The probability under the null model of a result at least as extreme as the observed result in the direction(s) of the alternative
Power	The probability of rejecting a false null for a specified alternative value
r / r^2	Correlation / fraction of variation in the observed y values explained by their least-squares line with x
Residual	Observed response minus predicted response