ST. MARY'S SCHOOL

Medford, Oregon

AP Statistics

Preparation Readings



Wizard Island, Crater Lake, Oregon

Before each class, read the assigned example and answer the two questions that follow. Bring your work to the next meeting.

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

How to prepare for class

Find tomorrow's meeting

Use the PR number your teacher assigns. For example, after meeting 1, complete the practice listed at the start of PR02. Then read PR02 and answer its two questions before meeting 2. Start PR01 in class with your teacher.

Complete the assigned practice

Use the listed workbook questions on a lesson night. The remaining problems are available for extra practice. On project, checkpoint, and review nights, the practice directions appear on the next reading page.

Work through the example

Before reading a calculation or conclusion, try the next step yourself. Compare your reasoning with the example. If a step is unclear, mark it so you can ask about it in class.

Answer the two checks

Write your answers on the reading page, in your notebook, or in your class document. Include the units for a calculation and identify the group or quantity you are describing. Bring your answers to class for your own use. The entry ticket draws on the previous lesson and this reading.

Budget the evening

Plan roughly 30-40 minutes for lesson practice and 12-15 minutes for the reading and both answers together. The two timed mock sections have lighter follow-up nights.

SOURCES

The primary instructional source is Introductory Statistics 2e by Barbara Illowsky and Susan Dean, published by OpenStax / Rice University. The credited readings adapt OpenStax concepts and examples to this AP course. The short checks and paired tickets test the assigned reading and the preceding lesson. The original textbook is available free at OpenStax.org.

https://openstax.org/details/books/introductory-statistics-2e

Course adaptation distributed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International. AP is a College Board trademark.

PR01 | Questions, Variables, and Categorical Data

What can a bicycle count tell us?

OPENING MEETING

Begin here in class. Use this reading and its checks during the orientation. Your teacher will explain the new ideas before guided practice.

READ FOR THE COMING MEETING

A statistical question concerns a group and anticipates variation in the observations. Descriptive statistics organize and summarize the observations collected. Statistical inference uses sample information to reason about a wider population, while recognizing uncertainty. Before calculating, identify the individuals or objects being studied, the variable recorded for each, and the population to which the question refers. A population is the entire group of interest; a sample is the subset actually observed. A parameter summarizes a population, whereas a statistic summarizes a sample.

OpenStax asks about the mean annual amount spent on school uniforms by families at a school. In its example, 100 families are randomly surveyed. One family is the observational unit. Annual uniform spending is a quantitative variable measured in dollars. The mean for all school families is the parameter, and the mean for the 100 surveyed families is a statistic. The three reported amounts, $65, $75, and $95, are individual observations, not three different variables. We cannot find the sample mean from these three amounts when 100 families were surveyed.

The same distinctions apply to categorical data. A categorical variable places an observation in a group, such as whether a bicycle has a working light. A quantitative variable records a numerical amount or count for which arithmetic has a meaningful interpretation. Writing a category as a number, such as a locker number, does not make the variable quantitative. The following bicycle example changes the variable from spending to light condition. Read for the population, sample, variable, and sample proportion.

Consider two questions a school transportation club could ask. “Does this bicycle have a working front light?” requires a single inspection. “What proportion of bicycles parked at school on Friday morning have working front lights?” anticipates different results across bicycles and calls for a summary. To investigate the second question, the club lists all 90 bicycles parked that morning, selects 30 at random, and checks each selected bicycle once. Each bicycle is an observational unit.

PR01 | Continued

Each row of the club's record identifies one bicycle. The light entry has two possible values, working or not working. These values describe categories. A rack number also identifies a category, even if the club writes it as a number. By contrast, the time in seconds required to test the light measures a quantity. A difference of five seconds has a physical meaning.

Suppose 21 of the 30 checked bicycles have working lights. A frequency table would have two rows, one for each light condition. Their frequencies are 21 and 9. For relative frequencies, divide each count by the number checked. The fractions 21/30 and 9/30 account for all 30 bicycles and add to 1. In a bar chart of these results, label the horizontal axis with the two conditions and the vertical axis with either counts or percentages. The club has observed the condition of every bicycle in its sample. The 60 bicycles it did not inspect could have a different mix of working and nonworking lights.

The bicycle study: keep the groups distinct

Quantity	Bicycles
Population listed	90
Sample checked	30
Working lights in the sample	21

Source and adaptation: OpenStax, Introductory Statistics 2e, §§1.1-1.3, Try It 1.1 (school uniforms) and the section on frequency tables. The bicycle study is an SMS example.

TWO READING CHECKS

1. What does one row in the light record represent, and what two values could the light entry take?

2. Why does 21 divided by 30 describe a statistic?

PR02 | Two Categorical Variables and Association

Which visitors used the trail map?

PRACTICE FROM THE LAST MEETING

S01.01-HW (W011). Complete 1, 2, 3, 5, 6.

READ FOR THE COMING MEETING

A two-way table records two categorical variables for the same individuals. Each individual belongs in one cell. Row and column totals are marginal counts, and the sum of all cells is the overall sample size. A joint relative frequency divides a cell count by the overall total. A conditional relative frequency divides by the total for a specified group. The denominator is part of the meaning of a percentage; a larger count alone does not show a higher rate.

In OpenStax's table of 100 hikers, 18 of the 45 women and 16 of the 55 men prefer the coastline. The overall coastline count is 34. Thus 18% of all sampled hikers are women who prefer the coastline, while 18/45=40% of sampled women prefer it. The corresponding percentage among sampled men is 16/55, about 29.1%. These percentages answer different questions from 18/100 or 16/100. Comparing conditional distributions, using each group's own total, is the appropriate way to compare preferences.

The hikers also illustrate a marginal relative frequency. Ignore the preference categories and look at the row total for women. Women account for 45/100 of the entire sample. Now ignore sex and read the coastline column total. It gives 34/100 of the sample. A joint percentage uses a cell, a marginal percentage uses a row or column total, and a conditional percentage starts by restricting the group. Naming the group first keeps these three questions distinct.

At a fictional trail orientation, a volunteer records each visitor's arrival session and whether the visitor picks up a printed map. Ten people attend the morning session and six take maps. Thirty attend the afternoon session and twelve take maps. To compare the sessions, write the four counts in a table. The morning row has six map users and four others. The afternoon row has twelve map users and eighteen others.

PR02 | Continued

Someone argues that the afternoon visitors were more inclined to take maps because they took twice as many. Count the opportunities, too. Six of ten morning visitors took a map. The afternoon had three times as many visitors as the morning. If the two sessions had the same map-taking rate, 18 of those 30 afternoon visitors would have taken a map. The actual afternoon count was 12. Compute the afternoon rate for the first reading check.

Two categorical variables are associated in a data set when the distribution of one changes across groups defined by the other. To show the local comparison, make a segmented bar for each session, with both bars running from 0% to 100%. Use the same order for “took” and “skipped” in each bar. In a mosaic plot, the afternoon group would also be three times as wide because it accounts for three times as many visitors. In both displays, the height of the “took” segment shows the rate within that session. These visitors chose when to attend. The difference in rates alone cannot establish that attending later changed anyone’s preference.

Printed maps at the trail orientation

Session	Took a map	Skipped the map
Morning	6	4
Afternoon	12	18

Source and adaptation: OpenStax, Introductory Statistics 2e, §§1.3 and 3.4, especially Example 3.21 (hiking preferences). The orientation and map example is an SMS example.

TWO READING CHECKS

1. What fraction of afternoon visitors took a map? Identify the group represented by the denominator.

2. What fraction of all visitors attended in the morning and took a map?

PR03 | Seeing a Quantitative Distribution

What disappears when times are grouped?

PRACTICE FROM THE LAST MEETING

S01.02-HW (W017). Complete 1, 2, 3, 4, 5.

READ FOR THE COMING MEETING

A distribution describes the values a quantitative variable takes and how frequently they occur. Before describing it, identify the variable and its units. A dotplot places one mark for every observation on a numerical scale. A stemplot also preserves individual observations, provided its key states what each stem and leaf mean. A histogram groups values into nonoverlapping numerical intervals. These displays summarize one quantitative variable; a bar chart instead compares categories.

OpenStax's supermarket-distance example contains 21 distances in kilometers. Many values are between 3 and 5 kilometers, while one is 12.3 kilometers. A stemplot key such as 3|2=3.2 km makes the units and decimal placement clear. Keeping the empty stems between 6 and 12 shows the gap before the distant supermarket. Removing those empty positions would visually compress a genuine gap. The observation at 12.3 km is worth investigating as a possible outlier, but a graph alone does not tell us that it is a recording error.

When asked to describe a quantitative distribution, address its shape, a typical value or center, spread, and unusual features in context. Use values and units where available. A long right tail means unusually large observations extend toward the right; it does not mean most observations are on the right. With a histogram, recognize the information lost within each bin. The next example lets you compare what a dotplot preserves with what a histogram groups together.

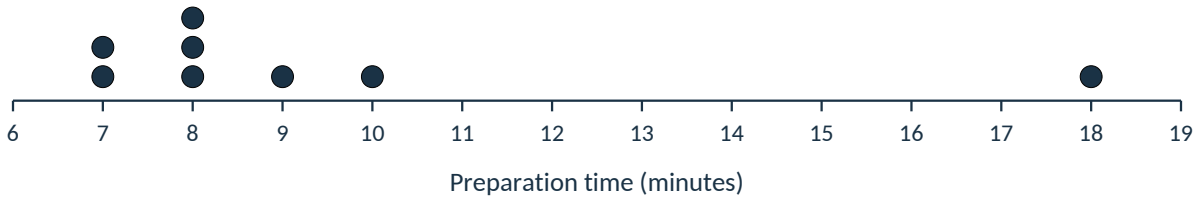
Suppose eight trays in a school kitchen take 7, 7, 8, 8, 8, 9, 10, and 18 minutes to prepare. To make a dotplot, number a horizontal line in equal one-minute steps. Place one dot above each observed time and stack repeats. For a stemplot, use tens digits for stems and ones digits for leaves. With the key 1 | 8 = 18 minutes, the two observations of 7 become two leaves of 7 on the 0 stem. Either display lets a reader recover the original times.

A histogram combines values into intervals called bins. With bins that start at 6, 10, 14, and 18 minutes and have width four, each time belongs to exactly one bin. Use a rule that includes the lower end and excludes the upper end. Then a ten-minute tray belongs to the bin from ten up to fourteen.

PR03 | Continued

Changing the bins can change the picture. The 6-to-under -10 bar counts six trays. Split it at 8 minutes and you get two trays from 6 up to 8 and four from 8 up to 10. A reader of the original histogram could recover neither of those counts from the single bar. A useful description of the raw times would say that seven of eight trays took 7 to 10 minutes, the median was 8 minutes, and one tray took 18 minutes. That distant observation gives the distribution a right tail. A sentence like "the times were uneven" would leave the reader guessing how much and in which direction.

Preparation times for eight trays



Source and adaptation: OpenStax, Introductory Statistics 2e, §§2.1-2.2, especially Example 2.2 (distances to supermarkets). The kitchen times are an SMS example.

TWO READING CHECKS

1. At the horizontal-axis value 8 minutes, how many dots are stacked? What does one dot represent?

2. With the stated endpoint rule, which bin contains the eighteen-minute tray?

PR04 | Center, Variability, and Resistance

One unusually long wait

PRACTICE FROM THE LAST MEETING

S01.03-HW (W023). Complete 1, 2, 3, 5, 6.

READ FOR THE COMING MEETING

The mean of a sample is the sum of its observed values divided by the number of observations. Every value affects it. To find a median, first order the data. With an odd number of observations, take the middle value; with an even number, average the two middle values. A resistant statistic changes relatively little when a few observations become extreme. Resistance is a property of the method, not a claim that a statistic never changes.

OpenStax illustrates the distinction with a town of 50 earners: 49 earn $30,000 each and one earns $5,000,000. The median income is $30,000. The mean is [49($30,000)+$5,000,000]/50=$129,400. The mean is mathematically correct, but it lies far above the income of 49 of the 50 people. For a question about a typical resident's income, the median better represents the middle of this strongly right-skewed distribution. For a question about total income divided equally, the mean still has a useful interpretation.

Spread tells us how variable the observations are around a typical value. The range subtracts the minimum from the maximum. The sample standard deviation summarizes distances from the sample mean using squared deviations and then a square root; it has the original measurement units. Unlike the median, both the mean and standard deviation are sensitive to extreme values. Choose a summary to answer the question, and retain the context rather than writing only "mean" or "median."

A school printer finishes five small jobs in 6, 7, 8, 9, and 10 minutes. If you have to describe a typical job to another student, both the mean and the median are eight minutes. The mean distributes the total time evenly across five jobs. The median is the middle value when the times are ordered.

PR04 | Continued

Now suppose the final job took thirty minutes because a paper jam delayed printing. The total time increases by twenty minutes, so the mean increases by four minutes. The middle job still took eight minutes. The median stays in place because only the largest value changed. This is one example of what statisticians mean when they describe a summary as resistant to extreme observations.

The center alone leaves part of the experience unexplained. Consider the original five jobs, with mean 8 minutes. Subtract 8 from each time to get deviations of −2, −1, 0, 1, and 2 minutes. They sum to zero because the deviations on either side of the mean cancel. Squaring gives 4, 1, 0, 1, and 4. Their sum is 10. To calculate the sample variance, divide by one less than the number of jobs. Here 10/4 = 2.5 square minutes. Taking the square root gives a sample standard deviation of about 1.58 minutes. It describes the spread of individual job times around their mean. Apply the same reasoning to the jammed printer. Moving that fifth observation farther away makes the spread larger, even though the median stays put.

Printer completion times in minutes

Scenario	First four jobs	Fifth job
Original	6, 7, 8, 9	10
Paper jam	6, 7, 8, 9	30

Source and adaptation: OpenStax, Introductory Statistics 2e, §§2.5-2.7, especially Example 2.27 (income in a small town). The printer times are an SMS example.

TWO READING CHECKS

1. After the jam, what are the mean and median completion times?

2. What is the range before the jam, and what is it after the change?

PR05 | Quartiles, Outliers, and Group Comparisons

Two trails with the same typical time

PRACTICE FROM THE LAST MEETING

S01.04-HW (W027). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

The five-number summary lists the smallest observation, first quartile (Q1), median, third quartile (Q3), and largest observation. Try it on seven ordered walking times of 10, 12, 13, 14, 16, 18, and 25 minutes. The middle time is 14. Using this course's median-of-halves rule, leave it out of both halves. The median of 10, 12, 13 is 12, and the median of 16, 18, 25 is 18. The five-number summary is therefore 10, 12, 14, 18, 25 minutes. The interquartile range (IQR) is 18 − 12 = 6 minutes, the width of the middle half of the distribution. When quartiles are supplied, use those values.

In OpenStax's two-class test-score comparison, the day class has median 74.5 points and quartiles 56 and 82.5. The evening class has median 81 points and quartiles 78 and 89. The evening class has a higher median by 6.5 points, while its IQR of 11 points is smaller than the day class's 26.5 points. This is a comparison: each statement relates the groups directly and names the quantity being compared. Two separate lists of statistics would leave the reader to make that comparison.

A value is flagged by the 1.5×IQR rule if it lies below Q1−1.5×IQR or above Q3+1.5×IQR. A modified boxplot plots such values separately and extends each whisker to the most extreme observed value still within the fences. A fence is a calculated boundary, so it need not be an observation or a whisker endpoint. A flagged value deserves attention, not automatic deletion. Boxplots also conceal details such as clusters and gaps within each quartile.

Two hiking groups keep records of how long an easy loop takes. Both groups have a median time of 24 minutes. For Group A the first quartile is 20 minutes and the third quartile is 28 minutes. Those two quartiles bracket roughly the middle half of the times. Their difference, eight minutes, is the interquartile range, often abbreviated IQR.

PR05 | Continued

A boxplot draws a box from the first to the third quartile and marks the median inside. Group B has first and third quartiles of 17 and 31 minutes. Its box would be wider on the same time scale. Knowing the two medians alone would have missed this difference in the spread of the middle half.

For Group A, the upper fence equals the third quartile plus one and a half times the IQR, which gives 40 minutes. Suppose 39 and 44 minutes were this group's two longest observations. The upper whisker would end at 39, and 44 would appear as its own point. A different screen sometimes flags observations more than two standard deviations from the mean. That screen uses the mean and standard deviation, so it can flag a different set of times. For a modified boxplot using the 1.5 IQR rule, use the quartile fences. Investigate why a time is unusual before deciding what to do with it.

Trail times in minutes

Group	First quartile	Median	Third quartile
A	20	24	28
B	17	24	31

Source and adaptation: OpenStax, Introductory Statistics 2e, §§2.3-2.4, Examples 2.14 and 2.24 (day and evening test scores), and §2.7. The trail times are an SMS example.

TWO READING CHECKS

1. What are the IQRs for Groups A and B? Include units.

2. Would a recorded Group A time of exactly forty minutes be marked above the upper fence?

PR06 | Relative Standing and Changing Units

A score depends on the group

PRACTICE FROM THE LAST MEETING

S01.05-HW (W031). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A standardized score measures how far an observation lies from its group mean in standard deviation units. Subtract the mean from the value, then divide by the standard deviation: z=(value −mean)/standard deviation. The sign tells the direction, and the absolute value tells the number of standard deviations. Because the original measurement units cancel, a z-score has no units. A positive z-score does not automatically mean a better performance: for a race time, a smaller value can be preferable.

In OpenStax's swimming example, Angie swims 50 meters in 26.2 seconds on a team with mean 27.2 and standard deviation 0.8 seconds. Beth swims in 27.3 seconds on a team with mean 30.1 and standard deviation 1.4 seconds. Angie's z-score is −1.25, while Beth's is −2. Beth is slower in seconds but farther below her own team mean. The comparison is relative to two different distributions. It does not by itself establish which athlete would win a race.

A change of units offers a way to check a z-score. In tenths of a second, Angie's time is 262, her team mean is 272, and the standard deviation is 8. Her z-score is still $(262 - 272)/8 = -1.25$. If a timer added the same 3 seconds to everyone's time instead, the mean would also rise by 3 seconds, so every distance from the mean would stay the same. More generally, a new scale $y = a + bx$ has mean $a + b\mu$ and standard deviation $|b|\sigma$, where μ and σ are the original mean and standard deviation. The absolute value keeps the spread nonnegative even when a negative multiplier reverses the order.

A volunteer packs 36 pears in a trial at one orchard station. Workers at that station average 27 pears per trial, with a standard deviation of 6 pears. The volunteer's count is nine pears above the mean. Dividing that difference by six gives 1.5. A z-score of 1.5 places the count one and a half standard deviations above its group mean.

PR06 | Continued

At a second station the mean is 32 pears per trial and the standard deviation is 4 pears. A different volunteer packs 40. Comparing 40 with 36 answers who packed more. To compare how far above their own stations' means the two volunteers finished, first find each excess in pears and then express it in that station's standard deviation units. The second volunteer's count can be closer to the mean in pears yet farther from the mean on the standardized scale. Use the figures in the first check to decide.

Percentile rank asks a different question. Under the convention we use, a score's percentile rank is the percentage of observations at or below it. If eighteen of twenty workers at the first station packed 36 pears or fewer, the percentile rank of 36 is 90. Notice what you needed to learn that rank. The mean, standard deviation, and score tell you a distance. To determine the observed percentile rank, you also need information about where the other observations fall.

Pears packed per trial

Station	Volunteer's count	Station mean	Station SD
First	36	27	6
Second	40	32	4

Source and adaptation: OpenStax, Introductory Statistics 2e, §§2.3 and 2.7, especially Try It 2.35 (two swimmers). The pear-packing comparison is an SMS example.

TWO READING CHECKS

1. In the second station, how many pears above the mean is 40? How many standard deviations is that?

2. Do workers who packed exactly 36 pears count toward the percentile rank of 36? Explain.

PR07 | Choosing a Random Sample

Choosing trees from an orchard map

PRACTICE FROM THE LAST MEETING

S01.06-HW (W035). Complete 1, 2, 3, 4, 5, 6.

READ FOR THE COMING MEETING

A sampling method is a procedure for selecting observations, not merely a description of who happened to respond. In a simple random sample of size n, every possible subset of n members of the population has the same chance of selection. Giving every individual the same chance is not, by itself, enough to establish this stronger property. Start with a complete sampling frame, give distinct labels, select labels with an appropriate random mechanism, and explain how repeated or out-of-range labels are handled.

OpenStax compares plans for estimating undergraduate tuition at a university. Randomly selecting students separately from each class year gives a stratified random sample. Randomly selecting whole class years and including all students in the chosen years gives a cluster sample. In stratification, every stratum contributes a sample; in cluster sampling, only selected clusters contribute observations. Stating which groups are selected and which people within them are included makes the distinction reproducible.

A systematic sample uses a random starting point and a fixed interval along a list. A convenience sample selects accessible individuals. Stratifying by a characteristic related to the response can improve precision, whereas cluster sampling can reduce travel or collection effort. Neither method becomes a simple random sample merely because it includes a random step. To justify a plan, connect its grouping and selection rules to the population and statistical question.

Suppose a grower wants to estimate the proportion of 240 pear trees that need a second inspection. The orchard map identifies every tree, including those far from the gate. A worker labels the trees 1 through 240, uses a random-number generator to select twelve distinct labels, and visits the corresponding trees. If every possible group of twelve has the same chance of selection, it is a simple random sample.

PR07 | Continued

A systematic sample could use the same map. Number the 240 trees in a fixed order. To select 12, choose a starting label at random from 1 to 20, then take every twentieth tree. A start of 7 gives 7, 27, 47, and so on through 227. Every tree has a 1-in -20 chance of selection, but trees 7 and 8 cannot both be selected, so this is not a simple random sample. If the map order repeats a pattern every 20 trees, this procedure could repeatedly sample the same position in that pattern.

Now suppose the orchard grows two pear varieties. Sampling some trees from each variety guarantees both are represented. The varieties are strata. For a different design, the grower selects two of twenty rows at random and inspects every tree in those two rows. The rows are clusters. The first design spreads observations across both varieties. The second can save walking, but trees in the same row may share irrigation and growing conditions. To identify the design, ask what was chosen by chance and whether some or all trees in each group were inspected.

Three orchard sampling plans

Plan	What is randomly selected?
Simple random sample	12 individual trees from all 240
Stratified sample	Trees from within each variety
Cluster sample	2 rows; inspect every tree in them

Source and adaptation: OpenStax, Introductory Statistics 2e, §1.2, Example 1.11 (undergraduate tuition). The orchard plans are SMS examples.

TWO READING CHECKS

1. Why include distant trees on the map even if walking to them takes longer?

2. Which procedure ensures that both varieties appear in the sample?

PR08 | Bias, Observational Studies, and Ethical Questions

Who gets a chance to answer?

PRACTICE FROM THE LAST MEETING

S02.01-HW (W041). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

Bias is a systematic tendency of a method to overestimate or underestimate a target. It differs from random sampling variability, the sample-to-sample fluctuation that remains under a sound random design. More observations usually reduce random variability, but they cannot repair a procedure that repeatedly excludes an important part of the population. In a bias explanation, name the source of the problem and explain how it could affect the response.

OpenStax considers estimating textbook spending among 10,000 part-time students. One sample includes only students in an organic-chemistry class, and another includes only older students in physical-education classes. Neither sampling frame covers the full target population. The chemistry students might buy more expensive texts, while students taking other courses might have different costs. These are plausible mechanisms for bias, not proof of an exact numerical amount of bias. A defensible answer explains the relevant difference rather than simply calling a sample “too small.”

Undercoverage occurs when part of the population is absent or poorly represented in the frame. Nonresponse occurs when selected individuals do not provide data, and response bias arises when the measurement or question systematically distorts answers. Voluntary-response samples can overrepresent people strongly motivated to answer. Ethical collection also requires appropriate consent, protection of identifying information, and an honest account of how data were obtained. An observational study records existing conditions; without random assignment, an association may have explanations besides a causal effect.

A town wants to estimate the proportion of residents who use a public walking path at least once a week. Staff post a survey at the path entrance on Saturday morning. The survey is convenient to conduct, but a resident who never visits the path has little chance of being included. If frequent users are easier to reach there, the survey could overstate weekly use.

PR08 | Continued

Suppose staff instead select 200 residents at random from a complete resident list and mail a neutral question. Only 100 reply, and 70 of those respondents report weekly use. The response rate is 100 out of 200. The observed weekly-use rate among respondents is 70 out of 100. Whether that second percentage describes all residents well depends partly on how the people who did not reply compare with those who did.

What might be learned about path use without assigning anyone a treatment? Staff could enroll residents now and follow them for four weeks to record future walks. That would be a prospective observational study. They could instead use existing records to compare past visits. That would be retrospective. A survey question asking about visits in the past seven days gives respondents a concrete period to recall. A neutral question and a way to answer privately help protect the quality of those reports. Each design still requires staff to explain who was included and who may be missing.

The mailed walking-path survey

Quantity	Count
Residents selected	200
Residents responding	100
Weekly users among respondents	70

Source and adaptation: OpenStax, Introductory Statistics 2e, §§1.2 and 1.4, especially Example 1.13 (textbook spending). The walking-path survey is an SMS example.

TWO READING CHECKS

1. Give a reason the entrance survey might overestimate weekly path use.

2. For the mailing, calculate the response rate and the use rate among respondents.

PR09 | Designing a Completely Randomized Experiment

A fair comparison on the greenhouse bench

PRACTICE FROM THE LAST MEETING

S02.02-HW (W045). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

An experiment deliberately imposes a treatment and observes a response. A treatment is a specific experimental condition, an experimental unit receives an assigned treatment, and a response variable measures an outcome. A comparison group helps separate the treatment effect from other changes that could occur during the study. Control means keeping other relevant conditions comparable. Replication means applying each treatment to enough independent experimental units to assess variation.

In OpenStax's aspirin example, 400 participants are randomly divided between an aspirin group and a placebo group. The treatment is the pill assignment, each person is an experimental unit, and the response is whether that person has a heart attack during the study period. A placebo permits a comparison with a treatment that resembles the active pill. Blinding participants or those measuring outcomes can limit expectation effects; whether blinding is possible depends on the treatment and measurement.

Random assignment uses chance to form treatment groups. It helps prevent a systematic association between treatment and preexisting characteristics and provides a basis for judging whether an observed difference is convincing evidence of an effect. It does not guarantee perfectly identical groups. Random selection addresses population representation, a different question. When asked to describe an experiment, specify units, treatments, an implementable assignment rule, relevant controls, the response and its measurement, and the comparison to be made.

A greenhouse class has 24 seedlings and wants to compare two watering schedules. Students could give one schedule to the twelve plants by the sunny window and the other schedule to the twelve plants against the back wall. If the groups grow by different amounts, water and light would both be plausible explanations. The location would interfere with the comparison.

PR09 | Continued

Instead, the class labels the seedlings 1 through 24, shuffles 24 labeled slips thoroughly, and draws 12 without replacement for schedule A. The remaining 12 get schedule B. Each plant receives the same total weekly amount of water. Only the frequency of watering differs. Students keep soil and pot size consistent and distribute the two schedules across bench positions. After four weeks they record each plant's change in height in centimeters and compare the groups' mean changes.

Watering frequency is a factor, a variable the class deliberately changes. The two schedules are its levels. If the class also varied fertilizer (with or without), it would have four treatments, one for each combination of schedule and fertilizer. To count experimental units, follow the assignment. A seedling gets its own assigned schedule, so that seedling is one unit. Measuring it repeatedly could show how its height changes over time. To learn how growth varies among plants given the same schedule, the class needs several independently treated seedlings.

Assignment on the greenhouse bench

Plan	Schedule A	Schedule B
By location	12 by the window	12 by the back wall
By random assignment	12 randomly assigned	Remaining 12

Source and adaptation: OpenStax, Introductory Statistics 2e, §1.4, Example 1.19 (aspirin and placebo). The greenhouse experiment is an SMS example.

TWO READING CHECKS

1. Name the treatments and the response variable in the redesigned experiment.

2. Why could the original window-versus-wall plan give a misleading comparison?

PR10 | Blocking, Matched Pairs, and Scope of Inference

Each volunteer tries both routes

PRACTICE FROM THE LAST MEETING

S02.03-HW (W049). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A randomized block design groups similar experimental units before randomly assigning treatments within every block. A blocking variable should be related to the response. Comparing treatments among similar units reduces some variation that would otherwise obscure the treatment comparison. Blocking is not a substitute for random assignment: assigning one whole block to each treatment would tie the treatment to the block characteristic.

A matched-pairs design is a special case. Each person may receive both treatments, or two similar individuals may form a pair with one assigned to each treatment. OpenStax describes participants completing mazes with scented and unscented masks. Each participant supplies outcomes under both conditions, and the order of conditions is randomly assigned. A participant's usual speed is less troublesome when that same person supplies both observations. Learning, fatigue, and carryover remain concerns, so the order and timing require deliberate control.

Analyze a pair through a difference defined in a fixed order. The sign then has one interpretation throughout the study. Random assignment supports a causal comparison for the participants when the design and evidence justify it. A random sample supports generalization to its sampled population. A volunteer experiment can have strong internal comparison while still requiring caution about generalizing to everyone. Use these distinctions to explain both what a design supports and what additional evidence it would need.

A map club compares two ways to guide volunteers to a course marker. One format uses symbols and the other uses written steps. Volunteers differ in walking speed and experience. If each volunteer tries both direction formats on comparable courses, the club can compare two times from the same person. The person serves as their own comparison in this matched-pairs design.

For example, Mira takes 7 minutes with the symbol directions and 9 minutes with the written directions. Her symbol-minus-written difference is negative two minutes. Calculating the difference in the same order for every volunteer makes a negative result consistently mean a shorter time with symbols. A comparison based on these differences accounts for some of the variation in volunteers' usual speed.

PR10 | Continued

A second attempt might be quicker simply because the volunteer has practiced. The club can randomly assign the order of formats and balance which course uses which format so that a particular format is not always second or always used on the easier course. A convincing design could support a causal comparison for the tested volunteers and conditions. If the volunteers came from a map club, applying the result to every walker in town would require further evidence.

How could the club judge a difference across several volunteers? Imagine that the directions make no difference to anyone's time. Within each volunteer's pair, reversing the random assignment of formats would reverse the sign of the difference. In a study that uses a separate fair coin for each person's assignment, the club could simulate that null explanation by independently flipping the signs and recalculating the average difference many times. An observed average far into the tail in the direction predicted before the experiment would be evidence against that explanation. This comparison uses the pairing throughout.

Keep the two times from one volunteer together

Volunteer	Symbols (min)	Written (min)	Symbols − written (min)
Mira	7	9	−2

Source and adaptation: OpenStax, Introductory Statistics 2e, §1.4, Example 1.20 (scented masks and mazes). The direction-format experiment and sign-flip illustration are SMS examples.

TWO READING CHECKS

1. Interpret Mira's difference of −2 minutes using the specified subtraction order.

2. Why would giving everyone the symbol directions on the second attempt be a concern?

PR11 | How much can a sample change the story?

Keep the selection record

PRACTICE FROM THE LAST MEETING

S02.04-HW (W055). Complete 1, 2, 3, 4, 5, 6.

READ FOR THE COMING MEETING

The project asks how well a random sample of ten books estimates the mean page count of an illustrative forty-book collection. The workbook supplies all forty page counts, but begin without calculating their population mean. A book is the observational unit, page count is the quantitative variable, and the mean page count of all forty books is the parameter. The ten sampled books will provide a sample mean that estimates it.

Label the books 01-40. Use a random-number generator to select ten distinct labels, skipping repeats, and save the labels before inspecting their page counts. Retain every selected book, including an unusually long one. A partner should be able to use the saved labels to recover exactly the same sample. Replacing a selected long book with a short one changes the procedure and can systematically lower the estimates.

How much can one substitution change an estimate? In a separate five-book example, replacing a 340-page book with a 240-page book lowers the sample total by 100 pages. The denominator stays at five, so the sample mean falls by 100/5 = 20 pages. On the class dotplot the group would move one dot 20 pages to the left. That dot represents a whole group’s sample mean.

PR11 | Continued

Different groups following the same valid method will usually obtain different samples and different means. That is sampling variability. In class you will compare the sample means with the full collection's mean and with a deliberately biased method, such as selecting the shortest books. One inaccurate random sample does not demonstrate that the method is biased. Bias concerns whether the method systematically favors estimates on one side of the population value over repeated samples.

Use the given collection as the entire population for this investigation; the page counts are fictional classroom data. The project is descriptive: ten of forty books is a 25% sampling fraction, so do not attach an independent-sampling standard-error formula or confidence interval to this exercise. Later lessons examine when those inferential calculations are justified.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S02.01 (R029).

TWO READING CHECKS

1. Why save the ten labels before inspecting the page counts?

2. Two groups use valid random selection and obtain different means. Does this alone establish bias? Explain.

PR12 | Checkpoint 1

Follow a comparison back to its data

PRACTICE FROM THE LAST MEETING

Using your project records, (1) reconstruct the mean from your ten saved labels, (2) compare its signed error with that of the group's deliberately biased sample, and (3) write one paragraph using the class display to distinguish variability from bias. Use the workbook population if you missed collection.

READ FOR THE COMING MEETING

Suppose two school cafeteria lines have recorded median wait times of 6.5 and 8.5 minutes, respectively. A student says that anyone in the first line will finish sooner. Can the two medians support that prediction? Each median describes the center of a group of observed wait times. It tells us nothing certain about which of two individual customers finishes first.

One group timed only its friends during the first ten minutes of lunch. Another randomly selected student arrivals throughout each lunch period on several days. The first group's records can describe its friends, while the second plan covers more of the population about which the students want to draw conclusions.

The lines may serve different meals or attract students with different schedules. Even a random sample of customers leaves those differences in place when customers choose their own lines. A comparison can establish which line had the shorter median wait in the sample. Attributing the gap to the line itself requires more evidence.

How could the cafeteria test a proposed change? Suppose the same menu is offered at both stations. Within each lunch period, staff could randomly assign each volunteer to a station, with equal numbers at each. This keeps the stations comparable across busy and quiet periods. Timing from the moment each customer joins the line until checkout gives everyone the same response definition. Random assignment supports a comparison of station effects for the volunteers. How the volunteers were recruited still limits whom they represent.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S01.05 (R021); S02.04 (R041).

TWO READING CHECKS

1. What does the difference between the reported medians tell you in minutes?

2. Name one reason timing only friends early in lunch could give an incomplete picture.

PR13 | Chance models and reproducible simulation

Three shots make one trial

PRACTICE FROM THE LAST MEETING

(1) Revise one description and one study-design answer from your returned checkpoint, marking the evidence added to each. (2) Sketch a dotplot for kitchen times of 2, 3, 3, 4, and 8 minutes. (3) Find the median and explain why selecting only the busiest shifts would limit the interpretation.

READ FOR THE COMING MEETING

A probability model specifies possible outcomes and probabilities under stated assumptions. An outcome is one possible result of the complete random process, and an event is a set of outcomes. A probability describes a long-run relative frequency under that model. Short runs can differ substantially from that frequency. Before a simulation, define the process to be repeated and the event or statistic to record from each repetition.

OpenStax's two-coin example has ordered outcomes HH, HT, TH, and TT. With independent fair tosses, each has probability 1/4. HT and TH are distinct because the first and second tosses differ. The event "at least one tail" contains HT, TH, and TT, so its model probability is 3/4. To simulate it, generate two independent equally likely head/tail outcomes, record whether that complete pair has at least one tail, and repeat the two-toss process many times. Counting individual tails would answer a different question.

Write the simulation instructions so a partner could repeat the experiment. Say how to generate one outcome, what counts as one complete repetition, what result to record, and how many repetitions to run. Decide which results count toward the event before running the simulation. In the soccer example, tally whole three-shot sets.

Suppose a soccer player makes 70% of her practice penalty shots over the long run. A coach wants to estimate how often she would make all three shots in a set. The coach proposes a model in which every shot has the same chance of success and the outcome of one shot has no effect on the next. A simulation can explore what would happen under those assumptions.

PR13 | Continued

For each shot, generate a digit from 0 through 9 independently of earlier digits. Make all ten digits equally likely. Assign 1 through 7 to a goal and 8, 9, and 0 to a miss. Generate three digits to represent one whole set. The digits 6, 1, and 4 represent a set with three goals. Put one mark beside this set because it meets the coach's criterion.

Repeat the complete set 100 times. If 35 sets have three goals, the estimated chance of a three-goal set is 35/100 = 0.35. Imagine that the player then scores all three in a real set. Would the coach have reason to doubt the model? About a third of the simulated sets did just as well, so that performance is quite plausible under the model. Before drawing a conclusion, the coach should still ask whether fatigue or changing conditions could affect the shots.

Digit model for one penalty shot

Outcome	Equally likely digits
Goal	1, 2, 3, 4, 5, 6, 7
Miss	8, 9, 0

Source and adaptation: OpenStax §§3.1-3.2 and Example 3.6, printed pp. 168-174 (PDF pp. 180-186). Source probability model adapted; reproducible simulation procedure developed for AP CED.

TWO READING CHECKS

1. What happens in the set represented by 2, 7, and 0? Does the set earn a mark?

2. Why is 100 the denominator in the coach's estimate?

PR14 | Events, complements, and overlapping categories

Two sign-up sheets, one group of hikers

PRACTICE FROM THE LAST MEETING

S03.01-HW (W063). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

The complement of an event contains all outcomes outside that event, so its probability is 1 minus the event probability. “A or B” includes outcomes in A, in B, or in both. “A and B” includes only their overlap. These phrases refer to membership in sets, so a Venn diagram or a table can clarify the event before a calculation. If A and B are mutually exclusive, their overlap is empty; otherwise simply adding their probabilities counts the overlap twice. Written as a probability rule, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

OpenStax gives a college example in which 40% belong to a club, 50% work part time, and 5% do both. The proportion in at least one group is $0.40+0.50-0.05=0.85$. The proportion in neither is $1-0.85=0.15$. The proportion in exactly one group is $0.40+0.50-2(0.05)=0.80$ because both copies of the overlap must be excluded. The three different answers correspond to three different events.

A two-way table gives another way to see the overlap. Imagine 100 students in the college example. Put the 5 who do both activities in one cell. That leaves 35 club-only students and 45 work-only students. The remaining 15 do neither. The four cells add to 100. Once every person belongs to exactly one cell, you can add whichever cells match the question without counting anyone twice.

Forty volunteers can sign up for either or both of two trail jobs. Twenty-four sign the brush-clearing sheet, eighteen sign the litter-pickup sheet, and eleven sign both. The coordinator wants to know the chance that a randomly selected volunteer signed up for at least one job. Adding the two sheet totals gives 42 names, more than the number of volunteers. Where do the extra names come from?

PR14 | Continued

The eleven volunteers who signed both sheets each contribute two names to the total. Count them once by subtracting one copy of the overlap. There are 24 + 18 − 11 = 31 distinct volunteers who signed at least one sheet. Each volunteer is equally likely to be chosen, so the probability of selecting someone in this group is 31/40.

The remaining nine volunteers signed neither sheet. Their event is the complement of signing at least one. Together these two groups account for all forty volunteers. Signing exactly one sheet is a different event because it excludes the eleven volunteers who signed both. Sketching the groups can help you decide which people belong in a count.

Sign-ups among 40 volunteers

Names counted on	Number of volunteers
Brush-clearing sheet	24
Litter-pickup sheet	18
Both sheets	11

Source and adaptation: OpenStax §§3.1, 3.3, 3.5, printed pp. 168-171,179-182,194-195 (PDF pp. 180-183,191-194,206-207), especially Example 3.29. Addition/complement rules adapted; trail counts are original.

TWO READING CHECKS

1. How many volunteers signed only the litter-pickup sheet?

2. What is the probability that the selected volunteer signed neither sheet?

PR15 | Conditional probability and independence

Which market visits belong in the denominator?

PRACTICE FROM THE LAST MEETING

S03.02-HW (W069). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A conditional probability restricts attention to outcomes satisfying given information. P(A given B) means the probability of A among outcomes in B. When P(B)>0, calculate it as P(A and B)/P(B). The denominator changes because the relevant set of possible outcomes has changed. P(A given B) and P(B given A) generally differ, even though both numerators use the same overlap.

In OpenStax's course-enrollment example, 75% take English, 30% take speech, and 22.5% take both. Among speech students, the English proportion is 0.225/0.30=0.75, equal to the overall English proportion. Among English students, the speech proportion is 0.225/0.75=0.30. Thus knowing one enrollment does not change the other probability in this model. The events are independent. They are not mutually exclusive, because some students take both.

Independence means P(A given B)=P(A), when the conditional probability is defined. An equivalent test is P(A and B)=P(A)P(B). Mutually exclusive events with positive probabilities are dependent: knowing one occurred rules out the other. When interpreting a two-way table, do not decide independence from similar-looking raw counts. Compare conditional probabilities using their correct totals, or compare a joint probability with the product of the marginals. The following sales records show why the denominator matters.

A farm stand records 100 visits. Of 60 morning visits, 18 involve a pear purchase. Of 40 afternoon visits, 24 involve one. Imagine choosing one recorded visit at random. The probability of a purchase is (18 + 24)/100 = 0.42. This is a marginal probability because it uses the purchase total across both times of day. The joint probability of a morning visit and a purchase is 18/100 = 0.18. Its numerator counts visits satisfying both conditions.

PR15 | Continued

Now suppose someone tells you that the chosen visitor came in the morning. Only 60 visits fit that information. Of those, 18 involve a purchase, giving a conditional probability of 18/60. If you learn instead that the visitor bought pears, the relevant group contains 42 visits. The question about whether that buyer came in the morning now calls for 18/42.

To test independence in these records, compare the morning purchase probability, 18/60 = 0.30, with the overall purchase probability, 42/100 = 0.42. The change shows dependence for a random visit drawn from these records. It does not tell the stand owner why morning and afternoon purchases differ.

Pear purchases at the farm stand

Visit time	All visits	Bought pears
Morning	60	18
Afternoon	40	24

Source and adaptation: OpenStax §§3.2-3.4, printed pp. 172-188 (PDF pp. 184-200), especially Example 3.10. Conditional probability and independence adapted; pear-sales case is original.

TWO READING CHECKS

1. What fraction of the afternoon visits involved a pear purchase?

2. What fraction of all 100 visits were afternoon visits with a pear purchase?

PR16 | Multi-stage chance processes and probability trees

Two names from a bowl

PRACTICE FROM THE LAST MEETING

S03.03-HW (W075). Complete 1, 2, 3, 5, 6.

READ FOR THE COMING MEETING

Start a probability tree by drawing one branch for each possible first outcome. From the end of each branch, draw the possible second outcomes. Label each second branch with the chance of its outcome given the first. The branches leaving each point should sum to 1. To find the chance of one complete path, multiply its branch probabilities. An event can include several different paths. Add the probabilities of all its paths, counting each only once.

OpenStax draws two marbles without replacement from a bag containing 3 red and 8 blue marbles. The probability of red first is $3/11$. Given red first, only 2 red remain among 10 marbles, so the probability of two red is $(3/11)(2/10)=3/55$. Given blue first, all 3 red remain among 10 marbles. The probability of blue then red is $(8/11)(3/10)=12/55$. The second-stage probabilities depend on the first draw because the first marble is not returned.

For one marble of each color, include both red-blue and blue-red. The chance of red-blue is $(3/11)(8/10) = 12/55$. Blue-red also has probability $12/55$, so the two paths together have probability $24/55$. Returning the first marble and remixing changes the tree. After red first, the chance of blue second becomes $8/11$, giving $(3/11)(8/11)$ for red-blue. Multiplication works in both trees because the second factor always represents the chance after the first outcome.

A club places six names in a bowl, four juniors and two seniors. A teacher mixes the slips, draws one name, sets it aside, and draws a second name. Suppose every remaining slip is equally likely on each draw. The chance that the first name is a senior is $2/6$. If a senior is drawn first, five names remain, including just one senior. The chance of another senior at that point is $1/5$.

PR16 | Continued

One path through these draws is senior followed by senior. To follow that entire path, both stages must occur. Its probability is $(2/6)(1/5)$, or $1/15$. The second factor uses what is left after the first draw. Drawing a junior first would leave a different mixture of names for the second draw.

The club might instead ask for one representative from each grade. There are two possible orders, senior followed by junior and junior followed by senior. A successful pair follows exactly one of those paths. Find the probability along each path by multiplying its stage probabilities, then add the two path probabilities.

Second draw after a junior is drawn, without replacement

Names remaining	Juniors	Seniors	Total
Original bowl	4	2	6
After a junior	3	2	5

Source and adaptation: OpenStax §3.5, Examples 3.24-3.25, printed pp. 189-192 (PDF pp. 201-204). Adapted from the marble-draw models with and without replacement. The club election is an original example.

TWO READING CHECKS

1. If the first student is a junior, how many seniors remain among how many names?

2. Find the probability of a junior followed by a senior.

PR17 | Random variables, distributions, and expected value

What does seven-tenths of a request mean?

PRACTICE FROM THE LAST MEETING

S03.04-HW (W081). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A random variable assigns a number to each outcome of a chance process. Define what the number counts or measures and the relevant time period. A discrete probability distribution pairs each possible value with its probability. Every probability lies between 0 and 1, and the probabilities sum to 1. The most likely value, the list of possible values, and the long-run average are three different features of that distribution.

OpenStax models the number of days a soccer team plays in a week as 0, 1, or 2 with probabilities 0.20, 0.50, and 0.30. The expected value is the probability-weighted average: $0(0.20)+1(0.50)+2(0.30)=1.1$ days per week. One week cannot contain 1.1 playing days, but an average across many weeks can approach 1.1. Averaging 0, 1, and 2 without weights would incorrectly give each outcome the same probability.

To find a cumulative probability, add the probabilities of all values meeting the boundary. For the soccer model, “at most 1 day” includes 0 and 1 and has probability 0.70. “At least 1 day” includes 1 and 2 and has probability 0.80.

How much do the weekly outcomes vary around 1.1 days? The deviations for 0, 1, and 2 days are −1.1, −0.1, and 0.9 days. Square each deviation so values on opposite sides of the mean do not cancel. Then weight by the chance of each outcome. The variance is $0.20(1.21) + 0.50(0.01) + 0.30(0.81) = 0.49$ days squared. Its square root, 0.70 day, is the standard deviation of the weekly count. Returning to days lets us describe the spread in the same units as the original counts.

A volunteer tracks how many map requests arrive during a ten-minute period at a trail kiosk. For a quiet morning, she proposes a model with only three possible counts. The chances of zero, one, and two requests are 0.50, 0.30, and 0.20. Call the count X. Before a period begins, its value is uncertain. Once that period ends, someone can record the whole-number value that occurred.

PR17 | Continued

The model assigns a probability to each possible value of X. The probabilities add to one because the possibilities cover every outcome in this model. To find the chance of at most one request, combine the probabilities for zero and one request. When you add probabilities up through a specified count, the result is a cumulative probability.

Imagine observing 100 such periods. Roughly 50 with no requests, 30 with one, and 20 with two would give 70 requests in total. This reasoning suggests a long-run average of 0.70 requests per period. The value 0.70 describes an average across many periods even though a volunteer must record a whole number in each one.

Map requests per ten-minute period

Count of requests	Probability
0	0.50
1	0.30
2	0.20

Source and adaptation: OpenStax §§4.1-4.2, printed pp. 228-236 (PDF pp. 240-248), especially Example 4.3 and Table 4.5. Expected-value explanation and soccer model adapted; kiosk model is original transfer practice.

TWO READING CHECKS

1. Calculate the probability of at most one request.

2. Use the kiosk model and its mean of 0.70. Calculate the standard deviation of the request count. Show how the probabilities enter your calculation and give the units.

PR18 | The binomial model

Eight packets, one count

PRACTICE FROM THE LAST MEETING

S04.01-HW (W087). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A binomial random variable counts successes in a fixed number of trials. Check four features: there is a fixed number n of trials; each trial is classified as success or failure; trial outcomes are independent; and the success probability p is the same on every trial. “Success” means the outcome being counted, even when that outcome is a failure of a device or a missed shot. Name it explicitly. A process that continues until a target number of successes occurs does not have a fixed number of trials.

OpenStax describes an altered coin with probability 0.25 of heads, flipped independently five times. Let X count heads. For exactly four heads, one possible order is HHHHT, with probability $(0.25)^4(0.75)$. The single tail could fall in any of five positions. Multiplying the probability of one order by 5 gives $P(X = 4) \approx 0.01465$. In general, $P(X = k)$ is the number of orders with k successes times $p^k(1-p)^{n-k}$.

“More than three heads” includes four and five. Five heads has only one possible order, with probability $(0.25)^5$. Adding the probabilities for four and five gives $P(X > 3) = 0.015625$. About 1.6% of five-flip sets would have more than three heads in the long run. A cumulative binomial calculation gives the chance of k or fewer successes. Thus $1 - P(X \le 3)$ gives the same answer. For “at least three,” the count three belongs in the event, so use $1 - P(X \le 2)$.

A teacher prepares eight packets of wildflower seeds for a trial. Each packet contains one seed. For this model, suppose every seed has a 25% chance of germinating and that one seed’s outcome gives no information about another’s. The teacher plants all eight and counts how many germinate. Each seed contributes one of two outcomes to the count.

PR18 | Continued

Here success means germination. The mean of a binomial count is np and its standard deviation is $\sqrt{np(1-p)}$. For the eight packets, the mean is $8(0.25) = 2$ germinated seeds per set. The standard deviation is $\sqrt{8(0.25)(0.75)} \approx 1.22$ seeds. Two describes the long-run average across sets of eight. The outcomes of individual sets can range from zero through eight.

Consider what could disrupt this model. If all packets share a tray that sometimes dries out overnight, learning that one seed failed could change your assessment of the others. If the teacher instead plants additional seeds until two germinate, the number planted is no longer fixed in advance. The question you ask and the way you run the trial both matter.

Before calling a count binomial

Feature	Question to answer
Fixed n	How many trials will be made?
Two outcomes	Which outcome counts as success?
Independence	Does knowing one outcome change the probability of another?
Common p	Is success equally likely on every trial?

Source and adaptation: OpenStax §4.3, printed pp. 237-241 (PDF pp. 249-253), especially Example 4.11. Binomial definition and altered-coin model adapted; source's stray "fair" label is corrected. Greenhouse example is original transfer practice.

TWO READING CHECKS

1. What counts as a success in the proposed eight-seed model?

2. For the eight independent seeds, a calculator gives $P(X \le 1) \approx 0.3671$ and $P(X \le 2) \approx 0.6785$. What is the probability that at least two seeds germinate? State which counts you exclude.

PR19 | Continuous models and the normal distribution

A band of board lengths

PRACTICE FROM THE LAST MEETING

S04.02-HW (W093). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

For a continuous random variable, probability is represented by area under a density curve over an interval. The total area is 1, and the density is never negative. A curve's height is not itself the probability of one exact value. A single point has no width, so its area and probability are 0 in a continuous model. Consequently, including or excluding one endpoint does not change a continuous probability, although it can change a discrete probability.

A normal model has a symmetric, bell-shaped density curve described by its mean and standard deviation. The mean locates the center; the standard deviation controls horizontal spread. OpenStax's example uses mean 50 and standard deviation 6. Approximately 68% of values lie from 44 to 56, 95% from 38 to 62, and 99.7% from 32 to 68. These are the intervals one, two, and three standard deviations from the mean. Under the normal model, the area outside the middle 95% is 5%, divided equally between the two tails.

The empirical rule is a useful approximation at these particular distances. It does not apply to every symmetric distribution or every data set with a known mean and standard deviation. To use a normal model, the problem must supply or justify its shape. Standardizing with $z=(x-\mu)/\sigma$ tells how many standard deviations a boundary is from the mean. It changes units and scale but preserves the relevant area. Sketch the model, label the center and boundary values, and shade the event before reading an area.

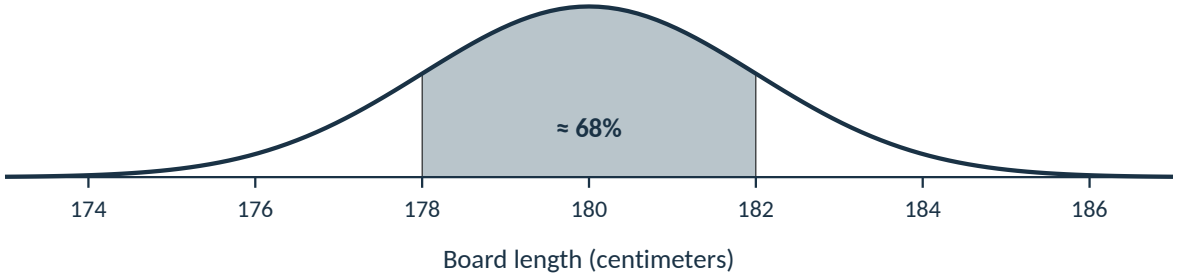
A shop cuts boards to a target length of 180 centimeters. Imagine measuring finished boards with an instrument precise enough to distinguish small differences. The shop uses a normal model with mean 180 centimeters and standard deviation 2 centimeters. On a sketch of this bell-shaped curve, the center sits at 180. The marks one standard deviation from the center lie at 178 and 182.

PR19 | Continued

Now imagine a second cutting process with the same mean of 180 centimeters and a standard deviation of 1 centimeter. For the first process, 179 to 181 centimeters covers just half a standard deviation on each side of the mean. For the second process the same interval covers a full standard deviation on each side, about 68% of the area. On a common horizontal scale, sketch the second curve narrower and taller. Both curves enclose the same total area, one.

An individual board with a length of 184 centimeters is two standard deviations above the mean. Its position follows from the four-centimeter difference divided by two centimeters per standard deviation. An instrument rounding to the nearest centimeter can display 184 cm for a board 183.8 cm long. A positive probability for the rounded reading is consistent with probability zero for any one exact continuous length.

Board lengths within one standard deviation of 180 cm



Source and adaptation: OpenStax §5.1 and §6.1, printed pp. 292-294 and 336-340 (PDF pp. 304-306 and 348-352), especially Example 6.5. Continuous-area interpretation and empirical rule adapted; local measurement model retained.

TWO READING CHECKS

1. What interval of lengths is within two standard deviations of the mean?

2. How far from the mean is a 178-centimeter board, in standard deviations and direction?

PR20 | Normal probabilities, percentiles, and cutoffs

How heavy must the top tenth be?

PRACTICE FROM THE LAST MEETING

S04.03-HW (W097). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

Normal calculations answer two kinds of question. A forward probability calculation starts with one or more values and finds an area. An inverse calculation starts with an area or percentile and finds a value. Both require a defined variable, a justified normal model, and a clear event. State whether the second model parameter is a standard deviation; do not confuse it with a variance.

OpenStax models final-exam scores with mean 63 and standard deviation 5. For $P(X>65)$, the boundary has $z=(65-63)/5=0.4$. A standard normal table gives about 0.6554 to the left of 0.4, so the requested upper-tail area is $1-0.6554=0.3446$. The result is a probability, not a score. For the 90th-percentile score, instead find $z\approx1.2816$ and convert back: $63+1.2816(5)\approx69.4$ points. This result is a score, with about 90% of modeled scores at or below it.

For an interval containing the middle C of a symmetric normal distribution, split the remaining $1-C$ equally between the tails. The middle 80% leaves 10% in each tail and uses the 10th and 90th percentiles as its boundaries. “Above,” “below,” “between,” and “middle” specify different shaded regions. A calculation is easier to check when the sketch and its expected direction come first. For example, a cutoff for the highest 10% must exceed the mean; an output below the mean signals an area or sign error.

A kitchen uses a normal model for the mass of a serving of rice, with mean 160 grams and standard deviation 10 grams. A cook asks what proportion of servings exceed 170 grams. Here the boundary, 170 grams, is already known. Shade the part of a bell curve to the right of that boundary. Its area represents the proportion of servings heavier than 170 grams. The boundary is one standard deviation above the mean. A standard normal table gives about 0.8413 to the left, leaving $1 - 0.8413 = 0.1587$ to the right. Under the kitchen’s model, about 16% of servings exceed 170 grams.

A different question starts with a proportion. The cook wants a cutoff such that the heaviest 10% of servings exceed it. Mark a boundary with a small shaded region to its right. That region contains 0.10 of the area, so 0.90 of the area lies to its left. A calculator that asks for the area to the left needs 0.90 to find this cutoff.

PR20 | Continued

For a standard normal curve, the boundary with 0.90 of the area to its left lies about 1.28 standard deviations above the mean. In the kitchen's units, that is about 160 + 1.28(10) = 172.8 grams. Sketching the tail first gives an immediate check that a cutoff for the heaviest tenth lies above the mean.

An interval uses two boundaries. For servings between 150 and 170 grams, the standardized boundaries are −1 and 1. The area to the left of −1 is about 0.1587. Subtract that from the area to the left of 1 to get 0.8413 − 0.1587 ≈ 0.683. This agrees with the empirical rule for a band one standard deviation on either side of the mean. Standardized positions also allow comparisons across settings. A serving mass of 170 grams is one standard deviation above its mean. The exam score of 65 was only 0.4 standard deviation above its mean, so the serving mass sits at the higher percentile in its own normal distribution.

Two questions about serving mass

Given	Find
Boundary of 170 g	Proportion above the boundary
Heaviest 10% of servings	Lower cutoff in grams

Source and adaptation: OpenStax §6.2, printed pp. 340-347 (PDF pp. 352-359), especially Example 6.8(a,c). Forward/inverse calculations adapted; rice-serving example is original transfer practice.

TWO READING CHECKS

1. Which area would a left-area calculator use to find the cutoff for the lightest 10%?

2. In the rice model, what does the number 172.8 measure?

PR21 | Checkpoint 2

Translate the event before calculating

PRACTICE FROM THE LAST MEETING

S04.04-HW (W103). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A repair team tests four independently manufactured lights. Suppose each light has probability 0.10 of failing inspection, and the model treats the four outcomes as independent. Let X count how many fail. The possible values of X are the whole numbers zero through four. A phrase such as “at least one fails” includes four different values of X.

The complement, zero failures, describes one simple sequence. Every light must pass. Under the stated model its probability is the product of four factors of 0.90. Subtracting that probability from one gives the probability of at least one failure. The same word “at least” would include more values if the sample were larger.

For these four lights, P(no failures) = 0.90 × 0.90 × 0.90 × 0.90 = 0.6561. The chance of one or more failures is therefore 1 − 0.6561 = 0.3439. The value 4 × 0.10 = 0.4 has a different meaning. It is the expected number of failures per group of four, averaged over many groups. A probability has no units and must lie between zero and one. An expected count is expressed in failures.

A shared voltage surge could make several failures occur together. If that is plausible, the product calculation needs additional justification. Write down what the random variable counts, the event you need, and the assumption behind any multiplication.

For a simulation, let digit 0 represent failure and digits 1 through 9 represent passing. Four independently generated digits make one simulated group. For an event worded “more than three,” a group with exactly three failures would miss the cutoff. Changing the wording to “three or more” includes that same group. Mark the boundary before counting qualifying groups.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S03.02 (R051); S04.02 (R068).

TWO READING CHECKS

1. List the values of X in the event “at least three lights fail.”

2. Which model assumption permits multiplying four pass probabilities to find the probability of no failures?

PR22 | Sampling distributions and randomization distributions

One dot for each sample

PRACTICE FROM THE LAST MEETING

(1) For five independent sensors with failure probability 0.20 each, draw a probability tree for the first two tests. (2) Use the complement to find the probability of at least one failure among five. (3) Give the expected number of failures and explain one way shared conditions could invalidate the calculation.

READ FOR THE COMING MEETING

A statistic is a number calculated from a sample. Its sampling distribution describes the values it would take over all possible random samples of a specified size from a population, using a specified sampling procedure. The distribution of individual observations answers a different question. Always ask what one dot represents: one person's response, a sample's summary, or a statistic calculated after a treatment reassignment.

OpenStax introduces the idea by repeatedly rolling five dice and recording the mean of each five-roll sample. For the sample 2,2,3,4,6, the sample mean is $17/5=3.4$. That whole sample contributes one value, 3.4, to the distribution of sample means. Plotting the five separate rolls would instead describe individual die outcomes. Repeating the complete five-roll sample many times builds an approximation to a sampling distribution. Increasing simulation repetitions makes that approximation more stable; it does not change the sample size of five.

The center of a sampling distribution indicates whether the statistic targets the parameter on average. Its spread measures sample-to-sample variability. A biased estimator systematically centers above or below the target; an individual estimate can miss the target even when its estimator is unbiased. Larger random samples often reduce variability. A randomization distribution uses a different chance process: keep observed responses fixed and reassign treatments according to the original experimental design under a no-effect null model. The card and assignment examples below separate these two processes.

A teacher prepares four equally accessible cards marked 2, 4, 6, and 8. A student draws a card, records its value, replaces it, mixes the cards, and repeats the draw. The mean of the two values is one statistic from one sample. Drawing 2 and then 8 produces a mean of 5. If a classmate obtains 4 and then 6, that sample also contributes a dot at 5 on a plot of sample means.

PR22 | Continued

The cards themselves form a population of four values with mean 5. The distribution of means from all possible two-draw samples describes how the sample mean varies under this sampling process. Different samples can produce the same mean. The height above 5 therefore depends on how many draw sequences give that value. With four possible cards per draw there are $4 \times 4 = 16$ equally likely ordered pairs. Four pairs have a mean of 5: (2,8), (8,2), (4,6), and (6,4). Thus the exact chance of a sample mean of 5 is $4/16$. A score of 5 never appears on a card, yet it appears repeatedly on the sampling plot.

A different simulation starts with outcomes already recorded in an experiment. Suppose the experiment randomly assigned students to two groups of equal size. Under a claim that assignment changed no student's outcome, keep each recorded outcome fixed, shuffle the group assignments according to the design, and recalculate the difference in group means. This produces a randomization distribution for that difference.

For example, suppose four recorded scores are 3, 5, 7, and 9. One reassignment puts 3 and 9 in treatment and 5 and 7 in control. This gives one difference, $6 - 6 = 0$. Another puts 7 and 9 in treatment and 3 and 5 in control, giving $8 - 4 = 4$. The scores stay fixed. Different assignments change which scores enter each average.

Two samples drawn with replacement

Sample	First card	Second card
First student	2	8
Classmate	4	6

Source and adaptation: OpenStax Chapter 7 introduction and §7.1, printed pp. 365-369 (PDF pp. 377-381). Repeated-sample definitions adapted. Randomization-distribution comparison is an original AP CED 2.12/3.1 bridge.

TWO READING CHECKS

1. Start a second plot for samples of four draws with replacement. A sample yields 2, 2, 2, and 6. Where does its dot go on this new plot? How does that value compare with the population mean?

2. Use the four fixed scores 3, 5, 7, and 9. A shuffle puts 5 and 9 in treatment and 3 and 7 in control. What difference (treatment minus control) would you plot? What does each shuffle preserve?

PR23 | Sampling distributions of proportions and their differences

Why a class survey might get 57% twice

PRACTICE FROM THE LAST MEETING

S05.01-HW (W111). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A sample proportion $\hat{p}$ is the number of successes divided by the sample size. Its population counterpart p is the true proportion of successes in the population. For samples of n independent observations, each with the same success probability p, $\hat{p}$ has mean p and standard deviation $\sqrt{[p(1-p)/n]}$. These are properties of a distribution of sample proportions. They do not describe variation among individual yes/no responses, nor do they imply that every sample proportion equals p.

OpenStax's binomial homework example assumes 50 independent students, each with probability 0.70 of submitting on time. The count X has mean $50(0.70)=35$ and standard deviation $\sqrt{[50(0.70)(0.30)]}\approx 3.240$. Divide the count by 50 to obtain the sample proportion. Its mean is 0.70 and its standard deviation is $3.240/50\approx 0.0648$. Rescaling the count explains why the proportion formula divides by the square root of sample size.

For a simple random sample without replacement, check that the sample is at most 10% of its population before using the independence-based standard deviation as an approximation. A normal approximation also requires enough expected successes and failures. Check $np \geq 10$ and $n(1-p) \geq 10$ using the assumed population proportion. Center and spread alone do not establish a normal shape.

Imagine a large student population in which exactly 60% would choose an early lunch. We know this population proportion only because we have specified it for a model. A teacher draws a simple random sample of 100 students and records the sample proportion choosing early lunch. If 57 choose it, that sample's proportion is 0.57. A second sample might also produce 0.57, or it might give another value.

Picture a dot for the proportion from each possible random sample of the same size. The distribution of these proportions is centered at 0.60. Increasing the sample size tends to reduce the spread. Here the standard deviation of sample proportions is about 0.049 for samples of 100 and about 0.0245 for samples of 400, assuming each sample is a small fraction of the population.

PR23 | Continued

Suppose the early-lunch proportions at two large schools are 0.60 and 0.40. Take independent simple random samples of 100, each smaller than 10% of its school. Let D be the first sample proportion minus the second. Across pairs of samples its mean is 0.60 − 0.40 = 0.20, or 20 percentage points. Each sample proportion has variance about 0.0024. If the first sample proportion runs high and the second runs low, both errors push D upward. Independent sources of variation add as variances even when we subtract the proportions.

The standard deviation of D is approximately $\sqrt{0.60(0.40)/100 + 0.40(0.60)/100} \approx 0.0693$, or 6.93 percentage points. Each school has at least 10 expected successes and 10 expected failures (60 and 40 at the first school, 40 and 60 at the second), so a normal approximation is reasonable. For instance, a difference below 0.10 would be about $(0.10 - 0.20)/0.0693 \approx -1.44$ standard deviations from the mean. The area below that boundary is about 0.075. Even with a true gap of 20 percentage points, about 7.5% of such sample pairs would show a gap under 10 points in that order.

Same lunch population, different sample sizes

Sample size	Center of $\hat{p}$	SD of $\hat{p}$
100	0.60	About 0.0490
400	0.60	About 0.0245

Source and adaptation: OpenStax §4.3 Example 4.12, printed pp. 238-239 (PDF pp. 250-251), and §8.3, pp. 420-425 (PDF 432-437). Count-to-proportion derivation and two-population difference are original bridges for AP CED 3.1/3.9.

TWO READING CHECKS

1. A fresh sample contains 80 students, and 52 choose early lunch. What sample proportion does this sample contribute to the plot? Show the fraction.

2. At two other schools, the early-lunch proportions are 0.65 and 0.50. Independent random samples give sample proportions with SDs of 0.03 and 0.04. For the first school minus the second, find the center and SD in percentage points.

PR24 | Sample means, differences in means, and the CLT

What varies when you average wait times?

PRACTICE FROM THE LAST MEETING

S05.02-HW (W117). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

For independent observations from a population with mean μ and standard deviation σ, the sample mean $\bar{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$. Averaging does not shift the target center. It reduces the spread of the statistic because unusually high and low observations can offset one another. The standard deviation of a statistic's sampling distribution is often called its standard error, although later we must distinguish a known standard deviation from an estimate based on sample data.

OpenStax models soccer-match durations as normal with mean 2 hours and standard deviation 0.5 hour. For a random sample of 50 independent match times, the sample mean is normal with mean 2 hours and standard deviation $0.5/\sqrt{50} \approx 0.0707$ hour. A probability about this average must use 0.0707, not the 0.5-hour spread of individual matches. Because the population model is normal, its sample mean is normal for any sample size under the independence assumptions.

When the population is not normal, the central limit theorem provides an approximately normal sampling distribution for the sample mean when n is sufficiently large under suitable conditions. The amount of data needed depends on skewness and unusual observations; the theorem does not make the original population normal. For two independent sample means, define the subtraction order, subtract the population means for the center, and add the two sampling variances for the variance. A larger total sample cannot repair a seriously inadequate sample in one group.

Suppose a transit agency's population of wait times has a mean of 12 minutes and a standard deviation of 6 minutes. A few very long waits make the distribution right-skewed. Consider one passenger's wait and then the average of 36 independently selected waits. Both quantities use minutes, but a single unusually long wait has a much smaller effect when it is averaged with 35 other waits.

Across many such samples, the mean of the sample averages is 12 minutes. The standard deviation of those averages is 6 divided by the square root of 36, or 1 minute. Each point on a plot of the averages represents an entire sample. A plot made from individual waiting times answers a different question about the spread of passengers' experiences.

PR24 | Continued

The central limit theorem describes how the distribution of sample means approaches a normal shape as sample size grows for independent draws from the same population with finite variance. Strong skew or extreme values can require larger samples for a good approximation. The underlying population of waits keeps its original shape.

Now compare Route A, the route above, with Route B, where the mean wait is 9 minutes and the standard deviation is 6 minutes. Take independent random samples of 36 waits on A and 64 on B, with independent waits within each sample. The sample means have standard deviations $6/\sqrt{36} = 1$ minute and $6/\sqrt{64} = 0.75$ minute. For the difference A minus B, the center is $12 - 9 = 3$ minutes. The sampling variances add, so the difference has standard deviation $\sqrt{1^2 + 0.75^2} = 1.25$ minutes. Reversing the route order changes the center to −3 minutes, but leaves the spread at 1.25 minutes. If each day supplies a matched pair of waits, analyze the within-day differences to account for the pairing.

Do not use individual spread for a mean

Quantity	Mean	SD
One modeled match duration	2 hours	0.5 hour
Mean of 50 independent durations	2 hours	About 0.0707 hour

Source and adaptation: OpenStax §§7.1 and 7.3, printed pp. 366-369 and 373-380 (PDF pp. 378-381 and 385-392), especially Example 7.2. CLT/standard-error explanation adapted; independent-mean difference is AP CED 4.6 bridge.

TWO READING CHECKS

1. Which quantity in the example has standard deviation 1 minute?

2. For two other routes, independent samples give sample means with the same SD, 2 minutes each. A classmate says the difference of the means must have SD zero. Calculate its SD and explain your result.

PR25 | What a confidence interval says

Which district proportions fit the bus survey?

PRACTICE FROM THE LAST MEETING

S05.03-HW (W121). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A point estimate uses a single sample statistic to estimate an unknown population parameter. An interval estimate adds information about uncertainty by reporting a range of plausible parameter values. Its endpoints depend on the sample, so another sample can produce a different interval. A confidence level describes the long-run success rate of the interval-producing procedure under its assumptions. The target population parameter is fixed during that repeated-sampling description.

OpenStax's smartphone example surveys 500 randomly selected adults in a large city and finds 421 owners. The sample proportion is 421/500=0.842. Its reported 95% confidence interval is approximately (0.810,0.874). A contextual interpretation is: we are 95% confident that between 81.0% and 87.4% of all adults in this city own smartphones. The interval estimates a population proportion. It is not a range containing 95% of the individual responses, which here are only yes or no.

To interpret the confidence level, imagine repeating the sampling procedure with the same city population. With each new sample, the sample proportion changes, so the endpoints can change too. If the method has 95% coverage, about 95% of the resulting intervals will contain the city's true ownership proportion in the long run. An individual interval either covers that value or misses it. People omitted from the sampling list would be omitted on every repetition. The confidence level cannot correct that bias.

Suppose a school randomly selects 100 students from a large district and asks whether they would use a later activity bus. In this sample, 58 say yes. The student council can report 58% as a point estimate of support among all district students. If the council drew a second random sample, some different students would answer the question, and the estimate would probably change.

PR25 | Continued

An interval estimate gives the council a way to express this uncertainty. Suppose an appropriate 95% confidence procedure produces an interval from 48% to 68%. The center is 58%, and the margin of error is 10 percentage points on either side. Values inside this interval are plausible population proportions in light of the sample. Both 49% and 60% are inside, so this sample leaves the majority question open.

You can see the distinction in a simulation because you know the population value you chose. Set a simulated population's support at 55%. One simulated sample gives an interval from 48% to 68%, which captures 55%. Another gives 56% to 74%, which misses. Across many samples, count how many intervals capture 55% and divide by the number of samples. In an actual district survey the population value is unknown, so the council reports the interval and its confidence level. It cannot check whether its particular interval captured the truth.

Three numbers, three different meanings

Quantity	Bus-study value
Sample proportion	58%
Interval for the population proportion	48%-68%
Long-run confidence level of the method	95%

Source and adaptation: OpenStax Chapter 8 introduction and §8.3 Example 8.10, printed pp. 405-407 and 421-422 (PDF pp. 417-419 and 433-434). Confidence definitions and source interval adapted; district-bus case is original.

TWO READING CHECKS

1. Which percentage describes the sample, and whose opinions does the interval estimate?

2. Explain why reporting a majority as established would go beyond these data.

PR26 | Build and justify a one-proportion interval

What makes an interval defensible?

PRACTICE FROM THE LAST MEETING

S06.01-HW (W129). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

Suppose each sampled adult answers a yes-or-no question. Define p as the proportion of the population who would answer yes, n as the sample size, and x as the number of yes responses. The point estimate is the sample proportion, $\hat{p} = x/n$. Across random samples, that estimate varies. Its estimated standard error, $\sqrt{\hat{p}(1-\hat{p})/n}$, describes the typical scale of that variation. A one-sample z interval extends a margin of error in either direction from the estimate. For 95% confidence, that margin is approximately 1.96 times the standard error.

Using OpenStax's 421 smartphone owners out of 500 adults, $\hat{p}=0.842$ and the estimated standard error is $\sqrt{[(0.842)(0.158)/500]} \approx 0.0163$. The 95% margin of error is about $1.96(0.0163)=0.0320$, giving (0.810,0.874). The sample contains 421 observed successes and 79 observed failures, both at least 10. A random sample from a city population of at least 5,000 also satisfies the usual 10% condition for sampling without replacement. A calculation alone does not verify that the sampling process was random.

Write each condition with the evidence supporting it. For a confidence interval, use observed successes x and observed failures n −x for the large-counts check. For independence when sampling without replacement, compare n with 10% of the population size. A large n cannot repair voluntary response or a missing part of the population. After calculating, state the confidence level, the population, the success definition, and the endpoints. The next case gives you practice distinguishing the design check from the count check.

Imagine that a town randomly selects 80 of its 4,000 adult residents without replacement. Of the 80 selected, 52 favor moving a bus stop and 28 prefer its current location. The fraction 52/80 estimates the share of all adult residents who favor the move. Before using the usual one-proportion interval, check how residents entered the study and what kinds of outcomes appeared in the sample.

PR26 | Continued

Check who had a chance to enter the sample and how they were selected. A random sample from a complete town list supports inference to adult residents. With sampling without replacement, the 10% condition lets us treat the selections as approximately independent. A separate check asks whether the normal approximation is reasonable. Count the observed yes and no responses and compare each count with 10. A random sample can pass one of these checks and fail another.

For the bus-stop survey, the point estimate is 52/80 = 0.65. The estimated standard error is about 0.0533, so the 95% margin is about 1.96 × 0.0533 = 0.105. The town can report that it is 95% confident that about 55% to 75% of all its adult residents favor the move. If the town had posted an open invitation online, the same result of 52 out of 80 would give exactly the same arithmetic. People with strong opinions might then have been more likely to respond. The design determines whether the town can defend that conclusion.

Observed counts for the bus-stop interval

Outcome	Count
Favor moving the stop	52
Favor current location	28
Sample total	80

Source and adaptation: OpenStax §8.3 Example 8.10, printed pp. 420-422 (PDF pp. 432-434). Wald proportion interval adapted with AP observed-count threshold 10 and explicit 10% check; plus-four method intentionally excluded.

TWO READING CHECKS

1. Identify the two counts to compare with 10 in this survey.

2. Show the population-size check in one line and state why it matters.

PR27 | Plan precision before collecting data

How many completed responses will the study need?

PRACTICE FROM THE LAST MEETING

S06.02-HW (W137). Complete 1, 3, 4, 6.

READ FOR THE COMING MEETING

Planning sample size starts with a desired margin of error and confidence level. For a population proportion, the planned margin is approximately $z^*\sqrt{[p^*(1-p^*)/n]}$, where p^* is a preliminary estimate used before collecting the new sample. Solving for n gives $n \geq (z^*/m)^2 p^*(1-p^*)$, where m is the desired margin expressed as a proportion. Round the result up to the next whole observation so that the planned sample is not too small.

OpenStax considers a phone company estimating the share of its customers aged 50 and older who use text messaging. It wants 90% confidence and a margin of three percentage points. With no preliminary estimate, use a planning proportion of 0.50. The product 0.50 × 0.50 is 0.25, compared with 0.60 × 0.40 = 0.24 or 0.80 × 0.20 = 0.16. Since the product reaches its maximum at 0.50, that choice gives enough observations across the full range of possible proportions. Using the 90% critical value 1.645 and m = 0.03 gives $n \geq (1.645/0.03)^2(0.25) \approx 751.7$. The company needs at least 752 completed responses.

At a fixed confidence level and similar proportion, reducing the margin of error by half requires about four times as many observations. Increasing confidence widens the interval unless the sample size also increases. These relationships involve sampling uncertainty under a suitable design. Collecting a much larger convenience sample can make a biased estimate appear precise without making it accurate. Also plan for nonresponse: the number invited may need to exceed the number of usable responses required. The survey below separates sample-size planning from improvements to the sampling method.

A trail organization wants a precise estimate of the proportion of local households that would use a proposed trailhead. A preliminary plan calls for 200 randomly selected households and gives a projected margin of error of eight percentage points. The board asks how many completed responses it would need for a margin of four points at the same confidence level.

PR27 | Continued

The eight-point margin describes one side of the interval. At a fixed center, switching from an eight-point margin to a four-point margin cuts its full width from 16 percentage points to eight. Keep the confidence level and planning proportion fixed as you compare sample sizes. A proposal to double the 200 responses would divide the eight-point margin by the square root of two, giving about 5.7 points. Use that relationship to work out the sample size the board actually needs.

Suppose the board also asks for greater confidence. At 90% confidence, the critical value is about 1.645. At 95% it is about 1.96. At the same sample size and planning proportion, the larger critical value gives a larger margin. A higher-confidence study with the same target margin therefore needs more completed responses. If turnout is low, sending extra invitations may help meet the count. If only regular trail users receive invitations, however, adding responses from that list still leaves other households out.

Planning the trailhead survey

Plan	Completed responses	Projected margin
Preliminary	200	8 percentage points
Goal	To determine	4 percentage points

Source and adaptation: OpenStax §8.3 sample-size discussion and Example 8.14, printed pp. 425-427 (PDF pp. 437-439). Formula and numerical planning example adapted; AP interpretation of precision retained.

TWO READING CHECKS

1. What sample size would cut the projected margin from eight points to four? Explain your multiplier.

2. Why do planners use 0.50 when the population proportion is unknown?

PR28 | Estimate a difference in population proportions

Choose the subtraction order first

PRACTICE FROM THE LAST MEETING

S06.03-HW (W141). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

Two towns can differ in their support for the same proposal. Name the success event and define both populations before subtracting. Let p_1 and p_2 be their true support proportions. If you choose the parameter $p_1 - p_2$, its sample estimate is the first sample proportion minus the second. Either sample could change on another draw. For independent samples, the variances add even when you subtract the estimates. The estimated standard error of the difference is $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$, where n_1 and n_2 are the two sample sizes.

OpenStax compares two medication groups with 20 of 200 and 12 of 200 adults still having hives after 30 minutes. The observed proportions are 0.10 and 0.06, so the first-minus-second estimate is 0.04, or 4 percentage points. For an interval, estimate each group's variability separately. The observed success/failure counts are 20/180 and 12/188. Before making a population claim, also verify how the groups were selected, that their observations are independent, and the relevant sampling fractions.

A confidence interval entirely above 0 supports a positive population difference in the stated order; one entirely below 0 supports a negative difference. If the interval includes 0, the data are compatible with no difference as well as the other values in that interval. It does not prove the proportions are equal. Proportions and percentage points also differ: a difference of 0.04 is 4 percentage points, not necessarily a 4% relative increase. The local example below keeps these interpretation details explicit.

Suppose independent random samples ask two large towns whether they favor an evening bus route. In North, 66 of 100 sampled residents favor it. In South, 60 of 120 favor it. Before comparing the towns, define the parameter in words. Let pN be the proportion of all North residents who favor the route and pS the corresponding proportion in South. Choose pN − pS as the difference to estimate.

PR28 | Continued

Suppose each town has at least 2,000 residents. Each random sample is then at most 10% of its town. The two samples are independent of each other, and each has at least 10 observed supporters and 10 observed nonsupporters. The North-minus-South estimate is 0.66 − 0.50 = 0.16. Its estimated standard error is $\sqrt{0.66(0.34)/100 + 0.50(0.50)/120} \approx 0.0658$. The observed difference is 16 percentage points, while the typical sampling variation in that difference is about 6.6 percentage points.

The 95% margin is 1.96 × 0.0658, or about 0.129. Extending that far on either side of 0.16 gives (0.031, 0.289). We are 95% confident that support for the evening bus among all North residents is about 3 to 29 percentage points higher than among all South residents. To report South minus North, reverse the sign of each endpoint and put the smaller number first. The direction of the conclusion in words should still match the survey.

Evening bus survey counts

Town	Residents sampled	Favor the route
North	100	66
South	120	60

Source and adaptation: OpenStax §10.3, Example 10.8, printed pp. 523-524 (PDF pp. 535-536). Medication counts adapted from the source test. The two-proportion confidence interval follows AP Statistics CED topics 3.10-3.11.

TWO READING CHECKS

1. Find the estimate for South minus North. Include units.

2. List the observed successes and failures in each town.

PR29 | Checkpoint 3

Know which quantity varies

PRACTICE FROM THE LAST MEETING

S06.04-HW (W147). Complete 1, 2, 4, 6.

READ FOR THE COMING MEETING

Imagine repeatedly drawing random samples of 100 households from one large town and recording the proportion that compost. One sample might yield 0.48 and the next 0.55. Each result estimates the same population proportion. A dotplot with one dot per sample proportion describes the variation in those estimates across samples.

A confidence interval uses one observed sample to estimate the population proportion. Its width depends partly on how much sample proportions of this size typically vary. For a large-sample interval, the estimated standard error uses the observed sample proportion and the sample size. The study design and observed counts matter when deciding whether that interval is appropriate.

Suppose a simple random sample of 100 from a town of 10,000 households includes 60 that compost. There are 60 successes and 40 failures, both at least 10, and the sample is under 10% of the town. The estimated standard error is $\sqrt{(0.60 \times 0.40/100)}$, about 0.049. For a 95% interval the margin is about $1.96 \times 0.049 = 0.096$. We are 95% confident that roughly 50% to 70% of households in this town compost. The interval estimates one proportion for the town. It does not predict the next household's yes-or-no response.

If the researchers switch to samples of 400 households drawn in the same manner, sample proportions will generally fluctuate less. An equally large voluntary response drawn from a town newsletter can still miss households that differ from the people who reply.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S05.02 (R084); S06.02 (R099).

TWO READING CHECKS

1. What does one dot represent in the proposed dotplot?

2. What quantity remains fixed across samples from this town at this time?

PR30 | Hypotheses, null models, and p-values

What counts as surprising under a claim?

PRACTICE FROM THE LAST MEETING

(1) Two independent SRSs from a large town find 52 of 100 and 208 of 400 households use a service. Compute both proportions and estimated standard errors. (2) Approximate both 95% margins using 1.96. (3) Explain the difference and identify a design flaw that more observations alone cannot repair.

READ FOR THE COMING MEETING

A significance test evaluates how strongly sample data conflict with a stated claim about a population parameter. The null hypothesis H_0 gives the reference value or equality used to form a chance model. The alternative hypothesis H_a gives the direction or kind of departure being investigated. Hypotheses describe population parameters, not the sample statistics already observed. Choose the alternative from the research question before inspecting which direction the data happen to favor.

OpenStax asks whether a population mean GPA differs from 2.0. The hypotheses are H_0: $\mu=2.0$ and H_a: $\mu\neq2.0$, where μ is the mean GPA in the defined population. “Differs” calls for a two-sided alternative. A question about a mean below 2.0 would instead use $\mu<2.0$ and count evidence in the lower tail. For AP work, write the null with equality at the reference value, even when a broad verbal claim includes an inequality.

A p-value is the probability, assuming the null hypothesis and model conditions are true, of obtaining a statistic at least as extreme as the observed one in the direction specified by the alternative. A small p-value indicates that the observed result would be unusual under that null model. It is not the probability that H_0 is true and not the probability the result occurred “by chance.” Compare it with a significance level selected in advance. A conclusion must connect that comparison to the population claim in the question.

A game designer says a spinner lands on blue with probability 0.25. A class suspects blue occurs more often. Let p be the probability of blue on one spin under the same conditions. The class writes H_0: $p = 0.25$ and H_a: $p > 0.25$. Before collecting data, the class selects a significance level of 0.05. Blue appears 15 times in 40 independent spins. If the designer's claim holds, the expected count is $40 \times 0.25 = 10$. The class needs to decide how unusual a count of 15 or more would be.

PR30 | Continued

The null hypothesis supplies a reference model with blue probability 0.25. To simulate one trial, generate 40 independent digits, each equally likely to be 1, 2, 3, or 4, and count the 1s as blue. Repeat that trial many times. Since the class suspected an increase, count simulated trials with at least 15 blues. The fraction of such trials estimates the p-value. The phrase 'at least' includes outcomes even further in the suspected direction.

Suppose 1,000 simulated sets of 40 spins give exactly 15 blues in 29 sets and more than 15 in 25 sets. Both groups meet the upper-tail rule. The simulation estimate is (29 + 25)/1,000 = 0.054. The denominator counts whole sets of 40 spins. If the true blue probability is 0.25, about 5.4% of sets in this simulation produce at least as many blues as the class observed. The estimate exceeds 0.05, so this run gives insufficient evidence of a higher rate at the chosen level. Since it falls close to the cutoff, a larger simulation would help check the decision.

Set up one simulated spin

Generated digit	Simulated outcome
1	Blue
2, 3, or 4	Other color

Source and adaptation: OpenStax §§9.1 and 9.4-9.5, printed pp. 462-463 and 467-470 (PDF pp. 474-475 and 479-482), especially Example 9.2. Hypothesis and p-value definitions adapted to AP equality-null conventions.

TWO READING CHECKS

1. Why is the null simulation built with probability 0.25?

2. Would a simulated count of 16 qualify as extreme? Explain using the class's question.

PR31 | Carry out a one-proportion z test

Compare the observed count to the null model

PRACTICE FROM THE LAST MEETING

S07.01-HW (W153). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A one-proportion z test compares a sample proportion $\hat{p}$ with a hypothesized population proportion p_0. The test statistic is $z=(\hat{p}-p_0)/\sqrt{[p_0(1-p_0)/n]}$. The denominator uses the null value because the test asks how the statistic would vary if that null model were true. In contrast, a one-proportion confidence interval estimates variability using the sample proportion. These two standard errors answer different questions.

OpenStax's household example observes 43 of 150 households with three cell phones and tests $p_0=0.30$ against a two-sided alternative. The sample proportion is $43/150\approx0.2867$. Under the null, expected successes are $150(0.30)=45$ and expected failures are $150(0.70)=105$, so both meet the AP threshold of 10. The standard error under H_0 is $\sqrt{[(0.30)(0.70)/150]}\approx0.0374$. The observed difference is only about -0.356 standard errors, giving a two-sided p-value about 0.7216.

At significance level 0.05, this p-value leads to failure to reject H_0. The sample does not provide convincing evidence that the household proportion differs from 0.30. That statement does not establish that the proportion is exactly 0.30. A complete solution defines p, states both hypotheses, verifies random/independent sampling and expected counts, shows the statistic or equivalent probability calculation, and gives a contextual conclusion. The supplied counts cannot establish the sampling design by themselves.

A town's planning benchmark is that 25% of residents use its library in a typical week. Before sampling, a researcher asks whether current use exceeds that benchmark and chooses a significance level of 0.05. Let p be the proportion of all town residents who use the library in a typical week. She tests H_0: $p = 0.25$ against H_a: $p > 0.25$. In a simple random sample of 120 from 10,000 residents, she finds 38 weekly users. The observed share is $38/120$, about 31.7%.

PR31 | Continued

Under that null model, the expected counts are 120 × 0.25 = 30 users and 120 × 0.75 = 90 nonusers. Both are at least 10, supporting the normal approximation. The sample is also less than 10% of the population. The reference variability for this test uses the claimed proportion, because the calculation describes the sampling distribution when the null model holds.

The standard error under the null is $\sqrt{0.25(0.75)/120} \approx 0.0395$. On the proportion scale, the observed gap from the benchmark is about 0.0667. Dividing by 0.0395 gives z ≈ 1.69. The researcher's question asks about an increase, so she finds the standard normal area to the right of the unrounded z statistic, about 0.046. This falls below 0.05. The sample provides evidence that more than 25% of town residents use the library in a typical week. If the sample proportion had instead fallen below the benchmark, the original direction of the question would still determine which tail to use.

Expected counts under the library-use null p = 0.25

Null-model outcome	Calculation	Expected count
Weekly user	120(0.25)	30
Not a weekly user	120(0.75)	90

Source and adaptation: OpenStax §§9.3-9.5, printed pp. 466-478 (PDF pp. 478-490), especially Example 9.18. One-proportion test adapted to AP null-count condition and equality notation.

TWO READING CHECKS

1. Which two counts support the normal approximation here, and how were they obtained?

2. Why is a sample count of 42 included among results at least as extreme as 38?

PR32 | Errors, power, and consequences

Two kinds of wrong decisions

PRACTICE FROM THE LAST MEETING

S07.02-HW (W159). Complete 1, 2, 4, 5.

READ FOR THE COMING MEETING

A significance test makes a decision from imperfect information. A Type I error rejects a true null hypothesis. A Type II error fails to reject a false null hypothesis. Describe each error by combining the decision with the true population situation; naming only an undesirable action does not distinguish them. The significance level α controls the Type I error probability for the stated null model, subject to the test's approximation and conditions.

OpenStax illustrates the distinction with climbing equipment. Suppose the null states that a piece of equipment is safe. A false alarm leads a climber to replace safe equipment. A missed warning lets unsafe equipment remain in use. In each case, identify both the actual condition and the decision. Knowing only that equipment was replaced tells you nothing about whether the decision was an error.

Power is the probability of rejecting H_0 for a specified true value in the alternative. It equals 1 minus the Type II error probability for that value. There is no single power value without specifying the effect to detect. With sample size and true effect fixed, lowering α generally reduces both false alarms and power. A larger well-designed sample can increase power while holding α fixed. Larger effects relative to variability are easier to detect. Evaluate the consequences in the actual decision setting, and distinguish statistical significance from practical importance.

A seed supplier promises that 80% of its seeds germinate. A grower tests that claim against the possibility that the true rate is lower. Two different mistakes matter to the grower. If the rate really is 80% and the test leads her to reject the claim, she might discard a satisfactory lot. If the rate is lower and the test fails to detect the shortfall, she might plant a poor lot.

PR32 | Continued

The first decision is a Type I error. It occurs when the null hypothesis is true and the test rejects it. The second is a Type II error. To quantify its chance, choose a particular alternative germination rate. Suppose repeated tests with a true rate of 70% detect the shortfall 85% of the time. The test then has power 0.85 at 70%, and the probability of a Type II error there is 0.15.

Picture the grower repeating the same test on 200 simulated lots at a chosen true germination rate. For each lot, she records whether the test detects a shortfall. The number of detections divided by 200 estimates power at that rate. Now move the true rate closer to 80%, keeping the sample size and decision rule the same. The possible results overlap more with results from satisfactory lots, making a slight shortfall harder to distinguish. If the grower makes the rejection rule stricter, fewer satisfactory lots are discarded, but more deficient lots can escape detection. Testing a larger random sample of seeds can improve detection at the same significance level.

Seed decision: null rate 80%, alternative below 80%

True condition	Reject the null	Fail to reject
Rate is 80%	Type I error	Correct decision
Rate is below 80%	Detect the shortfall	Type II error

Source and adaptation: OpenStax §9.2, printed pp. 464-465 (PDF pp. 476-477), especially Example 9.5. Error definitions adapted; parameter-based power and trade-offs expanded for AP CED.

TWO READING CHECKS

1. Describe a Type I error in the seed example.

2. At a true rate of 70%, what fraction of repeated tests would miss the shortfall in the example?

PR33 | Test a difference in population proportions

What would equal proportions predict?

PRACTICE FROM THE LAST MEETING

S07.03-HW (W163). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A two-proportion z test can assess H_0: $p_1=p_2$ against a directional or two-sided alternative. Under that null, both populations share a common success proportion. Estimate that common proportion by pooling the successes and sample sizes: $\hat{p}c=(x_1+x_2)/(n_1+n_2)$. The null standard error is $\sqrt{[\hat{p}c(1-\hat{p}c)(1/n_1+1/n_2)]}$. The test statistic compares the observed difference $\hat{p}_1-\hat{p}_2$ with the null difference of 0.

In OpenStax's medication example, 20 of 200 patients in group A and 12 of 200 in group B still have hives after 30 minutes. The pooled estimate is $32/400=0.08$. Under equal population proportions, each group of 200 would have 16 expected successes and 184 expected failures. The observed difference $0.10-0.06=0.04$ is about 1.474 null standard errors above 0. A two-sided test gives $p\approx0.1404$, which does not provide convincing evidence of a difference at $\alpha=0.01$.

Pooling is justified for this equality-null test because H_0 specifies one common proportion. It is not the standard error used to estimate an unrestricted difference in a confidence interval. Check the sampling or assignment process, independence between groups, independence within groups, and the null expected counts. State the difference in the defined order. Statistical evidence of an association does not alone show that a medication caused it; a causal interpretation needs a defensible randomized experimental design.

Imagine researchers randomly assign 100 individual volunteers to receive one of two reminder messages, 30 to message A and 70 to message B. In Group A, 20 of 30 volunteers send a confirmation. In Group B, 20 of 70 send one. The counts of confirmations match, but the observed proportions differ, about 67% and 29%. A test of equal success probabilities asks how large a gap might arise from chance if both messages had the same response probability.

PR33 | Continued

Under the equality claim, that shared probability is unknown. Estimate it by counting all the successes and dividing by the total number of volunteers. Here the pooled estimate is 40/100 = 0.40. Pooling this way gives each volunteer one observation. Simply averaging the two percentages would give the 30-person group as much weight as the 70-person group.

Under the common-probability model, Group A's expected confirmation count is 30 × 0.40 = 12. Group B's is 70 × 0.40 = 28. The same pooled estimate gives a null standard error of $\sqrt{0.40(0.60)(1/30+1/70)} \approx 0.107$. The observed difference, 20/30 − 20/70 ≈ 0.381, lies about 3.56 standard errors above zero. For a prespecified two-sided alternative, the p-value is about 0.00037. At a significance level of 0.05, this is convincing evidence that the message affects confirmations among these volunteers. Random assignment supports that causal interpretation. Generalizing to everyone who might receive a message would also require evidence that the volunteers represent them.

Confirmations by reminder message

Message	Volunteers	Confirmed
A	30	20
B	70	20

Source and adaptation: OpenStax §10.3 Example 10.8, printed pp. 523-525 (PDF pp. 535-537). Source counts and pooled-test reasoning adapted; causal limits made explicit.

TWO READING CHECKS

1. How would you calculate the pooled estimate from the two groups?

2. Give the expected numbers of nonconfirmations in A and B under equality.

PR34 | Compare categorical distributions with chi-square

Compare distributions across two towns

PRACTICE FROM THE LAST MEETING

S07.04-HW (W169). Complete 1, 2, 4, 6.

READ FOR THE COMING MEETING

A chi-square statistic compares observed counts with counts expected under a null model. For each category or cell, calculate $(\text{observed} - \text{expected})^2/\text{expected}$ and add the contributions. Squaring makes both positive and negative discrepancies increase the total. A large statistic indicates more disagreement with the null model, so the p-value is an area to the right on a chi-square distribution. Use counts, not percentages, in this calculation.

A test for homogeneity begins with independent samples from different populations or groups formed by random assignment. One categorical response is recorded for each person. If the populations share one response distribution, a category's overall share should apply to each group's total. Multiplying a group total by that share gives its expected count in the category. In table form, the calculation is (row total × column total)/grand total. An expected count can be fractional.

In an OpenStax example, separate random samples of 250 male and 300 female college students report one of four living arrangements. With two rows and four columns, the reference distribution has $(2 - 1)(4 - 1) = 3$ degrees of freedom. From the full table, $\chi^2 \approx 10.13$ gives a right-tail probability of about 0.0175. If the two population distributions were the same, about 1.75% of studies satisfying the model would give a statistic at least this large. At the stated 0.05 level, the data provide evidence that the distributions of living arrangements differ. To describe where they differ, inspect the category percentages within each group.

Suppose two large towns want to compare residents' preferred methods of paying for a bus ride. Each person chooses a card, phone, or cash. Researchers take independent random samples of 80 residents in West and 120 in East. Half of the 200 respondents combined prefer a card. What would the counts look like if the two towns shared the same distribution of preferences?

PR34 | Continued

West's expected card count is 80 × 0.50 = 40. East's is 120 × 0.50 = 60. Suppose 48 people in West actually chose a card. That cell alone contributes $(48 - 40)^2/40 = 1.60$ to the statistic. The observed count exceeds the model's prediction, so cards were more popular in the West sample than the common-distribution model predicts. A deficit of eight in that cell would make the same numerical contribution. Inspect the observed and expected counts to recover the direction.

Add the contributions from all six cells before making an overall decision. The two-by-three table has $(2 - 1)(3 - 1) = 2$ degrees of freedom. For this AP procedure, every expected cell count must exceed five. Check those counts, the independent random samples and, for sampling without replacement, that each sample is at most 10% of its town. Evidence against a common distribution would mean at least one category's population share differs. Individual category differences need further examination.

Two independent transit surveys

Town	Residents sampled	Response options
West	80	Card, phone, cash
East	120	Card, phone, cash

Source and adaptation: OpenStax §11.4, Example 11.8, printed pp. 578-581 (PDF pp. 590-593). College living-arrangement test adapted. For expected counts, the AP Statistics CED specifies greater than five in essential knowledge 3.14.D.1.iii.

TWO READING CHECKS

1. Why does East have a larger expected card count than West under the common-distribution model?

2. If the combined phone share is 0.30, what phone counts does the model predict in West and East?

PR35 | Investigate association with chi-square

Look for an association within one population

PRACTICE FROM THE LAST MEETING

S07.05-HW (W173). Complete 1, 2, 4, 5, 6.

READ FOR THE COMING MEETING

A chi-square test of independence examines two categorical variables measured on each individual in one sample. The null states that the variables are independent in the population; the alternative states that they are associated. This differs in design from a homogeneity test, which begins with separate population samples or treatment groups and compares one response distribution. The expected-count calculation and chi-square statistic can be identical even though the research designs differ.

OpenStax's driving example has 755 respondents, 70 reporting a speeding violation and 305 reporting cell-phone use while driving. Under independence, the expected number reporting both is $(70\times305)/755\approx28.28$. This is a model-based expected count, not necessarily a whole number and not a reported number of actual drivers. The formula for every cell is row total×column total/grand total. Use the observed cell counts, rather than these totals alone, to calculate the full statistic.

For an r-by-c table, the chi-square reference distribution has $(r-1)(c-1)$ degrees of freedom. Check that expected counts, not observed counts, are greater than 5 in every cell, and justify random sampling and independence of observations. A significant result supports an association in the population represented by the design. It does not measure the strength of the relationship or establish a cause. Examine conditional distributions and the largest contributions to describe how the observed pattern departs from independence.

A district takes a simple random sample of 200 students from 8,000. For each student the researcher records whether they live near a bike path and how they usually travel to school. For the analysis she groups travel mode as bicycle or other transportation. These are two categorical variables recorded for every student in one sample. The null hypothesis says that proximity to a path and usual travel mode are independent among the district's students. The alternative is that the two variables are associated.

PR35 | Continued

Suppose 60 of the 200 students usually bike, and 80 live near a path. If the variables are independent, the proportion who bike should be the same among students near a path as among all students. The expected number who both bike and live near a path is therefore $80 \times (60/200) = 24$. Suppose the actual count is 36. That cell would have 12 more students than the independence model predicts.

Now compare percentages the students actually reported. Of the 80 who live near a path, $36/80 = 45\%$ usually bike. Of the remaining 120 students, $24/120 = 20\%$ usually bike. The near-path cyclists' cell contributes $(36 - 24)^2/24 = 6$ to the test statistic. Adding all four contributions gives $\chi^2 \approx 14.29$ with one degree of freedom and a right-tail p-value of about 0.00016. All expected counts exceed five, and the random sample covers 2.5% of the district. At the 0.05 level, the data provide convincing evidence of an association among district students. The researcher observed where students lived. Distance to school, for example, could influence both neighborhood and travel mode. The association alone cannot establish whether building a path would cause more students to cycle.

Usual travel mode in the district sample

Residence	Bicycle	Other mode	Total
Near a path	36	44	80
Away from a path	24	96	120
Total	60	140	200

Source and adaptation: OpenStax §11.3 Example 11.5, printed pp. 574-575 (PDF pp. 586-587). Expected-count derivation adapted; association and sampling-design distinctions expanded for AP.

TWO READING CHECKS

1. Which two categorical variables did the researcher record for each student?

2. Use the given totals to find the expected number of cyclists who live away from a path.

PR36 | Checkpoint 4

Use the question to choose the comparison

PRACTICE FROM THE LAST MEETING

S07.06-HW (W179). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A coordinator might ask whether the proportion of hikers carrying refillable bottles exceeds a specified target. Each randomly sampled hiker contributes one yes-or-no response. The target gives a particular value for the population proportion under a null model. A sample result can then be compared with outcomes expected if that value holds.

A second coordinator might instead record both bottle use and trailhead for each hiker. A two-way table organizes these two categorical variables. Under a model of independence, the bottle-use pattern should be similar across trailheads, allowing for chance. The row and column totals determine the expected counts for individual cells.

For example, suppose 36 of 120 sampled hikers carry bottles. Under independence, the expected fraction carrying one at any trailhead is the overall fraction, $36/120$. If 50 of those 120 hikers began at the east trailhead, the expected number there with bottles is $50 \times (36/120) = 15$. If the observed count were 21, that cell's contribution to the chi-square statistic would be $(21 - 15)^2/15 = 2.4$. Statistical evidence depends on the full table. A single cell can show where the difference arose.

In either setting, an unusually large discrepancy needs an appropriate probability model and a defensible collection procedure. Record the investigative question, what one observation represents, and which quantities the null model holds fixed before calculating a test statistic.

The sampling plan helps name the question. One random sample of hikers, each classified by trailhead and bottle use, leads to a test of independence. Separate random samples at several trailheads, each classified by bottle use, lead to a test of homogeneity. The expected-count arithmetic is the same. For homogeneity, the null claim is that the distribution of bottle use is the same across the populations sampled. Check that all expected counts are at least five before using the chi-square approximation.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S07.05 (R128); S07.06 (R133).

TWO READING CHECKS

1. Which question compares one population proportion with a specified value?

2. Why collect both bottle use and trailhead in the second investigation?

PR37 | Estimating a population mean

What can sixteen pears tell us?

PRACTICE FROM THE LAST MEETING

(1) An SRS of 100 commuters from a large town contains 60 walkers and 40 cyclists. Thirty walkers and 30 cyclists carry a water bottle. Make the 2 × 2 table. (2) Find all expected counts under independence and the chi-square statistic. (3) State which variables are being compared and one sampling fact needed to draw a population conclusion.

READ FOR THE COMING MEETING

Suppose a grower wants the mean weight of all 800 pears in a shipment. She takes a simple random sample of 16 pears. Their mean weight is 210 grams and their sample standard deviation is 12 grams. Another random sample would probably give a different mean. How much uncertainty should she attach to her estimate of 210 grams?

The 12 grams describe how individual pear weights vary. To estimate how much the sample mean would vary over repeated samples of 16, use its standard error, $s/\sqrt{n}$. Here s is the sample standard deviation and n is the number of independent observations. The estimated standard error is $12/\sqrt{16} = 3$ grams. Each random sample is small relative to the shipment (16 of 800), so treating the observations as approximately independent is reasonable.

To build a 95% confidence interval, start at the sample mean and extend a margin of error on either side. When we estimate the population standard deviation from the sample, we use a t critical value. The t distribution is symmetric about zero. With small samples its heavier tails allow for uncertainty in the estimated standard error. A one-sample interval uses n − 1 degrees of freedom. For 15 degrees of freedom, the middle 95% of the t distribution lies between about −2.131 and +2.131.

Before using that multiplier, inspect a plot of the 16 weights for substantial skewness or outliers. If the shape is reasonably close to normal, the margin of error is $2.131 \times 3 \approx 6.4$ grams. The grower can report, “We are 95% confident that the mean weight of all pears in this shipment is between 203.6 and 216.4 grams.” A particular pear can weigh far more or less. The interval estimates one population mean.

PR37 | Continued

Try following the same calculation with a different sample size. OpenStax reports 15 sensory measurements with mean 8.2267 and standard deviation 1.6722. Under a normal population model and independent random sampling, the sample mean has estimated standard error $1.6722/\sqrt{15} \approx 0.432$. Now use 14 degrees of freedom: $8.2267 \pm 2.145(0.432)$ gives about 7.30 to 9.15. The two studies use different critical values because they have different numbers of observations.

Individual weights versus uncertainty in their mean

Summary	Value	Describes
Sample SD	12 g	Individual pear weights
Estimated SE, n = 16	$12/\sqrt{16} = 3$ g	Variation of sample means

Source and adaptation: OpenStax §8.2 Example 8.8, printed pp. 416-419 (PDF pp. 428-431). t-interval development and source summary adapted; random-sampling qualification added.

TWO READING CHECKS

1. Which number describes the spread of individual pear weights, and which describes uncertainty in the sample mean?

2. Why should the grower inspect a plot of the 16 weights?

PR38 | Confidence, precision, and defensible claims

How precise is a trail-time estimate?

PRACTICE FROM THE LAST MEETING

S08.01-HW (W185). Complete 1, 2, 3, 4, 5.

READ FOR THE COMING MEETING

Imagine a park manager has used a random sample of visits to obtain a 95% confidence interval of 48 to 56 minutes for the mean time hikers spend on a loop. What could the manager reasonably print on a sign? A proposed average of 50 minutes is consistent with the interval. An average of 60 minutes is outside it. Neither comparison tells an individual hiker how long their trip will last.

The interval is centered at 52 minutes, the sample mean. The distance from the center to either endpoint is 4 minutes, the margin of error. Precision refers to how tightly the interval locates the population mean. Its width is 8 minutes. A manager deciding whether the mean exceeds 45 minutes can use the entire interval as support. A decision that depends on whether the mean exceeds 54 minutes remains unresolved because the interval includes values on both sides of 54.

Where does 95% enter? Picture many managers independently drawing a random sample by the same procedure from the same population. Their sample means and endpoints would differ. If the conditions hold, about 95% of the intervals they construct this way will contain the same fixed population mean. After one interval is calculated, its endpoints are fixed too. The confidence level describes the method that produced them.

OpenStax's 15 sensory measurements illustrate what changes when the same data are analyzed at two confidence levels. Both intervals are centered at 8.2267 and use a standard error of about 0.432. For 14 degrees of freedom, a 90% t multiplier of 1.761 gives a margin of about 0.76. At 95%, a multiplier of 2.145 gives a margin of about 0.93. To achieve a higher long-run capture rate with those same measurements, we must allow a wider range on either side of the estimate.

PR38 | Part 1 of 2 | Student Reference R143 | Continue on the next page.

PR38 | Continued

The manager can improve precision at the original confidence level by sampling more visits. If variability stays similar, quadrupling the sample size roughly halves the standard error because the denominator is √n. Adding more early-morning hikers to an early-morning-only sample could give a narrow interval while still missing how long afternoon hikers take. A small margin of error accounts for sampling variability under the design.

Precision and confidence: change one feature at a time

Change	Usual effect, other features similar
Increase confidence	Wider interval
Increase sample size	Narrower interval
Increase variability	Wider interval

Source and adaptation: OpenStax §§8.1-8.2, printed pp. 407-419 (PDF pp. 419-431), especially Example 8.8 and confidence-width discussion. Precision principles adapted without adding a required known-σ mean procedure.

TWO READING CHECKS

1. What are the point estimate and margin of error for the loop?

2. How would changing to 99% confidence on the same sample affect precision?

PR39 | Paired observations: analyze the differences

Two scales, one crate

PRACTICE FROM THE LAST MEETING

S08.02-HW (W189). Complete 1, 2, 3, 5, 6.

READ FOR THE COMING MEETING

A produce buyer suspects that two scales give different readings. She takes a simple random sample of 14 crates from a shipment of 500 and weighs each crate on both scales. A heavy crate can raise the readings on both devices. How can the buyer compare the scales while accounting for which crates she happened to select?

Let d be scale A's reading minus scale B's reading for the same crate, in kilograms. A crate that reads 18.4 on A and 18.1 on B contributes +0.3 kilogram. If you plot these data for inference, put a dot at +0.3 on a number line labeled "A minus B (kg)." Each other crate contributes one more dot. Keeping the differences removes the crate's own weight from the comparison. The 28 readings now give 14 differences.

The parameter is the mean A-minus-B difference for all crates in this shipment. Suppose the 14 differences have mean 0.25 kg and standard deviation 0.28 kg. Check their plot for substantial skewness and outliers. The independent units are the sampled crates, and 14 is less than 10% of the shipment. If the shape looks reasonably normal, treat these differences as one quantitative sample. Their estimated standard error is $0.28/\sqrt{14} \approx 0.075$ kg. Using 13 degrees of freedom, the 95% margin is about $2.160 \times 0.075 = 0.16$ kg.

The interval is approximately 0.09 to 0.41 kg. We are 95% confident that scale A reads higher than scale B by 0.09 to 0.41 kilogram on average for these crates. Both endpoints are positive. Individual crates can still have negative differences. Reversing the order to B minus A would give an interval from −0.41 to −0.09 kg, with the same physical interpretation.

PR39 | Continued

OpenStax uses the same paired structure for pain scores recorded before and after treatment. A participant going from 6.6 to 6.8 contributes +0.2 in the after-minus-before order. A participant going from 6.5 to 2.4 contributes −4.1. The eight participants give eight differences.
Before-and-after scores show change over time. A causal claim about the treatment calls for a design that can separate its effect from other changes over that time.

Two measurements of the same crate

Measurement	Scale A (kg)	Scale B (kg)
Example crate	18.4	18.1

Source and adaptation: OpenStax §10.4 Example 10.11, printed pp. 527-529 (PDF pp. 539-541). Source matched-pair structure adapted for confidence intervals; causal claims kept distinct from before/after change.

TWO READING CHECKS

1. How many observations enter the analysis of differences, and why?

2. A crate gives a difference of −0.2 kg. Which scale read higher, and by how much?

PR40 | Estimating a difference between independent means

Comparing two campgrounds

PRACTICE FROM THE LAST MEETING

S08.03-HW (W195). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A ranger draws independent simple random samples of 36 completed stays from each of two campgrounds. Each file contains more than 360 stays, and each sampled party appears only once. Each response is the length of one stay in nights. Although the samples are the same size, a party at Pine has no natural partner at Oak. What can the ranger say about how the mean lengths of stay compare?

Define the parameter as the population mean at Pine minus the population mean at Oak. Suppose the sample means are 3.8 and 3.1 nights and their respective sample standard deviations are 1.2 and 0.9 nights. The observed difference is 3.8 − 3.1 = 0.7 night. In a new pair of samples the Pine mean might rise while the Oak mean falls. Both changes would increase the difference, so we need uncertainty from both samples.

Estimate each mean's standard error separately: $1.2/\sqrt{36} = 0.20$ night at Pine and $0.9/\sqrt{36} = 0.15$ night at Oak. For independent groups, add the variances and take the square root. The estimated standard error of the difference is $\sqrt{(0.20^2 + 0.15^2)} = 0.25$ night. Subtracting 0.15 from 0.20 would give a deceptively small value. Subtracting two means sets the direction of the comparison. It does not cancel uncertainty.

Check the distributions separately for severe skewness or extreme outliers. The independent, random samples each cover at most 10% of their population. If the distributions support t inference, use a two-sample t interval without pooling the variances. Technology uses both sample sizes and standard deviations to find about 64.9 degrees of freedom here. The 95% multiplier is about 2.00. The interval is $0.7 \pm 2.00(0.25)$, or about 0.2 to 1.2 nights.

PR40 | Continued

We are 95% confident that mean length of stay at Pine exceeds that at Oak by about 0.2 to 1.2 nights in these reservation populations. The statement concerns means. A particular Oak visitor can still stay longer than a particular Pine visitor. Visitors chose their campgrounds, so differences in travel plans or facilities remain possible explanations.

Length of sampled stays

Campground	Sample mean (nights)
Pine	3.8
Oak	3.1

Source and adaptation: OpenStax §10.1, printed pp. 512-519 (PDF pp. 524-531). Independent-group standard-error development adapted for AP two-sample t intervals; campground example is original.

TWO READING CHECKS

1. Why would sorting the two samples by length of stay and pairing the rows be inappropriate?

2. Interpret 0.7 night in this example.

PR41 | Testing a claim about one population mean

Are the bottles underfilled?

PRACTICE FROM THE LAST MEETING

S08.04-HW (W201). Complete 1, 3, 4, 5, 6.

READ FOR THE COMING MEETING

A one-sample t test compares an unknown population mean μ with a hypothesized value μ_0. State H_0: $\mu=\mu_0$ and an alternative matching the research question. The statistic $t=(\bar{x}-\mu_0)/(s/\sqrt{n})$ measures the signed distance from the null value in estimated standard errors. When conditions hold, compare it with a t distribution having $n-1$ degrees of freedom. Choose the tail from H_a before calculating the p-value.

OpenStax tests whether a population mean statistics score exceeds 65 points. Ten observed scores have mean 67 and standard deviation about 3.1972. The statistic is $(67-65)/(3.1972/\sqrt{10})\approx 1.978$ with 9 degrees of freedom. The upper-tail p-value is about 0.0396. At $\alpha=0.05$, this is evidence against the null in favor of a population mean above 65, assuming independent random sampling and an approximately normal score population. The p-value is the null-model probability of a t statistic this large or larger, not the chance that μ exceeds 65.

As with an interval, check the randomization, independence, sampling fraction when applicable, and shape evidence. A small sample requires particular attention to skewness and outliers. Report the parameter, hypotheses, conditions, calculation, p-value, decision, and contextual conclusion. Rejecting H_0 does not measure whether the difference matters educationally or practically. Failure to reject does not prove equality. The bottle-fill example applies the same reasoning to a lower-tail question.

A bottling line has a target mean fill of 500 milliliters. A supervisor wants to investigate possible underfilling. Before seeing the data, she chooses a one-sided question about whether the population mean is below 500 milliliters. She selects 25 bottles at random from a run of 2,000. The fill measurements show no marked skewness or outliers. The parameter μ is the mean fill of bottles in this run. Her hypotheses are H_0: $\mu = 500$ and H_a: $\mu < 500$, with fill measured in milliliters.

PR41 | Continued

Suppose the sample mean is 497 milliliters and its estimated standard error is 1.5 milliliters. The mean is 3 milliliters below the target, or two estimated standard errors below it. The test statistic keeps the direction of the difference, so its value is −2. A t model lets the supervisor judge how unusual this result would be if the population mean really were 500.

With 24 degrees of freedom, a t statistic of −2 cuts off about 0.0285 in the lower tail. Under a mean fill of 500 milliliters and the model assumptions, about 2.85% of samples would give a t statistic this low or lower. Values far above zero would point toward overfilling, the opposite of the supervisor's concern. At a significance level of 0.05, the data provide convincing evidence that the mean fill in this run is below 500 milliliters. Whether a shortfall of the estimated size warrants adjusting the line is a separate production decision.

Bottling-line measurements

Target mean (mL)	Sample mean (mL)	Estimated SE (mL)
500	497	1.5

Source and adaptation: OpenStax §9.5 Example 9.16, printed pp. 475-476 (PDF pp. 487-488). Source t-test and sample summaries adapted; sampling assumptions stated explicitly.

TWO READING CHECKS

1. Why did the supervisor choose the direction of the test before sampling?

2. Show how the supervisor obtained −2.

PR42 | Testing a population mean difference from paired data

Would a high-contrast map help?

PRACTICE FROM THE LAST MEETING

S09.01-HW (W209). Complete 1, 2, 3, 4, 5.

READ FOR THE COMING MEETING

A paired t test is a one-sample t test on a list of within-pair differences. Let μ_d be the population mean difference in the stated order. A usual null is H_0: $\mu_d=0$. The alternative can be $\mu_d>0$, $\mu_d<0$, or $\mu_d\neq0$, depending on the research question and subtraction order. Check the shape of the difference distribution and independence from one pair to another.

Consider the eight paired pain scores in OpenStax. The after-minus-before differences are +0.2, −4.1, −1.6, −1.8, −3.2, −2.0, −2.9, and −9.6. Their mean is −3.125 and their standard deviation is about 2.911. If the population mean difference were zero, how unusual would that average be? The estimated standard error is $2.911/\sqrt{8} \approx 1.029$. Thus the sample mean lies about 3.036 standard errors below zero. With seven degrees of freedom, the lower-tail p-value is about 0.0095 under the paired t model.

The model matters here. The eight people were randomly selected. The test treats people as independent, and OpenStax assumes a normal population of differences. That assumption matters with only eight pairs. The −9.6 observation is far from the others, so the numerical result is sensitive to that shape assumption. If the assumption is justified, there is evidence of a lower mean score after treatment. Everyone was measured before treatment and then after it. This design cannot by itself separate a treatment effect from other changes over time.

A trail designer wants to know whether a high-contrast map helps people find a marked junction more quickly. Each participant completes two comparable map tasks, one in each display style. For each person, subtract the high-contrast time from the standard-display time. A participant who takes 31 seconds with the standard display and 26 with the high-contrast display has a difference of 5 seconds.

PR42 | Continued

The resulting list contains one observation per participant. The null hypothesis says the population mean of these differences is zero. If the designer expects high contrast to be faster, the alternative says the mean difference is positive. An occasional negative difference is compatible with a positive population mean difference.

The designer has two separate questions to think through. First, how large is the sample mean difference relative to its estimated standard error? Second, what feature of the design could justify attributing a difference to display style? If everyone uses the standard display first, practice could speed up performance on the second task. For example, assign half the participants at random to use the high-contrast display first, and vary which route gets each display. Then one style will not always receive the benefit of a practice run or the easier route.

Two tasks for one participant

Map display	Time (s)
Standard	31
High contrast	26

Source and adaptation: OpenStax §10.4 Example 10.11, printed pp. 527-529 (PDF pp. 539-541). Paired t-test source example adapted with equality-null convention and explicit causal limits.

TWO READING CHECKS

1. For the participant with a 5-second difference, which display was faster?

2. Write the alternative in words using the stated subtraction order.

PR43 | Testing a difference between independent means

Is the mean wait longer at the North crossing?

PRACTICE FROM THE LAST MEETING

S09.02-HW (W215). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A two-sample t test compares two population means using independent samples. Define μ_1 and μ_2 with their populations and units. For H_0: $\mu_1-\mu_2=0$, calculate $t=[(\bar{x}_1-\bar{x}_2)-0]/\sqrt{(s_1^2/n_1+s_2^2/n_2)}$. The denominator estimates the standard deviation of the sampling distribution of the difference; it must include both groups. The sign retains the specified subtraction order.

In OpenStax's college example, 11 graduates of College A took a sample mean of 4 mathematics classes with s = 1.5. Nine graduates of College B took a sample mean of 3.5 with s = 1. The estimated difference is 0.5 class and its estimated standard error is $\sqrt{(1.5^2/11 + 1^2/9)} \approx 0.562$ class. For independent random samples from approximately normal populations, the unpooled test gives $t \approx 0.89$ and about 17.4 degrees of freedom. The upper-tail p-value of about 0.193 gives little evidence that graduates of College A take more mathematics classes on average. Half a class is less than one estimated standard error here.

Check each group's sampling or assignment process, independence, and shape evidence before using the reference distribution. Use the alternative to select a lower, upper, or two-sided tail. At a chosen significance level, compare the p-value with α and state whether the data provide convincing evidence for the specified population difference. Separate the size of the estimated difference from its statistical significance, and relate any causal claim to the design. In the crossing example, pedestrians chose their routes, so site differences can have several explanations.

A planner takes independent random samples of pedestrian waits at two crossings in a large archive of comparable visits, sampling less than 10% of the records at either site. Each sampled pedestrian appears once. The planner defines the difference as mean wait at North minus mean wait at South, in seconds. The question, chosen before sampling, is whether the population mean at North is greater.

PR43 | Continued

Suppose the planner samples 36 waits at each site and sees no marked skewness or extreme outliers in either sample. North has mean 26 seconds and standard deviation 8 seconds. South has mean 22 seconds and standard deviation 6 seconds. The estimated difference is 4 seconds, with estimated standard error $\sqrt{(8^2/36 + 6^2/36)} \approx 1.67$ seconds. Under a null difference of zero, this is $t \approx 2.40$. The unpooled two-sample t procedure gives about 64.9 degrees of freedom and an upper-tail p-value of 0.0096.

At $\alpha = 0.05$ the planner has convincing evidence that the mean wait at North is greater for the visits represented in the archive. A four-second difference would have weaker evidence in a study with more variable waits or smaller samples, because it would represent fewer standard errors. Whether four seconds matter depends on the planner's decision. Pedestrians chose their routes, so traffic timing and the mix of pedestrians remain possible explanations.

Observed pedestrian waits at two crossings

Quantity	Time
North sample mean	26 s
South sample mean	22 s
North-minus-South estimate	+4 s

Source and adaptation: OpenStax §10.1, Example 10.2, printed pp. 514-516 (PDF pp. 526-528). College summaries adapted from the source; the crossing case and its numbers are original.

TWO READING CHECKS

1. Define the null difference with a direction and units.

2. Could two studies with the same observed gap of 4 seconds yield different p-values? Why?

PR44 | Choosing inference and linking tests with intervals

Choose the question before the procedure

PRACTICE FROM THE LAST MEETING

S09.03-HW (W219). Complete 1, 2, 3, 5, 6.

READ FOR THE COMING MEETING

Choose an inferential procedure from the question, response variable, and study design. A categorical success/failure response leads to a proportion; a quantitative response leads to a mean. One sample, two independent groups, and paired observations require different analyses. A confidence interval estimates a parameter with uncertainty. A significance test evaluates evidence against a specified null value. The calculator menu should follow these decisions, not determine them.

OpenStax's one-proportion smartphone survey, independent medication groups, one-sample sensory measurements, and paired before/after scores illustrate these distinctions. For two numerical measurements on each participant, first form within-participant differences. For separate independent numerical groups, estimate two means and account for both standard errors. For success counts, keep successes and sample totals identified for every group. A data table with two columns can represent either paired measurements or two independent samples; its column layout alone does not answer the design question.

A two-sided t test at significance level α and the corresponding $100(1-\alpha)\%$ t confidence interval make consistent decisions when they use the same method, standard error, degrees of freedom, and assumptions. A null value outside the interval is rejected by that matching test; one inside is not rejected. Do not apply this rule indiscriminately to one-sided tests or procedures using different standard errors, such as some proportion tests and intervals. A one-sided p-value cannot be inferred by merely reading a two-sided interval without considering the relevant method and direction.

A school is considering two ways to print large posters. It could record whether each poster finishes within a minute, or it could record the actual time in seconds. The first response gives a success or failure and leads to a question about proportions. The second response is quantitative and leads to a question about means. The choice changes what the school can learn from its records.

PR44 | Continued

Suppose each of 20 designs is printed once by each method. A useful investigative question is, “For designs like these, what is the mean difference in printing time, in seconds, between methods A and B?” Define each difference as time with A minus time with B. Each design supplies a pair. A paired t interval can estimate the mean difference. If the school had independently sampled one set of designs for A and another for B, it would instead need a two-sample t interval. Twenty results per column could occur under either design.

Imagine a 95% paired t interval of −4 to −1 seconds for the mean difference. Zero lies outside it. The matching two-sided t test at the 5% level would reject a null mean difference of zero. The negative endpoints estimate how much less time A takes on average. To decide whether the faster method saves enough time, ask what saving would matter. A school needing at least 3 seconds per poster would still face uncertainty: the interval includes savings smaller and larger than 3 seconds.

Response and design determine the parameter

Response and structure	Target
Yes/no, one population	One population proportion
Numerical, one population	One population mean
Numerical, same units measured twice	Mean within-pair difference
Numerical, independent groups	Difference of population means

Source and adaptation: OpenStax Chapters 8-10, especially §§8.2,8.3,10.1,10.3,10.4, printed pp. 416-427 and 512-531 (PDF pp. 428-439 and 524-543). Method selection synthesized; matching-test/interval qualification added for AP.

TWO READING CHECKS

1. Which response would let the school estimate the fraction of posters that finish within a minute?

2. Which method is faster on average according to the interval, and by about how much?

PR45 | Does a design change increase flight time?

Count the assigned units

PRACTICE FROM THE LAST MEETING

S09.04-HW (W225). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

The project compares the fall times of long-wing and short-wing paper helicopters made by the class. The templates should differ only in wing length. Keep the paper type, body dimensions, release height, and timing rule the same. A paper-and-builder block supplies two comparable pieces. Randomly assign one piece to each design within that block, then randomize release order. The block links helicopters that share their material and builder.

The treatment is assigned to a helicopter, so the helicopter is the experimental unit. Three releases of that helicopter are repeated measurements of the same assigned unit, not three independently assigned helicopters. Average those releases to obtain one fall-time summary for that helicopter. Next subtract the short-wing average from the long-wing average within the same block.

For example, if a block's long-wing helicopter averages 1.63 seconds and its short-wing helicopter averages 1.39 seconds, its long-minus-short difference is +0.24 second. A positive difference favors a longer fall time for the long-wing design. Keep this subtraction order for every block. Twelve blocks give twelve differences for a paired t analysis, even though there are twenty-four helicopters and seventy-two releases.

Consider a possible source of error. If every long-wing helicopter is dropped first, students might get better at timing the later drops. Release order would then vary with wing design. A coin toss within each block can decide which design goes first. The same timing rule should apply even to a helicopter that falls surprisingly quickly.

PR45 | Continued

Before calculating, examine the distribution of the twelve differences for substantial skewness or outliers and consider whether blocks can be treated as independent. The test concerns a mean difference under the stated construction and release conditions. Random assignment supports a causal comparison within those conditions; it does not make one class's materials and builders representative of all possible helicopters. If class collection is unavailable, the workbook provides clearly labeled fictional differences for analysis practice. Never report those fallback values as measured class results.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S02.04 (R041); S08.03 (R147).

TWO READING CHECKS

1. With twelve blocks, two helicopters per block, and three releases per helicopter, how many differences enter the paired t analysis? Explain.

2. What does a long-minus-short difference of −0.2 second mean for a block?

PR46 | Checkpoint 5

Count the independent comparisons

PRACTICE FROM THE LAST MEETING

Using your helicopter project sheet, (1) identify the treatment assignment, experimental unit, and block, (2) annotate your difference plot and interval with units and subtraction order, and (3) write two claims, one supported under the testing conditions and one requiring new evidence. Label any fallback data as fictional.

READ FOR THE COMING MEETING

Two questions about delivery times can require different analyses. A technician might time the same ten printers before and after an update. Each printer supplies two related times. Defining a difference for each printer turns the question into one about the population mean of those differences.

Suppose one printer takes 34 seconds before and 30 seconds after. With after minus before as the order, its difference is −4 seconds. A negative mean difference would indicate faster operation after the update. Look for skewness and outliers in the differences themselves. The separate before and after distributions each omit the information about which times belong together.

Another technician might randomly assign twenty printers to one of two separate update conditions and record one delivery time from each printer. Each group now supplies its own set of independent observations. The population quantity of interest is a difference between two mean times, and variation within both groups contributes to the uncertainty.

The unit of measurement is seconds in both examples. The design determines whether a row pairs two values or belongs to a separate group. Before using a calculator, sketch what one row of data would look like and identify the independent units that supply information.

For a small paired sample, the t calculation uses the sample standard deviation of the differences divided by the square root of the number of pairs. For example, if nine independent pairs have a sample standard deviation of 6 seconds, the estimated standard error of the mean difference is $6/\sqrt{9} = 2$ seconds, with eight degrees of freedom. Timing more pages on the same printers would improve the individual measurements, but would not add independent printers.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S08.03 (R147); S09.02 (R161).

TWO READING CHECKS

1. For the first study, what would one row in a table of differences represent?

2. Why can the second design use a two-independent-sample approach?

PR47 | Seeing relationships in paired quantitative data

Which ride does the point represent?

PRACTICE FROM THE LAST MEETING

(1) Sketch data tables for a matched-printer study and a study with independent groups. (2) For six paired after-minus-before times of −3, −2, −2, −1, 0, and 2 seconds, plot the differences and compute their mean and standard error. (3) Write the parameter, null hypothesis of no mean change, and one condition to investigate before testing.

READ FOR THE COMING MEETING

A scatterplot displays two quantitative variables measured on the same observational units. Each point is an ordered pair (x,y) from one unit. Put the explanatory variable on the horizontal axis and the response on the vertical axis when that distinction is useful. Label both axes with the variable names and units. Sorting the two data columns independently would destroy the original pairings and create observations that were never collected.

OpenStax's vocabulary example pairs child age in years with vocabulary size in words: the displayed pairs are (3, 655), (4, 1,098), (6, 2,463), and (7, 3,195). Each point joins the age and vocabulary value from the same record. The displayed values increase together, giving a positive association. A graph allows us to ask whether the form is approximately linear, how closely points follow that form, and whether any points stand apart. These four points alone provide limited evidence about all children or ages outside the displayed range.

When describing a scatterplot, address direction, form, strength, and unusual features in context. "Positive" means higher values of one variable tend to occur with higher values of the other. Strength concerns scatter around the pattern, not the steepness of a line. Explanatory and response terminology does not establish cause and effect. Observational data can reflect lurking variables, selection effects, or both. The ride example below lets you practice keeping each record intact and describing what an unusual point might mean.

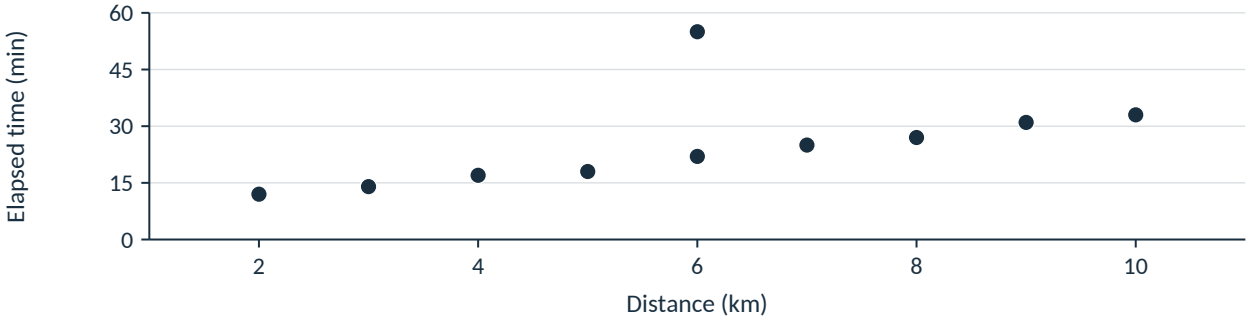
Suppose a cycling club records distance in kilometers and total time in minutes for ten rides. The club wants to use distance to anticipate time. On the scatterplot, find 8 km on the horizontal axis, then move up to the dot at 27 minutes. That dot represents a single observed ride. Two rides can share the same distance and have different times: at 6 km the club observed both 22 and 55 minutes.

PR47 | Continued

Follow the main band of dots from left to right. Among these rides, time generally increases with distance in a fairly tight, nearly straight pattern. The 55-minute ride lies far above the others near 6 km. The club could check the original record for a long stop or a steep route. Sorting the time column before graphing would assign that 55-minute time to the longest ride. The graph would then suggest a different story based on a ride nobody recorded.

Before drawing a conclusion, consider which rides the club recorded. Perhaps members chose short routes for casual outings and long routes for training. Ride purpose and terrain could affect elapsed time as well as distance. The plotted records can describe an association and suggest a prediction. They do not isolate the effect of adding a kilometer to a ride.

Distance and elapsed time for ten illustrative rides



Each dot is one ride. Two 6-km rides took different amounts of time.

Source and adaptation: OpenStax §12.2 Example 12.5, printed pp. 620-622 (PDF pp. 632-634). Source scatterplot data and interpretation adapted; ride case is original.

TWO READING CHECKS

1. Where would a 10-kilometer ride lasting 36 minutes appear, and what does that point represent?

2. What would a tight, upward, nearly straight band suggest about these rides?

PR48 | Correlation: strength of a linear association

A curved route can have zero correlation

PRACTICE FROM THE LAST MEETING

S10.01-HW (W233). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

The correlation coefficient r measures the direction and strength of a linear association between two quantitative variables. It lies between −1 and 1. Positive values indicate an upward linear tendency; negative values indicate a downward tendency. Values close to either endpoint indicate a strong linear relationship, while values near 0 indicate little linear association. An r of 0 does not rule out a strong curved relationship.

In OpenStax's third-exam/final-exam data, the reported correlation is about 0.6631. Students with higher third-exam scores tended to have higher final-exam scores, although the points did not lie exactly on a line. Correlation has no units because it is based on standardized values of both variables. Changing points to percentages by a positive linear rescaling leaves r unchanged. Interchanging the explanatory and response variables also leaves r unchanged, even though the regression equation used for prediction changes.

Always inspect the scatterplot before interpreting r. A single unusual point can change it substantially, and a curved pattern can make a small r misleading. Correlation does not establish a causal mechanism. Its sign also does not say whether a relationship is desirable; that depends on the variables. Keep r distinct from the slope: a slope measures predicted change in response units per explanatory unit, whereas r summarizes a unit-free linear association. The symmetric valley below shows a clear relationship with zero correlation.

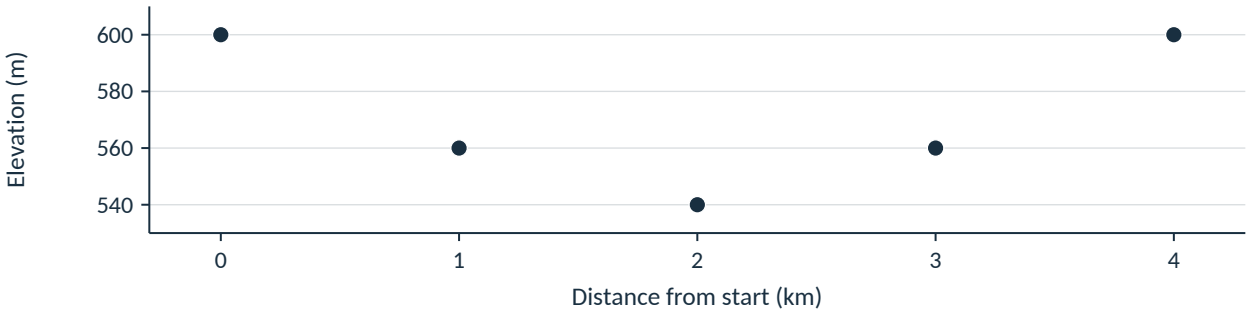
A cyclist starts on a ridge, descends into a valley, and climbs to another ridge. At distances of 0, 1, 2, 3, and 4 kilometers from the start, the recorded elevations are 600, 560, 540, 560, and 600 meters. Sketch the trend with your finger from left to right on the plot. One straight line cannot follow both the descent and the climb.

The middle distance is 2 kilometers. The two observations one kilometer to either side of that midpoint have exactly the same elevation, as do the two observations two kilometers to either side. One member of each pair pulls a fitted line downward and the other pulls it upward by an equal amount. The horizontal line at the mean elevation, 572 meters, is the best straight-line fit. The sample correlation is exactly zero, even though the points follow a clear curve.

PR48 | Continued

Suppose the cyclist changes kilometers to miles. Every distance is multiplied by the same positive constant, so distances below and above their mean stay the same numbers of standard deviations below and above it. Elevation has not changed either. Correlation therefore stays the same. By contrast, restricting attention to just the ascent changes which observations enter the calculation and would give a positive correlation.

Elevation along the illustrative cycling route



The points show a valley. Equal elevations on either side of 2 km balance the linear trend.

Source and adaptation: OpenStax §12.3, printed pp. 627-630 (PDF pp. 639-642). Correlation definition/properties adapted; curved-route example is original to expose a limitation of r. Correlation significance tests are outside this lesson.

TWO READING CHECKS

1. What shape would you see in the cyclist's scatterplot, and why might a single straight-line description be unhelpful?

2. Why does expressing distance in miles leave r unchanged?

PR49 | Linear predictions and their limits

What does the model predict at six kilometers?

PRACTICE FROM THE LAST MEETING

S10.02-HW (W237). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A fitted regression line predicts a response y from an explanatory value x. Write the fitted response as $\hat{y}=a+bx$. The slope b is the change in predicted response for a one-unit increase in x. Its units are response units per explanatory unit. The intercept a is the predicted response at $x=0$. That interpretation is useful only if 0 is meaningful and lies within a range for which the model is defensible.

OpenStax's exam-score example gives the rounded line $\hat{y}=-173.51+4.83x$, where x is a third-exam score out of 80 and $\hat{y}$ is a predicted final-exam score out of 200. Using this rounded equation at $x=66$ gives about 145.3 points. Its observed third-exam scores range from 65 to 75. Substituting $x=90$ gives 261.2, exceeding the possible 200-point final score and also using an impossible third-exam score. A calculator can produce that number, but it is not a defensible prediction.

Prediction outside the observed explanatory-variable range is extrapolation. Even within the range, the model estimates a tendency and does not guarantee an individual response. Examine the scatterplot and residual behavior, consider the populations and conditions that produced the data, and do not extend a fitted relationship merely because the equation accepts another input. A slope from observational data describes association, not automatically the effect of changing x. The shuttle model below provides a local context for these same distinctions.

A shuttle dispatcher wants to predict trip time from distance. Suppose that for routes between 2 and 10 kilometers, a fitted line predicts time in minutes as 5 plus 3 times the distance in kilometers. For a 6-kilometer route, the prediction is $5 + 3(6)$, or 23 minutes. Different trips of that length can still take different amounts of time.

PR49 | Continued

The 3 in the model has units of minutes per kilometer. Within the observed range, an additional kilometer corresponds to 3 additional minutes of predicted time. The 5 is the predicted time at zero kilometers. Because none of the recorded routes was shorter than 2 kilometers, the data give little basis for interpreting this intercept as an observed loading or waiting time.

A dispatcher could substitute 30 kilometers and get 5 + 3(30) = 95 minutes. The arithmetic is valid, but trips beyond the observed 2-10 km range may use faster roads or make more stops. This is extrapolation. For a route inside the range, such as 6 km, 23 minutes is a reasonable model prediction only if the route resembles those measured. Applying the line to emergency vehicles on a closed course raises a different concern: the vehicles and operating conditions have changed even if the distance is familiar.

Shuttle line: predicted minutes = 5 + 3 × kilometers

Distance	Model calculation	Relation to observed 2-10 km
6 km	5 + 3(6) = 23 min	Within range
30 km	5 + 3(30) = 95 min	Extrapolation; not justified by fit

Source and adaptation: OpenStax §§12.3 and 12.5, printed pp. 623-630 and 635-636 (PDF pp. 635-642 and 647-648), especially Example 12.11. Source regression prediction adapted; unsupported extrapolation highlighted.

TWO READING CHECKS

1. What is the predicted time for a 7-kilometer route? Show the substitution.

2. Why is the intercept a weak basis for claiming that the shuttle spends 5 minutes loading?

PR50 | Residuals and checking a linear model

How far off was the prediction?

PRACTICE FROM THE LAST MEETING

S10.03-HW (W241). Complete 1, 2, 3, 4, 6.

READ FOR THE COMING MEETING

A residual is observed response minus predicted response: $e=y-\hat{y}$. It is the signed vertical distance from a point to the fitted line. A positive residual means the observed value is above the line, so the model underpredicts. A negative residual means it is below the line, so the model overpredicts. Residuals retain the units of y. They are not horizontal distances and not ordinary recording mistakes.

For OpenStax's rounded exam-score line $\hat{y}=-173.51+4.83x$, a student with third-exam score 65 has prediction 140.44. The observed final-exam score in the source data is 175, giving a residual $175-140.44=34.56$ points. The model underpredicts that student's final score by about 34.6 points. Another student can have a negative residual even when the overall association is positive; residual direction and correlation direction answer different questions.

A residual plot shows explanatory values or fitted values on the horizontal axis and residuals on the vertical axis. A clear curve suggests that a straight line misses a systematic pattern. Changing vertical spread suggests prediction errors differ across the range. Investigate unusual observations using their original records before deciding how to handle them. A point far in the x-direction has high leverage; it is influential if its inclusion materially changes the fitted model. High leverage does not guarantee influence. The inspection-time example focuses on reading residual signs and patterns.

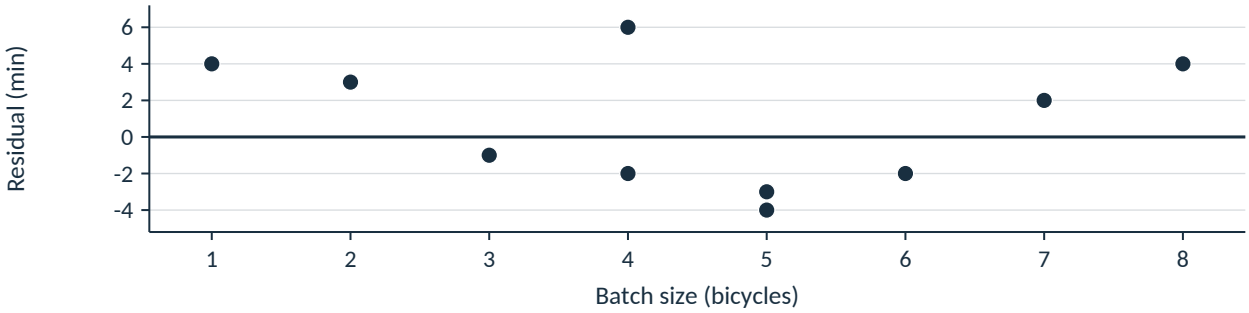
A mechanic wants to assess a line fitted using earlier batches of bicycle inspections. The line predicts minutes as 12 plus 7 times the number of bicycles. Now inspect a new batch of four. Its prediction is $12 + 7(4) = 40$ minutes, while the work actually took 46 minutes. The batch took 6 minutes longer than the model predicted.

To put this batch in a residual plot, keep its horizontal coordinate, 4 bicycles, and replace its vertical coordinate, 46 minutes, with $46 - 40 = +6$ minutes. At $(4, +6)$, the dot is above the zero line. Another four-bicycle batch that finishes in 38 minutes would appear at $(4, -2)$. A negative residual means the work took less time than the model allotted. Two batches can therefore share an input without sharing a residual.

PR50 | Continued

The plot shows results from ten new batches. Most small and large batches finish later than predicted, while most middle-sized batches finish earlier. The positive residual at four bicycles also deserves a closer look. A residual plot with points scattered around zero and roughly similar vertical spread across batch sizes would support keeping the line. Here the change in direction suggests that the mechanic should investigate whether setup and staffing work differently for batches of different sizes.

Prediction errors for ten new inspection batches



The horizontal line marks an exact prediction. Dots above it indicate longer times than predicted.

Source and adaptation: OpenStax §§12.3 and 12.6, printed pp. 625-626 and 636-643 (PDF pp. 637-638 and 648-655). Residual definition and diagnostic reasoning adapted; AP distinction between leverage and influence retained.

TWO READING CHECKS

1. For a batch with a prediction of 33 minutes and an actual time of 29 minutes, what is the residual? Interpret its sign.

2. Where would the example batch of four appear on the residual plot?

PR51 | Least squares and explained variation

How much did the line improve the predictions?

PRACTICE FROM THE LAST MEETING

S10.04-HW (W245). Complete 1, 2, 4, 5, 6.

READ FOR THE COMING MEETING

A least-squares regression line predicts a numerical response from a numerical explanatory variable. Technology needs the original pairs. Put the explanatory observations in one list or column and the corresponding responses in the next. Select a linear fit that includes an intercept and identify which list is x and which is y. Read the intercept and slope by their roles in the displayed equation: the coefficient multiplying x is the slope. Reversing the input lists would answer a different prediction question.

For least-squares regression with an intercept, r^2 is the fraction of the observed variation in y about its mean that is explained by using the linear relationship with x. OpenStax's exam-score data have $r \approx 0.6631$, so $r^2 \approx 0.4397$. About 44% of the variation in final-exam scores in those data is explained by the linear regression on third-exam score. The remaining 56% is variation around the fitted line. This is not the percentage of students whose scores were predicted correctly.

The baseline for r^2 predicts the sample mean response for every observation. Compare that baseline sum of squared deviations with the fitted-line residual sum of squares: $r^2 = 1 - SSE/SST$. Neither a high r^2 nor a small residual spread establishes causation or excuses extrapolation. A high r^2 can also coexist with a systematic curved pattern, so retain the graphical check. The print-shop comparison below shows the fraction explained and the fraction remaining using small totals that are easy to verify.

A print-shop manager compares two rules for predicting job time from page count. The simpler rule predicts the sample mean time for every job. The fitted line gives each page count its own prediction. Check that the line passes through the mean point: at the sample mean page count, it predicts the sample mean time. This does not require any individual job to sit at that point.

PR51 | Continued

How do we compare the two rules fairly? For every job in the same data set, subtract predicted time from observed time, square that residual, then add the squared residuals. For example, errors of +2 and −2 seconds cancel if added directly, even though both predictions missed. Squaring yields 4 + 4 = 8 square seconds. The least-squares line minimizes the total across all straight lines. A miss twice as large contributes four times as much to that total.

Suppose the mean-only rule has a total squared error of 40 square seconds, while the fitted line has a total of 12. Of the original 40, the line eliminates 28. The explained fraction is $(40 - 12)/40 = 0.70$. Thus the linear relationship with page count explains 70% of the observed variation in print time about its mean. Each model's prediction still refers to seconds. The fraction compares two quantities in the same squared units and therefore has no units.

Compare the same print jobs under both rules

Prediction rule	Sum of squared errors
Sample mean for every job	40 s^2
Fitted line using page count	12 s^2
Reduction from baseline	28 s^2, or 70%

Source and adaptation: OpenStax §12.3, printed pp. 625-630 (PDF pp. 637-642), especially least-squares and coefficient-of-determination discussions following Example 12.6. Definitions and source r^2 value adapted.

TWO READING CHECKS

1. Why square residuals before adding them?

2. Which prediction supplies the baseline sum of 40 seconds squared?

PR52 | Checkpoint 6

Read a prediction in its units

PRACTICE FROM THE LAST MEETING

S10.05-HW (W249). Complete 1, 2, 3, 4, 5.

READ FOR THE COMING MEETING

A line fitted to a sample of delivery routes predicts minutes of driving from route length in kilometers. Suppose its equation is predicted time = 5 + 2 × length, and observed lengths range from two to twelve kilometers. The slope of two describes a change of two predicted minutes for one more kilometer within the observed relationship.

For a five-kilometer route the line predicts fifteen minutes. A driver who takes eighteen minutes has a residual of three minutes because residual equals observed time minus predicted time. A positive residual places the observation above the fitted line when time is on the vertical axis.

Route length alone cannot account for every difference in travel time. Intersections, traffic, or loading delays may contribute to the residuals. Before using the line for a particular route, check its length against the recorded range and examine the residual plot for a pattern.

What would count against the straight-line model even if its predictions are sometimes close? If short and long routes usually have positive residuals and medium routes usually have negative ones, the model underpredicts at both ends and overpredicts in the middle. That systematic bend suggests a curved relationship. A scattered band around zero with fairly constant width gives more support for using a straight line over the observed distances.

The fitted line also passes through the point of sample means. Suppose a different set of routes averages 8 kilometers and 19 minutes, with fitted slope 1.5 minutes per kilometer. A route 2 kilometers longer than average is predicted to take 3 minutes more than average, or 22 minutes. Writing the equation as predicted time = 19 + 1.5 × (length − 8) makes that comparison visible.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S10.03 (R183); S10.04 (R187).

TWO READING CHECKS

1. What are the units of the slope in this model?

2. How would you interpret a residual of −2 minutes for a route?

PR53 | Diagnostic and evidence inventory

Use one answer as evidence

PRACTICE FROM THE LAST MEETING

(1) For lengths 2, 4, 6, and 8 km with observed times 10, 12, 15, and 23 minutes, use predicted minutes = 5 + 2 × kilometers to compute predictions and residuals. (2) Plot residuals against length with a labeled zero line. (3) Identify the largest absolute error and state one limitation of using this small set to evaluate the model.

READ FOR THE COMING MEETING

A student writes, “The second group is faster,” after comparing two lists of finishing times. The teacher asks which feature of the lists supports that claim. The student can calculate both medians correctly but omitted the numbers and units from the comparison. That omission points to a writing target the student can practice.

Try repairing an answer in a different setting. A report says “the second shift is slower,” and the two observed median packing times are 12 seconds for the morning shift and 17 for the afternoon. A sentence that can stand on its own is “The afternoon shift took 5 seconds longer per parcel at the median, 17 seconds compared with 12 seconds in the morning.” A reader now has the groups, the measurement, the statistic, the units, and the evidence.

A different student compares the largest times and calls them medians. Adding units would leave the mistaken choice of statistic intact. This student needs practice finding a meaningful center and deciding what the center describes. Both students can improve their next response by identifying the specific step where their evidence failed.

During diagnostic review, save one original sentence or calculation, explain the difficulty, and write a revision. Then use a short new example to check whether you can apply the idea without copying the correction.

Use the same approach to diagnose a design answer. Someone who writes “we randomized” may know the procedure but have left out the steps. Someone who assigns all experienced workers to one method has a different problem. The groups already differ in experience. A new reader should be able to tell how workers received their methods and why the design supports the conclusion.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S01.04 (R017); S02.03 (R037).

TWO READING CHECKS

1. What evidence suggests the first student needs a writing target?

2. What error should the second student address first?

PR54 | Comparing distributions

Compare the middle half

PRACTICE FROM THE LAST MEETING

Revisit your returned C1 free response, parts b and d. (1) Rewrite each answer without looking at the key. (2) Mark one design step or population limitation that you made explicit. (3) Write a new two-sentence example that demonstrates the same skill in a different study.

READ FOR THE COMING MEETING

A school shop records assembly times for two kinds of kits. For Kit A the median is 18 minutes, with quartiles 14 and 22. For Kit B the median is 20 minutes, with quartiles 18 and 23. Saying that the times are “about the same” skips two separate comparisons a reader could actually use.

A median describes a typical assembly time. The interquartile range (IQR) is the distance from the lower quartile to the upper quartile, where the middle half of the observations lies. For example, another kit with quartiles of 10 and 16 minutes has an IQR of 16 − 10 = 6 minutes. Calling 6 minutes its full range would require knowing that 10 and 16 were the minimum and maximum, which the quartiles alone cannot establish.

Think about what an extreme delay would do. In a long list of sorted times, imagine that the largest time changes from 30 minutes to 90. The total increases by 60 minutes, so the mean rises. The observations in the middle of the sorted list stay where they are, so the median and the quartiles stay the same. That resistance is useful when describing typical performance. A plot still matters if the manager also needs to see unusually long delays.

When comparing the original two kits, give two separate statements. Use the medians to say which kit typically took longer. Use the IQRs to say which kit had a more variable middle half. Both comparisons should include the size of the difference in minutes.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S01.04 (R017); S01.05 (R021).

TWO READING CHECKS

1. Which kit had the larger median assembly time, and by how much?

2. Compute the IQR for each kit and interpret the difference.

PR55 | Sampling and scope

Check who could enter the sample

PRACTICE FROM THE LAST MEETING

(1) Complete the two C1 sample-checkpoint items on IQR and transforming a variable (MCQ 2 and 3), justifying each answer and identifying the error in each alternative. (2) For kit times in PR54, express both medians and IQRs in seconds. (3) Explain how a common two-minute setup delay would affect these summaries.

READ FOR THE COMING MEETING

A town wants to know what fraction of its households have a working smoke alarm. A survey team draws addresses from a list of its water customers. Before generating random numbers, the team should ask which households the list actually covers. Some residents might share service through a landlord and never appear as individual customers.

Random selection from the list gives listed units a known chance of selection. It cannot reach an omitted household. Even a complete list leaves a second issue if households that answer differ from those that decline. The team should document contact attempts and the definition of a completed response.

Suppose the team obtains a complete address list with one record per household, including apartments. Label all records, generate the required number of distinct random labels, and save the selected list. Make the same follow-up attempts at every selected address. If a neighbor offers to take an absent resident's place, the neighbor's availability does not give that household the original household's selection probability. Record the outcome for the selected address.

A privacy plan can also improve the study. The team can separate the contact list from the dataset of alarm responses, use temporary study IDs, and report aggregate results. A useful design explains how the sample was chosen and what information needs protection.

Random sampling answers "Who enters the study?" In an experiment that tests whether reminders improve alarm maintenance, a second random step would decide which participating households get reminders. Drawing the sample permits generalization to the covered population under suitable response conditions. Assigning reminders permits a causal comparison within the experiment.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S02.01 (R029); S02.02 (R032).

TWO READING CHECKS

1. Give one reason the water-customer list might underrepresent households.

2. Which issue remains if a complete household list is used but many selected households never respond?

PR56 | Experimental design workshop

Make an assignment procedure executable

PRACTICE FROM THE LAST MEETING

A town wants to estimate the proportion of households using a curbside service. (1) Define the population, response, and reference period. (2) Write a reproducible random-sample plan, including how you would account for apartments and nonresponse. (3) Explain which records to separate to protect privacy.

READ FOR THE COMING MEETING

Suppose twenty-four volunteers will compare two keyboard layouts. People who type quickly at the start may also type quickly under either layout. Before assignment, the instructor gives everyone the same baseline typing test and forms two groups of twelve using those scores.

Within each group the instructor labels the volunteers 01 through 12, selects six distinct labels at random for layout A, and assigns the remaining six to layout B. The same passage length, practice period, and scoring rule apply to both layouts. Someone else could carry out those directions without deciding who looks best suited to either treatment.

Grouping similar volunteers can improve precision when baseline speed predicts the later response. The comparison still depends on assigning both layouts within each group. A plan that gives every fast typist A and every slower typist B confounds layout with baseline speed.

A classmate should be able to reconstruct one block from your record. In a demonstration block of six volunteers, suppose the random draw for A is 02, 04, 05. Volunteers 01, 03, 06 receive B. Before typing begins, define the response, for example correctly typed words in a two-minute passage. Compare the two layouts among volunteers with similar starting speed. If everyone is taught layout A with a tutor and layout B without one, any score difference could also reflect the tutoring.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S02.03 (R037); S02.04 (R041).

TWO READING CHECKS

1. How many volunteers in each baseline group receive layout B?

2. Give one condition the instructor should keep consistent across the two treatments.

PR57 | Probability and simulation

Build one trial before repeating it

PRACTICE FROM THE LAST MEETING

(1) Rewrite this weak plan: “Use the new layout with experienced typists and compare their scores with beginners.” Specify the response, blocks, and assignment procedure. (2) Make and record one valid assignment for two groups of six. (3) Explain the confounding in the weak plan and one implementation detail to hold fixed.

READ FOR THE COMING MEETING

A trail counter sometimes misses a bicycle when riders pass close together. A simplified model assumes each of four well-separated bicycles is detected independently with probability 0.80. The question is how often the counter records at most two of the four bicycles. A simulation can estimate this probability if one trial matches the event in the question.

Use a random digit from 0 through 9 for each bicycle. Let digits 0 through 7 mean detected and 8 or 9 mean missed. Read four digits for one trial and count detections. A trial with one or two detections qualifies, and a trial with none qualifies too.

For example, the digits 2, 7, 6, 9 produce three detections in a four-bicycle group. This trial falls outside the event “at most two.” If another run uses 120 digits, it contains 30 trials of four bicycles each. Seven qualifying trials would yield an estimated probability of 7/30, about 0.23. The event is defined for a group of four, which is why the denominator counts groups.

After many trials, divide the number of qualifying trials by the total number of trials. Different runs will give different estimates. Repeating the calculation more often reduces simulation noise, while dependence between actual bicycles would still challenge the model.

When you change to a question about unusually many detections, change the rule for a qualifying trial too. The simulation still generates whole groups under the assumed detection rate. What changes is the end of the distribution whose outcomes you count.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S03.01 (R047); S03.02 (R051).

TWO READING CHECKS

1. Why do eight of the ten digits represent detection?

2. For digits 8, 1, 9, 4, does this trial qualify?

PR58 | Normal and binomial models

Decide what the variable measures

PRACTICE FROM THE LAST MEETING

(1) Design a simulation for three independent parcels, each arriving on time with probability 0.70; estimate the chance at least two arrive on time. (2) Run at least twenty trials and save the first three coded trials. (3) Compare your estimate with the exact probability and explain one model assumption.

READ FOR THE COMING MEETING

A manufacturer records two quantities at a bottling station. One is the number of underfilled bottles among the next twenty. The other is the volume in milliliters of one bottle. The first quantity takes integer values from zero to twenty. The second is measured on a continuous scale.

A binomial model for the count requires a fixed number of trials, two outcomes per bottle, a common probability of underfilling, and independent outcomes. If the fill mechanism drifts during the run, the probability might change. If a valve sticks for several bottles, their outcomes might be linked.

A normal model for individual volume requires evidence or an explicit assumption about the distribution of volumes. A numerical mean and standard deviation alone do not establish a normal shape. State the variable and translate the requested event before choosing the model parameters.

Suppose the underfill probability stays at 0.05 for twenty independent bottles. To find the chance of at least two underfills, mark 2 through 20 on a count axis. The complement contains only 0 and 1. $P(0) = 0.95^{20}$, and $P(1) = 20 \times 0.05 \times 0.95^{19}$ because any one bottle could be the only underfill. Subtracting both probabilities from 1 gives about 0.264. Subtracting the cumulative probability through 2 would wrongly discard groups with exactly two underfills.

For a different question, suppose individual bottle volumes follow a normal model with mean 500 mL and standard deviation 3 mL. A volume of 494 mL is $(494 - 500)/3 = -2$ standard deviations from the mean. Shade the area below that value to find the probability of less than 494 mL, about 0.023. Here the outcome is a measurement on a continuous scale, so “below” and “at or below” have the same probability.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S04.02 (R068); S04.04 (R075).

TWO READING CHECKS

1. Which variable in the example can take a value such as 499.6?

2. How could a sticking valve undermine the binomial count model?

PR59 | Sampling distributions

Name the object behind each dot

PRACTICE FROM THE LAST MEETING

(1) Eight independently packed cartons each have a 0.15 probability of a damaged item. Compute P(at least one affected carton). (2) Separately, filling weight follows a normal model with mean 500 g and SD 4 g. Find P(weight > 506 g). (3) Explain why the inputs and units differ.

READ FOR THE COMING MEETING

Imagine a greenhouse with an inventory of several thousand plugs. A worker weighs a simple random sample of twenty plugs. A dotplot of those twenty weights has one point for each plug in the sample. The mean of the twenty weights is one summary of the sample.

Now imagine drawing another sample of twenty, computing its mean, and repeating the process many times. A plot of the resulting means has one point for each complete sample. It describes how the sample mean varies for samples of this size from this inventory. Both plots use grams on the horizontal axis, so units alone cannot identify them.

Consider just three small samples of four plugs for a demonstration. If their sample means are 11.0, 10.5, and 11.0 grams, the sampling display gets three dots. Two are stacked at 11.0. The twelve individual measurements belong to the underlying samples. To simulate what happens when the sample size rises, draw larger samples from that same inventory and again keep one mean per sample. The new means should cluster more tightly around the inventory mean.

A randomization distribution answers a different question. For example, a researcher might shuffle randomized treatment labels on observed plant outcomes under a no-effect null model and recalculate a treatment-group difference. Describe what changes across repetitions before naming the distribution.

On a blank sketch, decide first what one horizontal-axis value represents. “Plug weight (g),” “mean weight of 20 plugs (g),” and “difference between treatment means after shuffling (g)” communicate three different experiments. The vertical stack counts how often a value or a small range of values appears.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S05.01 (R081); S05.03 (R089).

TWO READING CHECKS

1. In the plot of sample means, what does a single point represent?

2. Which procedure in the reading changes treatment labels while keeping observed outcomes?

PR60 | Confidence-interval meaning

Trace the confidence claim

PRACTICE FROM THE LAST MEETING

(1) Draw and label a possible sample distribution of ten travel times and compute its mean. (2) Describe how to build a sampling distribution of that mean and sketch how its spread would change for larger independent samples. (3) Give a specific example showing why a larger sample can retain bias.

READ FOR THE COMING MEETING

A town uses a random sample of residents to estimate the proportion who used a new footbridge in the previous month. A report gives a 95% confidence interval of 0.34 to 0.46. The estimate concerns one population proportion. Individual residents have yes-or-no responses, so the endpoints cannot describe the middle 95% of individual responses.

The interval procedure relies on a long-run idea. Under its assumptions, about 95% of intervals built from many fresh random samples by this same method would contain the population proportion. The particular sample has already produced its particular interval. We can report confidence in that estimate without knowing whether it captured the true value.

For an interval built as estimate plus or minus a margin of error, the estimate is halfway between the endpoints. For instance, an interval from 18% to 28% is centered at 23% with a margin of 5 percentage points. Subtracting 18 from 28 gives the full width. Apply that distinction to the footbridge interval when you answer the checks.

To decide whether the usual one-proportion z interval is appropriate, look back at the observations. Suppose a fresh simple random sample of 80 from 5,000 residents has 28 users and 52 nonusers. Both observed counts are at least ten, and sampling without replacement draws far less than one tenth of the population. A voluntary online poll could return the very same counts. The calculation would look identical, but self-selection could distort the estimate of use among all residents.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S06.01 (R095); S06.02 (R099).

TWO READING CHECKS

1. What is the population quantity addressed by this interval?

2. Calculate the margin of error in percentage points.

PR61 | Comparing proportions

Keep the order of the groups

PRACTICE FROM THE LAST MEETING

(1) A poll estimates tree-planting support at 0.58 with margin of error 0.05 at 95% confidence. Give the interval and a contextual interpretation. (2) Explain the repeated-sampling meaning of 95%. (3) Plan a conservative sample size for margin 0.05 at 95% confidence using $z^* = 1.96$.

READ FOR THE COMING MEETING

Two independent random samples ask households in two large districts whether they use curbside recycling. In North, 48 of 80 say yes. In South, 35 of 70 say yes. A planner defines the difference as the North population proportion minus the South population proportion.

The sample estimates are 0.60 and 0.50, so the observed North-minus-South difference is 0.10, or ten percentage points. For an interval estimate, the two samples each contribute their own estimated variability. The standard error combines the separate contributions because the samples were independently selected. Keep the same subtraction order when interpreting the interval.

If an interval for this difference includes zero, the data at that confidence level remain compatible with equal population proportions as well as some differences. The size of the positive and negative values included matters when deciding which practical differences the study can distinguish.

Here is the calculation for a separate pair of independent random samples. Suppose 36 of 100 households in one large district and 48 of 100 in another recycle. Each sample is a small fraction of its population and each has at least ten successes and ten failures. The first-minus-second estimate is $0.36 - 0.48 = -0.12$. The estimated standard error is $\sqrt{[(0.36 \times 0.64)/100 + (0.48 \times 0.52)/100]}$, about 0.069. The variances add even though the two proportions are subtracted. An estimate plus or minus 1.96 standard errors gives an approximate 95% interval from −0.256 to 0.016.

A planner could report that the first district's recycling proportion is estimated, with 95% confidence, to be between 25.6 percentage points lower and 1.6 points higher than the second district's. Reversing the order of the districts changes the signs and swaps which endpoint comes first. The evidence about recycling behavior stays the same.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

In the Student Reference: S05.02 (R084); S06.04 (R106).

TWO READING CHECKS

1. What does a negative North-minus-South difference mean here?

2. Why should each sample contribute to the uncertainty of the estimated difference?

PR62 | Hypothesis-test reasoning

What would surprise us if the claim held?

PRACTICE FROM THE LAST MEETING

(1) Use the recycling counts from PR61 (North: 48 of 80 households, South: 35 of 70) to calculate the unpooled SE and a 95% interval for North minus South. Check observed successes and failures and describe the sampling assumptions. (2) Write a contextual conclusion. (3) Compare that conclusion with one for an interval of (−0.16, −0.04) under the same subtraction order.

READ FOR THE COMING MEETING

A print shop promises that 90% of its jobs finish by noon. A customer suspects the on-time rate is lower. Let p be the population proportion of the shop's jobs that finish by noon. Her hypotheses are H_0: $p = 0.90$ and H_a: $p < 0.90$. She specifies that direction before looking at her sample. Her own sample proportion could fall below 0.90 just by chance.

Suppose a random sample of 60 jobs includes 49 that finished by noon. Model the jobs as independent and assume the shop meets its benchmark. How often would 60 jobs give a result at least this unfavorable? Under the binomial model with 60 trials and success probability 0.90, the probability of 49 or fewer on-time jobs is about 0.034. Counting only samples with exactly 49 on time would leave out the samples that disagree even more strongly with the promise.

At the 0.05 significance level, the sample provides convincing evidence that the proportion of the shop's jobs finished by noon is below 0.90. That 0.034 describes possible samples drawn under the benchmark. A report saying that the shop has a 3.4% chance of meeting its promise has reversed the question. If the customer had suspected a higher on-time rate, larger counts would support her alternative and she would use the upper tail instead.

Source and adaptation: OpenStax, Introductory Statistics 2e, §9.4, p-value reasoning. The print-shop data are a hypothetical SMS example.

TWO READING CHECKS

1. Which quantity is the shop's claim about?

2. Why did the customer count results at or below her observed sample proportion?

PR63 | Errors, power, and decisions

Two possible mistakes in a decision

PRACTICE FROM THE LAST MEETING

Test whether the print shop falls below its 90% on-time benchmark (PR62), using a hypothetical p-value of 0.12. Sketch a null sampling distribution with the relevant tail shaded. Interpret 0.12 in context and write the decision at 0.05. Revise the interpretation for a customer asking about a departure in either direction.

READ FOR THE COMING MEETING

A greenhouse manager treats for pests when evidence suggests the infestation rate exceeds 5%. Her null hypothesis puts the population rate at 5%. A Type I error means she orders treatment even though the benchmark holds. This error costs money and exposes plants to an unnecessary treatment. A Type II error means she overlooks an elevated infestation. The resulting crop damage depends partly on how far above 5% the true rate is.

Suppose she keeps the same sample size, test and sampling design. Lowering the significance level makes it harder to trigger treatment and reduces the probability of a Type I error. For a particular elevated rate, that change generally reduces power, the chance of detecting the elevation.

A larger well-chosen random sample gives her another option. It can increase power for a specified elevated rate while she keeps the same significance level. Choosing a rule requires her to weigh the consequences of the two errors.

For example, suppose planning software estimates 0.78 power to detect a true 9% infestation with her proposed sampling plan. In 100 repetitions with the infestation rate fixed at 9%, she would expect treatment to be triggered about 78 times and a damaging infestation to be missed about 22 times. If the true rate were 6%, she would need a new power calculation. A 6% rate is closer to the benchmark and usually harder for the same test to detect.

Source and adaptation: OpenStax, Introductory Statistics 2e, §9.2, test outcomes and power. The greenhouse decision and planning figures are hypothetical.

TWO READING CHECKS

1. What action would constitute a Type I error here?

2. What does power measure if the true infestation rate is 9%?

PR64 | Chi-square analysis

One table, two sampling designs

PRACTICE FROM THE LAST MEETING

Return to the greenhouse. Make a decision-by-truth table for ‘treat’ and ‘do not treat’ against the benchmark and an elevated rate. Describe each mistake in context and recommend one adjustment if missed infestations are especially costly. Explain why a decision to leave the plants untreated cannot establish the true rate.

READ FOR THE COMING MEETING

A coordinator independently samples 40 riders from each of two bus routes and records whether each rider uses a paper pass. A total of 20 riders report a paper pass. Under the hypothesis of the same pass-use distribution on both routes, the pooled paper-pass proportion is 20/80. The expected number of paper-pass users among the 40 on either route is $40 \times 20/80 = 10$.

The coordinator compares observed counts with expected counts in all four cells. For each cell the chi-square contribution is the squared observed-minus-expected difference divided by the expected count. Large discrepancies in either direction increase the statistic. This is why evidence against the null lies in the right tail.

Here the route samples were selected separately. The homogeneity question asks whether the pass-use distributions match across routes. If she had drawn one sample of riders and recorded both route and pass type on each rider, the design would instead support a test of independence between those two categorical variables.

Suppose 13 riders on the first route and 7 on the second use paper passes. With 40 riders per route, the expected counts are 10 paper and 30 other passes on each route. The observed counts are 13 and 27 on the first route, 7 and 33 on the second. The four contributions are 9/10, 9/30, 9/10 and 9/30. Their sum is 2.40. A two-row, two-column table has $(2 - 1)(2 - 1) = 1$ degree of freedom. The right-tail probability is about 0.12, so at 0.05 these data give insufficient evidence of a difference between routes. This calculation assumes independent random samples, small sampling fractions when sampling without replacement, and expected counts of at least 5 in every cell.

Source and adaptation: OpenStax, Introductory Statistics 2e, §§11.3-11.4, independence and homogeneity. The bus-pass tables are original SMS examples.

TWO READING CHECKS

1. A later survey samples 30 riders on route A and 50 on route B. Of all 80 riders, 60 have no paper pass. Under the null model, what no-pass count do you expect for A? Show the calculation.

2. Why can a below-expected count increase the chi-square statistic?

PR65 | Mid-review assessment and corrections

Finding the first decision that went wrong

PRACTICE FROM THE LAST MEETING

For the later survey, 14 of 30 riders on route A and 6 of 50 on B have paper passes. Build the observed and expected tables. Find the statistic and degrees of freedom. State the sampling and expected-count conditions to check.

READ FOR THE COMING MEETING

A student concludes that a randomly assigned change in the seating arrangement caused a higher mean discussion score for the participants. She then extends her finding to all students in the county. The participants were volunteers from one club. The first sentence could be defensible if a well-controlled comparison provided convincing evidence of a treatment effect. The second sentence requires evidence about population selection that the design does not supply.

A correction should identify the earliest point where the argument loses support. Replacing ‘county’ with ‘club volunteers in the experiment’ repairs the scope. Reviewing a calculation would leave the reasoning problem untouched. A useful transfer example changes the design. If the researchers sampled students at random across the county, which population claim could they now consider?

Now change only how the students enter the study. A random county sample assigned at random to two seating plans could support a claim about the effect for the county population, provided the study is well conducted and the evidence is convincing. If the same county sample simply reports existing seating choices, differences in scores might instead reflect who usually chooses each seat. Ask two separate questions when correcting your response: Who could have entered the study? Who decided which arrangement each student received?

In class you will reattempt selected earlier questions from memory. Save the original response and mark exactly where the new reasoning differs. Then check whether you can recognize the same decision in a fresh example.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. Which feature supports a cause-and-effect claim for the participants?

2. What first error appears in the student’s second sentence?

PR66 | Means and t procedures

Estimating variability when the population spread is unknown

PRACTICE FROM THE LAST MEETING

Select a reattempted free-response part or two related multiple-choice items. Copy the earliest sentence or calculation that went wrong, write a corrected chain of reasoning, and create a new two- or three-sentence example that changes one design feature. Explain whether your revised conclusion still holds.

READ FOR THE COMING MEETING

A technician randomly selects 12 bottles and measures the fill volume in each. She wants an interval for the population mean fill volume. The population standard deviation is unknown, so she estimates the standard error of the sample mean with the sample standard deviation divided by the square root of 12.

The one-sample t procedure accounts for uncertainty in estimating that spread. With 12 independent observations it uses 11 degrees of freedom. A t distribution with few degrees of freedom has heavier tails than the standard normal distribution. At the same confidence level, its critical value leaves room for additional uncertainty.

Twelve is a small sample. Before using the interval, the technician checks the selection process and plots the volumes to look for strong skewness or outliers. A single unusually large fill can affect both the mean and the sample standard deviation. She should discuss its effect and investigate its measurement history before choosing how to analyze the data.

For example, suppose the 12 fills average 503.0 mL with sample standard deviation 3.46 mL. The estimated standard error is $3.46/\sqrt{12} \approx 1.00$ mL. If the conditions hold, a 95% interval uses a t critical value of 2.201 with $12 - 1 = 11$ degrees of freedom. The endpoints are approximately 503.0 ± 2.20, or 500.8 to 505.2 mL. We are 95% confident that the population mean fill volume lies in this range. In class you will meet a related problem with two measurements on each person. Form one difference per person first. The same one-sample t reasoning then applies to the differences, including the check of their distribution.

Source and adaptation: OpenStax, Introductory Statistics 2e, §§8.2 and 10.4, inference for a mean and for paired differences. Bottle data are hypothetical.

TWO READING CHECKS

1. Why is the sample standard deviation used here?

2. What two kinds of evidence matter before using t with 12 observations?

PR67 | Paired versus independent designs

Where the pairs come from

PRACTICE FROM THE LAST MEETING

A random sample of 16 bottles has mean 502 mL and s = 8 mL. The population is large, and a dotplot shows a single roughly symmetric cluster without outliers. Use $t^* = 2.131$ for 95% confidence. Build and interpret the interval, then explain what you would check if an extreme new value appeared.

READ FOR THE COMING MEETING

A crew tests two calibration routines on the same ten sensors. Each sensor experiences both routines, with their order randomized. Some sensors read slightly high under either routine. Comparing the two readings within each sensor removes much of this sensor-to-sensor difference. Define a difference as routine B minus routine A. A positive difference means the B reading was higher for that sensor.

The ten differences are the observations used for a one-sample t analysis of the mean difference. Variation among those differences determines the standard error. Measurements on different sensors still need an appropriate independence assumption.

OpenStax's pain-score example shows why the identifiers matter. One person's scores went from 6.6 to 6.8, while another's went from 8.1 to 6.1. Their after-minus-before changes were +0.2 and −2.0 points. Sorting just those two before scores and two after scores separately would pair 6.6 with 6.1 and 8.1 with 6.8, giving changes of −0.5 and −1.3. Both calculations give the same mean change of −0.9 point. Only the original pairs show how much each individual's score changed. Consequently, the spread of the differences and their standard error depend on keeping the subject identifiers.

Imagine a second crew assigns one set of sensors to routine A and a separate set to routine B. There is no meaningful match between a particular A sensor and a particular B sensor. The crew would estimate a difference of two population means using the separate group means, spreads and sample sizes. Equal group sizes alone do not create pairs.

Source and adaptation: OpenStax, Introductory Statistics 2e, §10.4, Tables 10.14-10.15 (two subjects' paired pain scores). The sensor comparison is an original SMS example.

TWO READING CHECKS

1. What does a B-minus-A difference of −3 units mean?

2. How many observations enter the paired analysis of ten sensors measured twice?

PR68 | Inference-method selection

Choose the quantity before choosing the procedure

PRACTICE FROM THE LAST MEETING

Design two studies comparing reading times on paper and on a screen, one paired and one using independent groups. Define each parameter and subtraction order. Sketch the data table and show which sample means, spreads and sample sizes would enter the analysis.

READ FOR THE COMING MEETING

A parks office can record trail access in two ways. It can mark each sampled trail day as open or closed. Or it can record the number of hours the trail was open. The first version supports a question about a proportion of days. The second supports a question about a mean number of hours. The unit of observation is still a trail day, but the measured response has changed.

Next identify how many groups contributed data and how observations are related. Two trails checked on the same set of dates may have similar weather. In a comparison of opening hours, subtracting the two values within each date can account for that pairing. Two unrelated date samples require a different analysis.

Finally ask what the researcher wants from the data. A confidence interval estimates a quantity and reports uncertainty. A hypothesis test measures evidence against a specified claim. Writing the parameter in words often resolves which operation the question calls for.

Try saying the question without naming a procedure. A conservancy has one random sample of autumn trail days and a measured number of visitors for each day. The population standard deviation is unknown. “How many people visit per day, on average?” leads to a one-sample t interval if the observations meet the independence and shape conditions. “Is mean attendance lower than last year’s posted benchmark of 80?” leads to a lower-tail one-sample t test against 80, assuming the same conditions and treating the posted benchmark as fixed. The measurements and sample are identical in the two questions. The difference is whether you need an estimate or a decision about a specified value.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. Which response would support an estimate of a population proportion?

2. What would be paired in the opening-hours comparison?

PR69 | Regression and predictions

What a prediction leaves unexplained

PRACTICE FROM THE LAST MEETING

Write one question about trail openings answered with a proportion and one answered with a mean. Specify the population, response and parameter for each. Change one question to compare two groups and explain how the sampling or pairing changes the required procedure.

READ FOR THE COMING MEETING

Suppose a fitted line predicts elapsed minutes from the distance of a short delivery. Its slope is 2.4 minutes per additional kilometer. Over the observed distance range, the line predicts that a trip one kilometer longer will take 2.4 additional minutes on average. A trip predicted to take 18 minutes actually takes 21. Its residual is observed minus predicted, or +3 minutes.

A residual plot shows whether the fitted line repeatedly misses in the same direction across parts of the distance range. A curved pattern suggests the linear form misses a systematic change. A growing spread suggests the error size depends on distance. Either feature matters when judging a prediction.

A second summary is the coefficient of determination, r^2. Suppose $r^2 = 0.49$ for a different set of these trips. Using the fitted line instead of predicting the same mean time for every trip accounts for 49% of the observed variation in delivery times. The remaining 51% is variation around the line. A large r^2 can occur even if a few faraway points strongly influence the fit, so examine the scatterplot and residuals as well. This percentage alone cannot tell you whether a particular predicted time is close to the actual time.

If the sample distances ran from 1 to 6 kilometers, a prediction at 22 kilometers extends far beyond the observed range. Traffic routes, stop patterns or travel speeds could change there. The fitted equation still returns a number, but the sample supplies little evidence about its accuracy at that distance.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. Interpret a residual of −2 minutes for one delivery.

2. What would a curved residual pattern suggest?

PR70 | Small investigation: question and plan

Write the plan before observing results

PRACTICE FROM THE LAST MEETING

A model predicts delivery time in minutes as 5 + 2.4d for distance d in kilometers, fitted from trips of 1-6 km. Predict the time at 5 km and find the residual for an actual time of 20 minutes. Sketch a residual plot that would make linear prediction questionable. Evaluate a request for a prediction at 20 km.

READ FOR THE COMING MEETING

Two groups want to compare how quickly different designs of paper tabs can be sorted. They agree that elapsed seconds for a fixed pile is their response. Without a shared rule, one timer might start when a student picks up the first tab while another starts when the instructions end. Define the starting signal, stopping rule and pile size before collecting times.

The design also determines what comparisons are possible. If each participant sorts both designs, record an anonymous pair identifier and randomize the order. If each participant sorts just one design, randomize the design assignment and compare group summaries. Decide in advance how to handle a dropped pile or a stopwatch failure.

A short written plan records the question, the observational unit, variables, collection method and intended analysis. A second group should be able to collect the same kind of data using the plan alone. Keep personal identifiers off the data sheet and choose a task everyone can safely perform.

Here is one design choice worth settling on paper. If everyone sorts design A first and design B second, any practice with sorting is tied to design B. A shorter second time could reflect that practice. Tossing a coin for each volunteer to decide which design comes first gives both designs a chance to benefit from practice. Record that order and the two elapsed times. An analysis of the within-person differences can then address the comparison for these volunteers under the study conditions. A claim about all students at the school would also require a defensible way of selecting them.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. What does a fixed timer start rule contribute?

2. Which column links two trials from one participant without recording the participant’s name?

PR71 | Small investigation: collection and analysis

Keep a record another group can reproduce

PRACTICE FROM THE LAST MEETING

Use your approved in-class plan. Make a blank table with one row per observation, headers, units and condition or pair codes. Write five sentences explaining your population, response, allocation or selection, intended display or calculation, and the scope of your eventual claim.

READ FOR THE COMING MEETING

One team's stopwatch file has values of 8, 9, 11 and 0 seconds. The zero came from a failed timer. Entering it as a real completion time would pull the mean down. Following the team's advance rule, the recorder marks that attempt as missing and records why. The original row remains in the record so another group can see how many attempts were made.

For those four attempts the recorder has three usable times. Their mean is $(8 + 9 + 11)/3 \approx 9.3$ seconds. Including the zero would produce 7.0 seconds and give the false impression of a faster group. Report both the number of valid times and the number of failed attempts. A genuine slow time raises a different issue: if the stopwatch worked and the participant finished, follow the planned rule for completed trials even if the point looks unusual.

A clean table gives every measurement a unit and identifies which rows belong together. For a paired task, use the same subtraction order for every pair. For independent groups, display each group's distribution so that spread and unusual observations remain visible.

For example, suppose a volunteer takes 8 seconds under design A and 10 under design B. With B minus A as the stated order, the difference is +2 seconds. Design B took longer on this trial. Reporting only "a difference of two" would leave the reader guessing about the direction. Before plotting, check a raw pair and a computed difference against each other in this way.

Before making a claim, check the graph against the data table. Count observations, verify a few plotted values and reproduce a summary from the recorded numbers. A convenience sample from one class can support a description of that class's observations. Extending the result to a broader population requires a design that supports the extension.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. How should the failed zero-second trial appear in the record?

2. Why count observations on both the graph and the table?

PR72 | Small investigation: reporting and critique

What a reader needs to judge a claim

PRACTICE FROM THE LAST MEETING

Finish your de-identified table and graph, then write a claim-evidence-limitation paragraph. Report a quantitative comparison with units and enough information about missing trials or labels for another group to reconstruct your result.

READ FOR THE COMING MEETING

A team reports that its volunteers sorted one set of paper shapes faster than another. To evaluate the claim, a reader needs to know what ‘faster’ meant, how the two sets were assigned or ordered, how many trials contributed and how large the observed difference was. A difference of two seconds has meaning only alongside the variability and the measurement rule.

Two questions lead to different kinds of critique. For reproducibility, ask whether someone with the same table could reproduce a graph or summary. For scope, ask which people, tasks or settings the design permits the report to discuss. A team may record its volunteers carefully and still have limited evidence about other classes.

Choose a revision that follows from the critique. If timer operators applied different stopping rules, clarify and standardize the rule. If the volunteers came from a single class, narrow the population claim or design a selection plan for a broader study.

Consider this draft: “Design B is faster for all students.” Suppose eight volunteers each sorted both designs in random order. Their B-minus-A times averaged −2 seconds and ranged from −7 to +3 seconds. A more useful claim is, “Among these eight volunteers, sorting design B took two seconds less on average. The B trial was slower for at least one volunteer.” A reader could verify the mean and range from the eight recorded differences. Deciding whether a similar advantage applies in another class calls for information about how the volunteers were selected.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. Which details help a reader interpret a two-second difference?

2. What kind of critique is ‘Could we recreate this graph from the table?’

PR73 | Practice examination: MCQ

A plan for a timed multiple-choice sitting

PRACTICE FROM THE LAST MEETING

Individually select one limitation of your finished investigation. Propose a feasible design change, explain how it addresses that limitation and rewrite one claim for the original and improved designs.

READ FOR THE COMING MEETING

The course practice section contains 42 multiple-choice questions in 90 minutes. That is a little over two minutes per item on average. Some questions can be answered by recognizing a design or reading a graph. Others require several steps with a calculator. A fixed two-minute cutoff on every item could spend too much time on a short item and too little on a difficult one.

Use one or two time checks to see whether there is room to revisit a flagged item. For example, suppose 26 items have been attempted by minute 54. Sixteen items and 36 minutes remain, or $36/16 = 2.25$ minutes per remaining item before allowing time for a return pass. A brief mark such as “reread graph axis” tells you what to do if you return. Marking one particular uncertainty is more useful than simply recording that an item seemed hard.

During the sitting, keep responses independent and use the technology permitted by your instructor. Afterward, make a short note about where time or uncertainty affected your work. Leave the original choices available for the later analysis.

Source and adaptation: College Board, AP Statistics Course and Exam Description, effective fall 2026, Exam Overview, printed p. 156. The pacing example concerns the SMS practice section.

TWO READING CHECKS

1. What is the average time available per item in this practice section?

2. Why mark uncertainty before scoring?

PR74 | Timed mock, Section II

Make written reasoning visible under time pressure

PRACTICE FROM THE LAST MEETING

Without opening a key or revising your answers, record one time checkpoint and one place where you felt uncertain. Describe whether reading, method selection, calculation or time pressure caused the uncertainty.

READ FOR THE COMING MEETING

The course's free-response practice section has four questions in 90 minutes. Twenty-two and a half minutes per question is an overall average. The parts differ in length, so read each question's verbs and identify what evidence the response must supply. 'Calculate,' 'justify' and 'interpret' each ask for a different product.

For a calculation, label the quantities entered and give units when they apply. For an interpretation, connect the result to the variables or population. For a justification, a reader needs a detail about the actual design. For example, a technician compares two thermometers by using both on each water sample. "The temperatures are paired" only names a method. "Both thermometers measured the same water sample, so subtracting their readings compares instruments at a common temperature" explains why. If different samples were used with each thermometer and there was no matching rule, that argument would fail.

A clear short response can earn credit when a reader can follow the decision and the supporting evidence. On a Chromebook, plain text such as 'sample mean' or 'sqrt(n)' can communicate the intended notation. Use the last few minutes to check that each requested part has an identifiable response and that the direction of every comparison is clear.

Source and adaptation: College Board, AP Statistics Course and Exam Description, effective fall 2026, Exam Overview, printed p. 156, and task verbs, p. 159. The thermometer example is an original SMS illustration.

TWO READING CHECKS

1. What are the total time and approximate average time per question?

2. How does an interpretation differ from a calculation?

PR75 | Practice-exam analysis

Find the first lost point

PRACTICE FROM THE LAST MEETING

Keep the submitted responses as written. Without consulting solutions, estimate how you divided the 90 minutes, identify a part that needed more or less time and write one change you could try on a future timed response.

READ FOR THE COMING MEETING

A student selects a sound method, enters an incorrect sample size in a calculator and then interprets the resulting interval coherently. A single input error has affected the final number. Another student obtains the right number but cannot explain which population it describes. Their final scores could match even though they need different practice.

When you score your practice responses, record where credit was earned and where the first unsupported step appeared. Keep the original wording next to the correction. Classify the main problem as a conceptual misunderstanding, method choice, condition check, calculation or communication issue. One response can involve more than one category.

Look for repeated errors across questions and distinguish them from independent slips. If the same missing parameter affects several responses, one focused change may improve several future answers. Use rubric language to explain which part of a revised response now supplies the missing evidence.

For example, suppose a student reports a confidence interval of 0.31 to 0.45 and says that between 31% and 45% of the surveyed households selected recycling pickup. The survey itself already gives one known sample percentage. If households were randomly sampled from the town and the conditions hold, the range estimates the proportion of all town households that would select pickup. Rework the interpretation while preserving any correctly computed endpoints. If an input produced the wrong endpoints, follow that same error forward and recompute each affected quantity.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. Which error category best fits the incorrectly entered sample size in the first example?

2. Why retain the original response next to the correction?

PR76 | Targeted reteaching and transfer

Test a correction in a new setting

PRACTICE FROM THE LAST MEETING

Choose one scored free-response question. Mark each earned and missed rubric component, revise the missed work and justify each point you expect the revised answer to earn. Identify the earliest error and one fresh task that would check the repair.

READ FOR THE COMING MEETING

A student can recite the corrected solution to a paired-data question involving two readings from each sensor. A new question compares two independent sets of sensors. The same topic appears in both questions, but the observational structure changes. To show transfer, the student must recognize that the new design calls for separate group information.

Start with one error category from your practice analysis. Write the rule that would have helped, then identify the cue in a new problem that activates the rule. For example, ‘the same sensor under both settings’ signals a meaningful pair. ‘Ten sensors in each group’ establishes the group sizes and leaves the pairing question open.

Here is a way to test your own rule. One team reports four readings for each of two settings. For a paired experiment, a sensor ID must link each of the four readings under A to its companion under B. Only then can you calculate four changes and describe their spread. If a different set of four sensors supplied each column, two columns of equal length would look much the same on paper. In that case the separate spreads of the two groups belong in the analysis. A correct method choice must survive that change in the description.

Attempt the new problem before consulting an answer. Compare your reasoning as well as the result with the key. If the result is correct for the wrong reason, create a nearby case in which that reason would fail. That second case helps locate what still needs attention.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. What piece of evidence would establish a meaningful pair in the sensor example?

2. Why could a correct result still leave a gap in understanding?

PR77 | Cumulative explanation conference

Explain a method to a partner

PRACTICE FROM THE LAST MEETING

Use the two error categories identified in class. From the corresponding unit-review questions, complete one previously unused question in each area. Label the cue that determines your method, show your reasoning and compare the two results with the keys.

READ FOR THE COMING MEETING

In an explanation conference, a partner can ask a question that a written calculation leaves unresolved. Suppose someone says, ‘I used a confidence interval because the calculator had that option.’ A partner might ask what quantity was estimated, which observations supplied the data and what feature of the design justified the method. These questions make the statistical reasoning audible.

A short reference sheet can support the exchange. For a one-proportion interval, it might list the population proportion, selection process, independence check, success and failure counts, calculation and contextual interpretation. Each item is a cue for a sentence the speaker can defend with evidence from the problem.

For example, suppose an SRS of 50 of a town’s 2,000 library cardholders includes 18 who use audiobooks. A speaker could say, “I want the proportion of all cardholders who use audiobooks. The sample was random and 50 is below 10% of the population. We observed 18 users and 32 nonusers, both at least 10, so the large-count condition for a one-proportion z interval holds.” The estimate is $18/50 = 0.36$. The 95% margin of error is approximately $1.96 \times \sqrt{(0.36 \times 0.64/50)} = 0.133$. We are 95% confident that between about 23% and 49% of the town’s library cardholders use audiobooks. If just three cardholders in another sample used audiobooks, the small success count would call the normal approximation into question.

Practice saying the whole chain aloud using one familiar example. Then let your partner alter one feature, such as changing independent samples to matched observations. An explanation is stronger when you can say which part of the method changes and explain the design reason.

Source and adaptation: Original SMS example. College Board, AP Statistics Course and Exam Description, effective fall 2026, Learning Objectives 3.3.B-C and 3.4.A.

TWO READING CHECKS

1. What question could a partner ask first when the speaker names a calculator procedure?

2. What makes an independence check informative in an explanation?

PR78 | Readiness check and final consolidation

Use the final check to plan the next independent attempt

PRACTICE FROM THE LAST MEETING

Revise your reference sheet using two questions your partner asked. Complete one unresolved checkpoint free-response part, then annotate where your answer identifies the method, supporting evidence and contextual meaning.

READ FOR THE COMING MEETING

A student reviewing a scored question may recognize every line of the solution and still struggle to begin the same kind of question alone. A stronger readiness check begins with a closed-reference attempt. Record the first method choice and a brief reason for it before comparing with the key.

After scoring, distinguish secure skills from those that need another independent attempt. ‘Review inference’ is too broad to guide a study session. ‘On an unfamiliar paired study, define the within-person difference, check the design and interpret a mean-difference interval’ gives a specific task and a way to judge the result.

A goal based on your own work might instead concern regression. Suppose your last answer called a positive residual an overprediction. Before seeing a key, choose a new delivery from the workbook and write down the observed and predicted times with units. Your success criterion is that the subtraction and the sentence agree: if the actual trip took longer, the residual must be positive and the model predicted too few minutes. If only one part is right, your next attempt should target that specific mismatch.

In the final meeting you will revisit difficult questions from across the course and a selected written response. Use evidence from your own work to name one remaining uncertainty, one suitable practice item and the reasoning a successful answer must include. A short targeted plan can guide any independent study you choose afterward.

Source and adaptation: Original SMS course preparation and review, aligned to previously taught course content. Classroom cases are illustrative unless students collect their own data.

TWO READING CHECKS

1. Why record a method choice before looking at the key?

2. What makes the paired-study goal specific enough to check?