ST. MARY'S SCHOOL

Medford, Oregon / 2026–2027

Introduction to Statistics

Student Reference

Explanations, worked examples, and visual guides for Applied Math Mod 2.



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

Adapted for St. Mary's School from its AP Statistics materials and Introductory Statistics 2e by Barbara Illowsky and Susan Dean, OpenStax, Rice University. Examples, activities, and sequencing are customized for Applied Math. The original textbook is free at OpenStax.org.

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

OpenStax textbook © 2026 Rice University. This adaptation uses CC BY-NC-SA 4.0. School insignia and previously approved artwork retain their original ownership.

https://creativecommons.org/licenses/by-nc-sa/4.0/

USING THIS BOOK

Working with statistical evidence

Keep this reference for explanations, worked examples, and study. Your workbook contains class activities, homework, the next lesson's reading and two self-checks. Its formula appendix is at the back. Match the Day and Unit labels in the two books.

When	What to do
The night before	Complete the named homework sheet. Turn the page to the reading for the next meeting. Answer its two self-checks, compare with the model answers, and keep that sheet in your binder. Read PR01 in class on Day 1.
At the start of class	Complete the entry ticket. Use the discussion to repair a mistake or connect the reading to the new question.
During the lesson	Follow the teaching and worked example here, then immediately practice in the Workbook. Complete the independent response and exit check yourself.
When revising	Name the group and measured outcome. Cite numerical evidence and explain how it supports the claim. Keep the conclusion within the limits of the design.

An answer that stands on its own

Suppose a dispatcher records 20 shuttle runs on each of two routes. The median time was 14 minutes on Route A and 19 minutes on Route B. The middle half of times spanned 3 minutes on A and 8 minutes on B. In these observations, Route A was typically faster and its times were less variable.

> **CHECK YOUR WORDING**
>
> Cover the question. Can a new reader tell which routes were compared, what was measured, and why the conclusion follows?

Collecting and reporting class data

Before collecting, agree on one observation, the measurement rule, and how the class will pool records. Save the raw counts or measurements with the class total. In your conclusion, distinguish what your group actually observed from a claim about a wider population.

School and regional examples use constructed data unless the task calls for live collection. Label any supplied fallback data when using them in place of class observations.

CONTENTS

Find a lesson

R identifies a page in this reference. PR identifies a preparation reading in your workbook. Each lesson begins at the page listed below.

Day	Lesson	Reference
01	What exactly are we measuring?	R005
02	What a display makes visible	R007
03	Compare a typical value and its spread	R009
04	Compare observations across different groups	R011
05	Which claims does the study design support?	R013
06	Checkpoint and revision: describe and compare	R015
07	Reading a scatterplot	R017
08	Predicting a value and checking the error	R019
09	Does traffic change the recommendation?	R021
10	Where does the model have support?	R023
11	Writing a model recommendation	R025
12	One sample, one estimate	R027
13	Why sample size changes spread	R029

CONTENTS

Find a lesson | continued

R identifies a page in this reference. PR identifies a preparation reading in your workbook. Each lesson begins at the page listed below.

Day	Lesson	Reference
14	What 95% confidence means	R031
15	Using an interval to make a decision	R033
16	How far can an estimate reach?	R035
17	Simulate a benchmark	R037
18	How strong and how large?	R039
19	Two ways a decision can go wrong	R041
20	Power at a stated alternative	R043
21	Defend a decision rule	R045
22	Repair a statistical explanation	R047
23	Write a decision memo from a new case	R049
24	Independent performance task	R051
25	Revision and independent transfer	R053

Quick reference

R055 Choosing a defensible claim

R056 Notation and calculations

Use the same day number across the package

For example, read PR08 after Day 7. At Day 8, complete ET08, study the Day 8 lesson here, and use Day 8 classwork in the Workbook. Complete HW08 after class, then read PR09.

UNIT 1 / DAY 01 • DATA AND STUDY DESIGN / T1

What exactly are we measuring?

Define a measurable question and write a claim whose population, variable, and evidence are clear.

A statistical question asks about a group and anticipates variation among its members. "What percentage of this class's time estimates meet our accuracy rule?" is statistical because trials may have different outcomes. Before collecting data, define a case: the person, object, or event that contributes one record. The variable is the characteristic recorded for each case. One shuttle run might have several variables: travel time in minutes, arrival time on the clock, and whether it arrived late.

A quantitative variable records a numerical count or measurement, such as elapsed seconds. A categorical variable places a case in a group, such as hit or miss. Give units for measurements and define the categories. A rule such as "more than five minutes late" excludes a run exactly five minutes late. The measurement and boundary rule must stay the same for every case.

The population is the whole group your question concerns; a sample is the subset studied. A statistic summarizes a sample, while a parameter summarizes the population. For example, to study the mean GPA of last year's graduates, the population is all those graduates, a sample is the graduates selected for study, and the variable is each graduate's GPA. The sample mean is a statistic used to estimate the population mean.

Adapted from OpenStax, §1.1, Example 1.2, p. 8. Access for free at openstax.org.

A proportion divides the count meeting a condition by the total number of cases being described. Multiply by 100 for a percentage. State what the denominator counts: 12 successful trials out of 20 trials is 60% of trials, not necessarily 60% of students. Before extending a result to a larger population, ask how the cases were selected.

Three entries from a fictional run log

Run	Scheduled arrival	Actual arrival	Travel time
A	8:10	8:13	17 min
B	8:10	8:16	15 min
C	8:10	8:15	18 min

With the rule "more than five minutes late," only run B is late. Clock times and trip duration measure different things.

1

UNIT 1 / DAY 01 • DATA AND STUDY DESIGN / T1 • CONTINUED

What exactly are we measuring?

Two shuttle headlines

A dispatcher proposes "The morning shuttle is unreliable." A driver proposes "Most trips finish quickly." An illustrative log records trip duration and minutes behind schedule for eight morning runs. All lasted less than 20 minutes; three arrived over five minutes late.

1. Set the case to one morning run. Trip duration measures elapsed travel time. Minutes behind schedule measures lateness against the timetable.

2. Define a late run as one that arrives more than five minutes late. The late fraction in the log is 3/8 = 37.5%.

3. All eight trips took less than 20 minutes, so the two observations can coexist. We do not know how the eight runs were selected.

All eight logged morning runs took under 20 minutes, and three arrived more than five minutes behind schedule. These observations describe the logged runs.

KEY IDEA	Specify the case, the measurement rule, and the denominator before interpreting a statistic.
WATCH FOR	A short trip can still arrive late. Verify that the recorded variable answers the question being asked.

Sources and adaptation. Concepts adapted from OpenStax, Introductory Statistics 2e, §1.1, pp. 5–8; GPA illustration adapted from Example 1.2 (p. 8). Shuttle and stopwatch cases are original classroom examples.

UNIT 1 / DAY 02 • DATA AND STUDY DESIGN / T2

What a display makes visible

Build a dotplot and histogram from the same quantitative data, then describe a feature supported by the display.

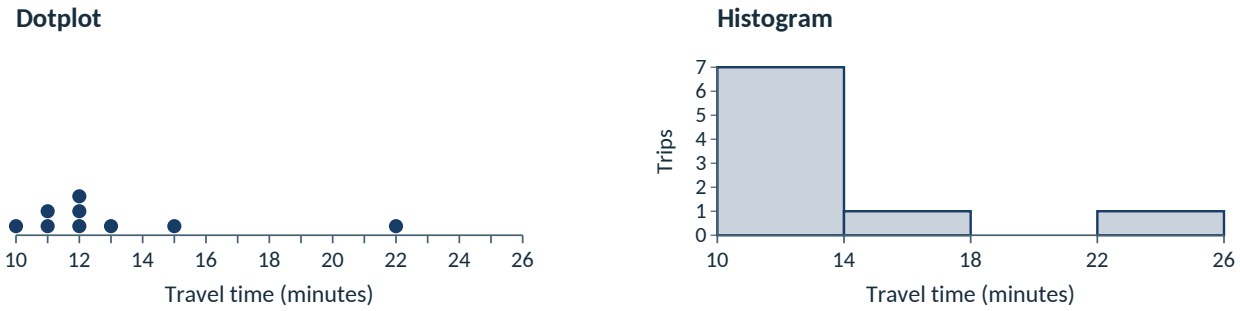
A distribution shows which values a variable takes and how often they occur. Begin with the cases, variable, and units. A dotplot places one dot for each observation above its value on a numerical scale. Stack repeated values vertically: three dots at 12 minutes represent three separate 12-minute trips. Keep equal numerical distances equally spaced. A horizontal gap means no observations occurred at those values.

A histogram groups quantitative values into intervals called bins. With equal-width bins, each bar's height gives the number of cases in that interval; adjoining bins touch. We use a left-inclusive, right-exclusive rule: [10,14) means at least 10 but less than 14. Thus 14 goes in the next bin. Label the horizontal axis with the measured variable and units, and the vertical axis with the count of cases. The bar counts must add to the sample size.

Describe what the graph shows before giving it a name. Locate clusters, gaps, and isolated values with numbers and units. A distribution is right-skewed when its longer tail extends toward larger values; left-skewed means the longer tail extends toward smaller values. A roughly symmetric distribution has similar shapes on either side of its center. A small sample may have no simple shape.

The mean equals the sum of all measurements divided by the number of observations. It summarizes a center, but it cannot show every cluster, gap, or extreme. Histograms simplify a display by grouping values, so the exact original measurements usually cannot be recovered from the bars. Changing the bin boundaries may also change the apparent shape. Inspect the individual values when a small data set is available.

Minutes for nine illustrative trips



The dotplot keeps individual times. The histogram counts trips in four-minute intervals.

1

UNIT 1 / DAY 02 • DATA AND STUDY DESIGN / T2 • CONTINUED

What a display makes visible

The 22-minute run

Illustrative trip times in minutes are 10, 11, 11, 12, 12, 12, 13, 15, and 22.

1. The dotplot has three stacked dots at 12. Eight of the nine times lie between 10 and 15 minutes, with one at 22.

2. Use equal-width bins [10,14), [14,18), [18,22), and [22,26). Include the left endpoint and exclude the right. The counts are 7, 1, 0, and 1; the 22-minute run belongs to [22,26).

3. The mean is 118/9 ≈ 13.1 minutes. The empty interval and isolated long trip still matter to someone planning a connection.

Most of these nine runs finished in 10–15 minutes. One took 22 minutes, leaving a gap and a tail on the long-time side.

KEY IDEA Read the horizontal scale as a measurement and the vertical stack or bar as a count.

WATCH FOR A dot at 12 represents one 12-minute trip. It does not represent 12 trips.

Sources and adaptation. Display concepts adapted from OpenStax, Introductory Statistics 2e, §1.1 classroom dotplot activity (pp. 5–6) and §2.2 (pp. 75–77); mean from §2.5 (pp. 98–99). Numerical trip and wait examples are original practice data.

UNIT 1 / DAY 03 • DATA AND STUDY DESIGN / T2

Compare a typical value and its spread

Choose summaries for two distributions and support a comparison with both center and spread.

A center describes a typical value; a measure of spread describes how much observations vary. Use both when comparing groups. The mean is the sum divided by the count, so every value affects it. To find the median, order the observations: take the middle value for an odd count, or average the two middle values for an even count. The median depends on positions in the ordered list. Moving an already-largest value farther out changes the mean but leaves the middle positions unchanged.

The range is maximum minus minimum. It uses only the endpoints and can change greatly with one extreme value. The interquartile range, IQR = Q3 − Q1, describes the width of the middle half. In this course, split the ordered data into lower and upper halves; omit the overall median when the count is odd. Find Q1 as the median of the lower half and Q3 as the median of the upper half. State this convention when using software, because quartile methods can differ.

Standard deviation (SD) describes spread around the mean and has the original measurement units. It is zero only when all values are identical, and increases as values become more spread out from their mean. It is not the average signed difference: positive and negative differences would cancel. A calculator uses squared differences to obtain SD. This course supplies SD or uses technology; interpreting it matters more here than doing the formula by hand.

OpenStax gives 15 student reports of daily exercise. The five-number summary is 0, 20, 40, 60, and 300 minutes. The median of 40 minutes and the IQR of 40 minutes describe the typical report and the middle half. The 300-minute report still matters: it stretches the range to 300 minutes. With roughly symmetric data and no extreme values, a mean and SD can be useful instead.

Adapted from OpenStax, §2.3, Example 2.22, pp. 93–94. Access for free at openstax.org.

Same eight runs, one changed duration

Longest run	Mean	Median	IQR
16 min	9.5 min	8.5 min	3.5 min
24 min	10.5 min	8.5 min	3.5 min

Only the longest run changes in this what-if example. The other seven measurements stay fixed.

1

UNIT 1 / DAY 03 • DATA AND STUDY DESIGN / T2 • CONTINUED

Compare a typical value and its spread

A shuttle delay changes one average

Eight illustrative shuttle durations in minutes are 6, 7, 8, 8, 9, 10, 12, 16. Replace the 16-minute run with a 24-minute run in a what-if calculation.

1. Originally, the sum is 76 minutes and the mean is 9.5. Raising one observation by 8 minutes raises the total to 84 and the mean to 10.5.

2. The middle two times are still 8 and 9, so the median stays at 8.5 minutes.

3. Q1 = (7 + 8)/2 = 7.5 and Q3 = (10 + 12)/2 = 11. The IQR is 3.5 minutes both times. The maximum changes and the range increases from 10 to 18 minutes.

Changing the longest of eight shuttle runs raised the mean by one minute. The median and IQR stayed fixed, while the longer run remained important for riders.

KEY IDEA Use comparable evidence about typical value and variability, then qualify the comparison by the observed groups.

WATCH FOR A larger maximum does not by itself imply a larger IQR. Locate the quartiles.

Sources and adaptation. Center and spread explanations adapted from OpenStax, Introductory Statistics 2e, §2.3 (pp. 86–87), §2.5 (pp. 98–100), and §2.7 (pp. 107–109). Shuttle, frog, and waiting-time data are original classroom examples. Textbook example adapted from OpenStax, §2.3, Example 2.22, pp. 93–94. Five-number summary and exercise context retained; interpretation refocused on resistance.

UNIT 1 / DAY 04 • DATA AND STUDY DESIGN / T3

Compare observations across different groups

Calculate and interpret a signed standardized score using the mean and SD of the relevant group.

A raw difference tells how far a value lies from its group mean in the original units. Whether that difference is large depends on the group's usual variation. A standardized score divides the signed difference by that group's positive standard deviation: z = (value − group mean) / group SD. Always use the mean and SD for the same reference group.

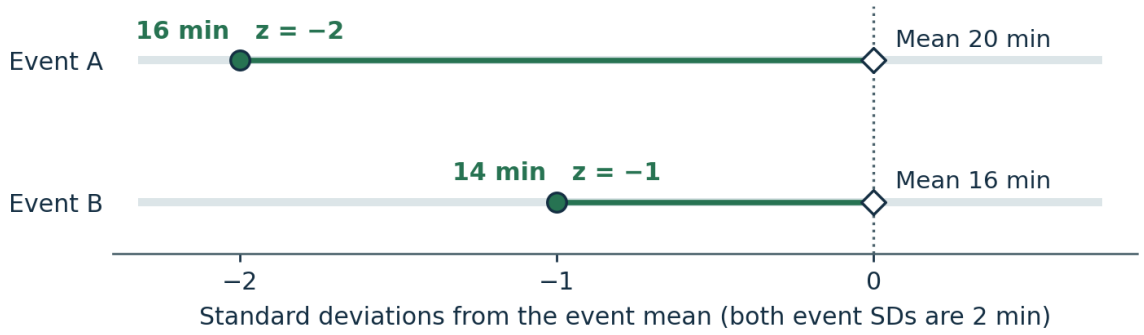
For a concrete example adapted from OpenStax, a supermarket's waiting times have mean 5 minutes and SD 2 minutes. Rosa waits 7 minutes: (7 − 5)/2 = +1, one SD above the mean. Binh waits 1 minute: (1 − 5)/2 = −2, two SDs below the mean. Binh's raw difference is −4 minutes; the z-score expresses that difference as two group-sized SD steps below the mean.

Adapted from OpenStax, §2.7, pp. 107–108. Access for free at openstax.org.

A positive z-score means above the mean, a negative score means below, and zero means at the mean. The absolute value gives distance in SDs. The measurement units cancel in the division, so z itself is unitless; interpret the distance as a number of standard deviations. If SD is zero, every group value is the same and this division is undefined.

To recover a raw value, reverse the calculation: value = group mean + z × group SD. Standardization does not require a bell-shaped distribution. However, a z-score alone does not tell the percentage of observations below the value; that requires more information about the distribution. Below-average time may be desirable, while below-average distance may not be. Use the context when explaining a comparison.

Time relative to each runner's own group



Dots mark the observed times and open diamonds mark the group means. Runner B is faster by the clock, while runner A is farther below the relevant mean.

1

UNIT 1 / DAY 04 • DATA AND STUDY DESIGN / T3 • CONTINUED

Compare observations across different groups

Two runners and two reference groups

In a fictional comparison, Runner A takes 16 minutes in an event with mean 20 minutes and SD 2 minutes. Runner B takes 14 minutes in a different event with mean 16 minutes and SD 2 minutes. Compare their recorded times and their positions within their own event groups.

1. For A, $(16 - 20)/2 = -2$. A's time was two SDs below the mean for A's event.

2. For B, $(14 - 16)/2 = -1$. B's time was one SD below the mean for B's event.

3. B ran the shorter clock time. A finished farther below the mean of the relevant event group.

Runner B has the shorter recorded time, 14 rather than 16 minutes. Relative to their own event groups, A's time is two SDs below the mean and B's is one SD below. A is farther below its own reference mean; different event conditions prevent declaring one runner universally better.

KEY IDEA Always name the reference group; identify the sign, distance in SDs, and practical meaning.

WATCH FOR A z-score is not a percentile. A comparison of SD distances alone does not give a percentage of runners beaten.

Sources and adaptation. Standardized distance and the supermarket example adapt OpenStax, Introductory Statistics 2e, §2.7, pp. 107–108, and §6.1, pp. 336–337. The workbook uses the three instrument prices, means, and SDs from chapter 2, exercise 104, pp. 146–147, with new interpretation questions. Runner and packing cases are classroom examples.

UNIT 1 / DAY 05 • DATA AND STUDY DESIGN / T1

Which claims does the study design support?

Decide whom a study can describe and whether a difference can reasonably be attributed to a treatment.

First ask how the cases entered the study. Random selection uses a chance mechanism to choose cases from a defined population. For example, number every student on a complete roster and use a random-number generator to select students. If selected students respond, the sample can support an estimate for that rostered population. The estimate may still differ from the population value because different samples contain different people.

An open poll attracts volunteers. People with strong opinions may respond more often, even when many people answer. A random sample can also lose groups through nonresponse. In an OpenStax fieldwork example, the interviewer finds nobody home at four selected addresses. If those residents work different hours from the people reached, their answers may differ. A return visit gives the selected households another chance to answer.

Adapted from OpenStax, §1.4, Example 1.22(b), p. 38. Access for free at openstax.org.

Next ask whether researchers assigned a condition. An observational study records existing choices or characteristics. An experiment imposes treatments, the conditions being compared, and measures a response. For a reminder study, the treatment might be reminder versus no reminder; the response is whether a book is returned on time. Random assignment uses chance to place participants into treatment groups. It helps balance other influences on average, but does not guarantee identical groups in one experiment.

A confounding variable is another factor that differs across the compared groups and could help explain their response difference. In a self-selected checklist study, experience could affect both checklist use and packing time. Random assignment makes a causal comparison more defensible when other conditions are comparable. An observed treatment difference could still arise by chance, so later lessons examine how strong the evidence is. Selection supports population scope; assignment supports causal reasoning. A study can use either, both, or neither.

1

UNIT 1 / DAY 05 • DATA AND STUDY DESIGN / T1 • CONTINUED

Which claims does the study design support?

Two distinct uses of chance

Study feature	Question it addresses	Check first
Random selection	Who can a sample describe?	Who is on the roster? Who responded?
Random assignment	Can we study a treatment effect?	Which cases were assigned? What else differed?
Neither	What was observed in these cases?	How could selection or confounding matter?

A study can use either process, both, or neither. Follow the actual design.

Does a reminder increase returns?

In an illustrative study, 40 volunteer borrowers enter a trial. A shuffle assigns 20 to receive a reminder and 20 to receive none. Sixteen and twelve, respectively, return a book on time.

1. Calculate on-time rates: 16/20 = 80% with a reminder and 12/20 = 60% without it. The difference is 20 percentage points.

2. Assignment by chance supports studying the reminder's effect among these volunteers, assuming the groups were treated alike otherwise. A sample difference can still arise by chance.

3. The borrowers volunteered. Other library users, such as those who rarely borrow, may respond differently.

In this volunteer trial, the reminder group's on-time return rate was 80%, compared with 60% without reminders, an observed difference of 20 percentage points. Random assignment supports evaluating a reminder effect among these participants. Further analysis must assess chance variation, and volunteering limits extension to all borrowers.

KEY IDEA Trace how people entered the study and, separately, how any treatment was assigned.

WATCH FOR Randomly assigning volunteers does not turn them into a random sample of the population.

Sources and adaptation. Study design adapts OpenStax, Introductory Statistics 2e, §§1.2 and 1.4 (pp. 34–38). The missed-households case adapts Example 1.22(b), p. 38. Reminder, shuttle, and checklist studies are classroom cases.

UNIT 1 / DAY 06 • DATA AND STUDY DESIGN / T1–T3

Checkpoint and revision: describe and compare

Independently use distribution summaries, relative standing, and study design to write and revise a supported claim.

Use the worked example for feedback and revision after the individual checkpoint or memo.

A statistical answer should let a reader understand the evidence without seeing the question. Name the observed groups and the measured outcome. When comparing distributions, report both a center and a spread in the same units. "Pine's median wait was 6 minutes, compared with Oak's 9 minutes" gives a checkable comparison; "Pine was better" does not explain which outcome or criterion matters.

Choose the calculation that answers the question. A proportion describes the share of cases meeting a condition. A median describes a middle value, and an IQR describes middle-half spread. A z-score describes an observation relative to its own group mean and SD. A shorter raw time and a more negative z-score answer different questions, so identify which comparison you intend.

Finally, match the conclusion to the study. A result from volunteers describes those respondents and may not represent all residents. Random selection can support a population estimate, with uncertainty. Random assignment can support a causal comparison, but the observed difference still needs evaluation. Neither a larger mean nor a larger median means every observation in one group exceeds every observation in the other.

For revision, cover the original question and read your answer alone. Circle the group, underline the numerical evidence, and check the final inference. Repair the first missing link: a wrong denominator, a missing unit, an unexplained comparison, or a claim beyond the design. Keep correct work and add only the detail needed to make the reasoning clear.

Which evidence answers which question?

Question	Evidence to use	An interpretation needs
Which group is typically faster?	Comparable centers in minutes	Group names and direction
Which group varies more?	Comparable SDs or IQRs	Spread with units
How far from its own average?	Signed z-score	Group and SD direction
Who else could this describe?	How cases entered the study	Population scope and selection limits
What caused a difference?	How conditions were assigned	A causal claim the design supports

Use the visual for feedback and revision after the individual checkpoint.

1

UNIT 1 / DAY 06 • DATA AND STUDY DESIGN / T1–T3 • CONTINUED

Checkpoint and revision: describe and compare

Revision clinic after the checkpoint

A draft about two observed transit stops reads, "This one is better." The illustrative medians are 6 and 9 minutes at Pine and Oak, respectively; the IQRs are 4 and 2 minutes.

1. Replace the ambiguous word "one" with the stop name and the outcome. The observed median wait was three minutes shorter at Pine.

2. Add spread. The middle-half span was 4 minutes at Pine and 2 at Oak.

3. Keep the claim with the observed waits. The records alone do not explain what caused the difference.

Among the observed waits, Pine's median was shorter than Oak's (6 vs. 9 minutes), but its middle half was more spread out (IQRs of 4 vs. 2 minutes).

KEY IDEA	An interpretation can be concise and still give the reader a group, an outcome, checkable numerical evidence, and an honest scope.
WATCH FOR	Do not replace "average" with "every." A difference in group centers can coexist with overlapping distributions.

Sources and adaptation. Unit review draws on OpenStax, Introductory Statistics 2e, §§1.1–1.4, 2.3, 2.5, and 2.7. Transit-stop, trail-survey, packing, and repair examples are original practice cases.

UNIT 2 / DAY 07 • PREDICTION AND ROBUSTNESS / T4

Reading a scatterplot

Describe a relationship between two measurements and name one reason it might arise.

A scatterplot displays two quantitative variables measured on the same cases. Each point keeps one case's two measurements together. For a shuttle trip, the pair (3, 16) means three miles and 16 minutes. Sorting the distance and time columns separately would break those pairs and could create a relationship that the actual trips never showed.

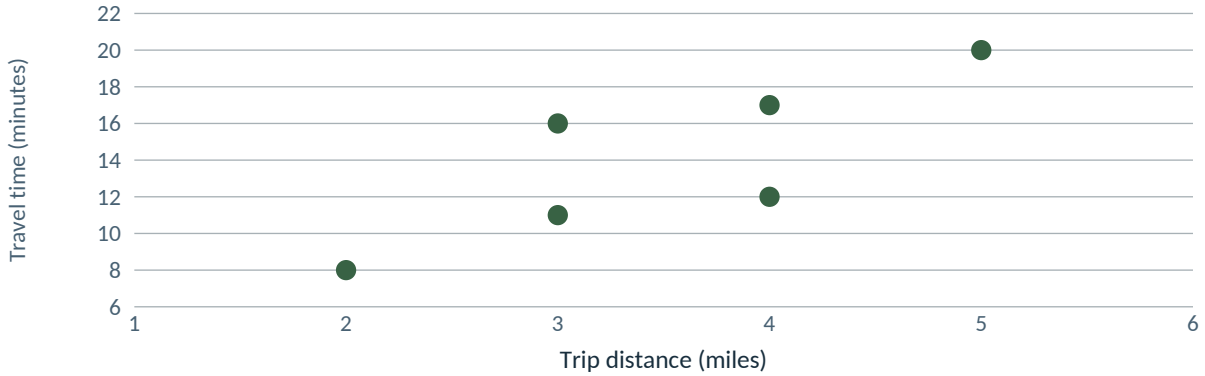
Put the explanatory variable, used to help explain or predict a response, on the horizontal axis. Put the response variable on the vertical axis. These roles describe the question being asked; calling distance explanatory does not establish that changing distance causes a particular change in time. Label both axes and give their units.

Describe the direction, form, and strength of the pattern. In a positive association, larger explanatory values tend to accompany larger responses; in a negative association, larger explanatory values tend to accompany smaller responses. Form asks whether the pattern is roughly straight or curved. Strength asks how closely points follow that pattern. Also identify unusual points or separate groups. Steepness alone does not measure strength.

Use evidence from the points. Two trips with the same distance can take different times, so a positive association does not promise an exact time. Here both three-mile and four-mile trips vary by five minutes. A useful description names the cases and variables, gives the overall pattern, and supports it with a comparison.

Finally, ask how the cases were collected. Existing shuttle routes may differ in traffic, stops, and distance. Such outside variables can help explain the association. A scatterplot alone does not separate those influences or establish a cause.

Distance and time for six shuttle trips



Two observations at three miles and two at four miles show variation among trips of the same length.

2

Reading a scatterplot

Two trips of the same length

Six illustrative shuttle trips give the pairs (2, 8), (3, 11), (4, 12), (3, 16), (4, 17), (5, 20), with distance in miles first and travel time in minutes second.

1. Read the point (3, 16) as one three-mile trip that took 16 minutes.

2. Travel time generally rises from the two-mile trip at 8 minutes to the five-mile trip at 20 minutes.

3. The two three-mile trips took 11 and 16 minutes. Distance leaves a five-minute difference to explain.

In this illustrative six-trip shuttle log, longer trips generally took more time, with a roughly upward pattern. The two three-mile trips differed by five minutes, so distance alone would be insufficient for an exact schedule.

KEY IDEA Each point represents one case. Support an overall description with at least one comparison of actual points.

WATCH FOR An upward pattern leaves open several explanations. Ask how cases were selected and which other variables could affect both measurements.

Sources and adaptation. Explanation adapted from OpenStax, Introductory Statistics 2e, §12.2, "Scatter Plots," printed pp. 620–623. Basketball independent practice adapts Try It 12.5, pp. 621–622, preserving its data and revising the question to distinguish association from causation. Shuttle and bead examples are SMS instructional cases.

UNIT 2 / DAY 08 ● PREDICTION AND ROBUSTNESS / T4

Predicting a value and checking the error

Use a given line for an in-range prediction and interpret the sign and units of a residual.

A prediction line summarizes a roughly linear relationship. Write it as predicted y = a + bx: x is the explanatory value, predicted y is the model's response, a is the intercept, and b is the slope. Substituting the input gives a prediction, not a record of what actually happened. A fitted line has been chosen from data; a supplied practice line lets us learn how to use and check a model without calculating the fit ourselves.

A worked example in OpenStax gives a tax preparer's fee: $31.50 per job plus $32 per hour. Two more hours add exactly $64 under that price rule. For the model predicted travel minutes = 4 + 2d, two more miles add four predicted minutes. The slope describes the change in predictions; actual trips can differ by more or less.

Adapted from OpenStax, §12.1, Example 12.3, p. 619. Access for free at openstax.org.

The intercept is the prediction when the input equals zero. The equation above returns four minutes at zero miles. If the data contain only two- to six-mile trips, that intercept is needed to position the line, but the observations do not establish a four-minute travel time for a zero-mile trip.

A residual compares an observation with its own prediction: residual = observed response − predicted response. A positive residual lies above the line and means the model underestimated the response. A negative residual lies below the line and means the model overestimated it. The size of the residual, ignoring its sign, gives the size of this miss.

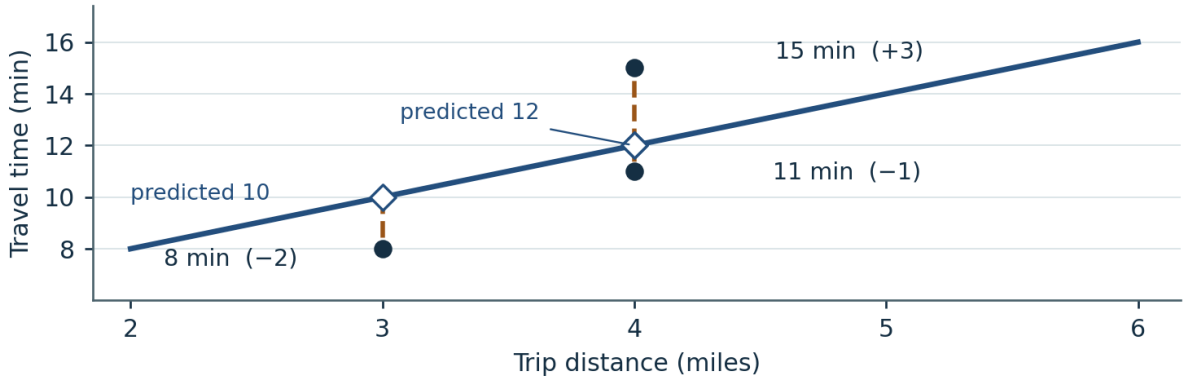
Use response units for a residual and response units per explanatory unit for a slope. A residual of −3 minutes is a three-minute overestimate, not a negative travel time. One small residual does not establish that a model predicts all cases well; inspect several observations and their conditions.

UNIT 2 / DAY 08 • PREDICTION AND ROBUSTNESS / T4 • CONTINUED

2

Predicting a value and checking the error

Distance, predictions, and signed residuals



Diamonds on the supplied line show predicted times. Filled dots show observations. Parentheses give observed minus predicted in minutes.

The four-mile run

In this illustrative case, a supplied model for comparable shuttle trips of two to six miles is predicted travel time = 4 + 2d minutes, where d is distance in miles. One four-mile trip took 15 minutes.

1. For four miles, predict 4 + 2(4) = 12 minutes.

2. The slope is 2 minutes per mile. Within the observed range, each additional mile adds two minutes to the predicted time.

3. The residual for the recorded run is 15 − 12 = +3 minutes. Its plotted point is three minutes above the prediction at four miles.

The line underestimated this four-mile shuttle run by three minutes. The slope describes predicted travel time across distances, and a single run can depart from that prediction.

KEY IDEA Subtract the prediction at the observation's own input from its observed response. Use response units for the residual and response units per input unit for slope.

WATCH FOR A negative residual can describe a perfectly ordinary positive travel time. It indicates a time shorter than the model predicted.

Sources and adaptation. Line, slope, and intercept explanations adapted from OpenStax, Introductory Statistics 2e, §12.1, printed pp. 617–620; residual interpretation from §12.3, pp. 623–626. The shuttle, bead, printer, and bicycle models are SMS practice cases. Textbook example adapted from OpenStax, §12.1, Example 12.3, p. 619. Tax-preparation fee and exact rates retained; comparison with a fitted relationship and extra-hour calculation are original.

UNIT 2 / DAY 09 • PREDICTION AND ROBUSTNESS / T5

Does traffic change the recommendation?

Compare two plausible model specifications and identify a decision that depends on which one is used.

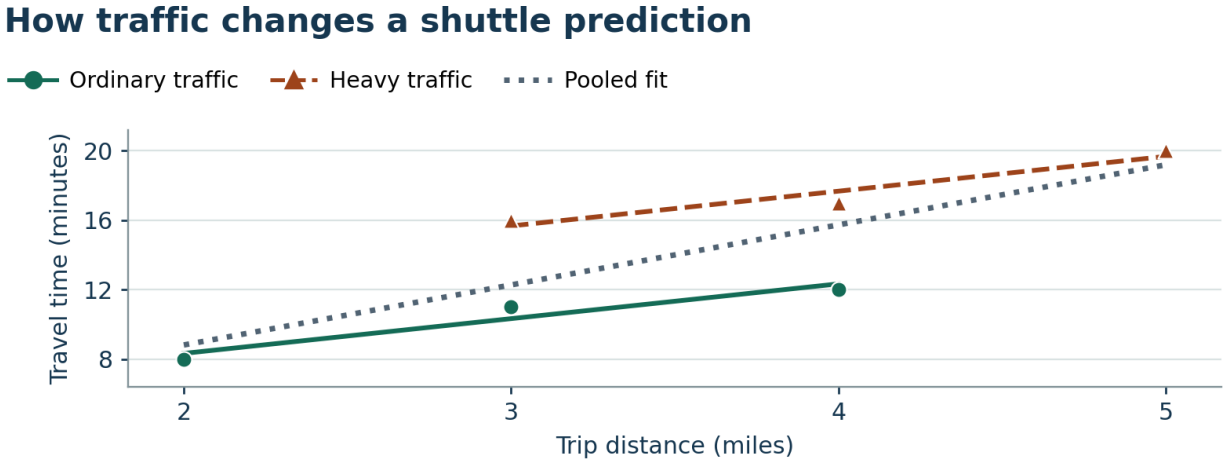
A model specification states what information a prediction uses and how it uses that information. Our pooled shuttle line combines all trips and uses distance alone. A traffic-specific model also uses whether traffic was ordinary or heavy. It can give two trips of the same distance different predictions.

The traffic model here uses two parallel lines: both have the same slope, but heavy traffic has the larger intercept. That common slope is an assumption. It says an additional mile has the same predicted time increase in both groups. A different model could allow different slopes, but we would need enough appropriate data to check that extra detail.

Greater detail can help when the extra information matters, but more complicated does not automatically mean better. A model can follow a small dataset closely and still predict new cases poorly. Pooling is simpler but can conceal a real group difference. Choose specifications for a reason connected to the process, then compare their predictions with new observations.

A robustness check asks whether a particular conclusion survives a reasonable change in the analysis. Hold the trip distance and time limit fixed, change the specification, and repeat the decision. If both specifications give the same decision, that decision is robust to this change. This does not prove it would survive every other change or that individual trips meet the limit.

Specify which conclusion you mean. All three shuttle lines predict more time for longer trips, so the direction agrees. Yet the decision about a 15-minute slot can change when traffic is included. Do not choose the model merely because it gives the answer the dispatcher prefers.



The pooled line uses all six trips. The traffic-specific fits share a slope and are drawn only over their respective observed distance ranges.

2

Does traffic change the recommendation?

Can a four-mile trip fit a 15-minute slot?

Return to the six supplied shuttle records. The ordinary-traffic pairs are (2, 8), (3, 11), (4, 12); heavy-traffic pairs are (3, 16), (4, 17), (5, 20). Rounded fitted lines give pooled time = 1.91 + 3.45d, ordinary time = 4.33 + 2d, and heavy time = 9.67 + 2d. Time is in minutes; d is miles. A trip is accepted on its prediction if that prediction is at most 15 minutes.

1. At d = 4, the pooled prediction is 1.91 + 3.45(4) = 15.71 minutes, just above the limit.

2. The ordinary-traffic prediction is 12.33 minutes and the heavy-traffic prediction is 17.67 minutes. The ordinary trip fits the slot on this model.

3. The observed ordinary four-mile trip took 12 minutes. Its traffic-specific residual is about −0.33 minute. A typical-time prediction still allows individual variation.

All three lines predict increasing travel time as distance increases. For an ordinary four-mile trip, however, the 15-minute decision changes: the pooled line predicts 15.71 minutes and declines it, while the ordinary-traffic line predicts 12.33 minutes and accepts it. That decision is sensitive to including traffic.

KEY IDEA	Test robustness by changing one plausible specification and repeating the very same calculation or decision.
WATCH FOR	A larger pooled slope does not establish a larger within-group change. Here heavy traffic is more common among the longer observed trips.

Sources and adaptation. Uses the prediction and residual foundation in OpenStax, Introductory Statistics 2e, §§12.3 and 12.5. The common-slope traffic model, model-comparison questions, and introductory robustness discussion are SMS extensions; the six shuttle records are constructed classroom data.

UNIT 2 / DAY 10 ● PREDICTION AND ROBUSTNESS / T5

Where does the model have support?

Check the observed range and conditions before defending a forecast.

Interpolation predicts at an input within the observed range. Extrapolation predicts beyond that range. Always use the range for the relevant group and process: a 2.5-mile heavy-traffic trip is outside a heavy-traffic log spanning three to five miles, even if ordinary trips include shorter distances.

The arithmetic can be correct while the prediction lacks support. OpenStax's exam-score example gives predicted final points = −173.51 + 4.83x, where x is a third-exam score. Observed x-values run from 65 to 75. Entering 80 gives 212.89 predicted final points, although the final has only 200 possible points. This is an extrapolation, and the impossible output reveals a limit of extending that line.

Adapted from OpenStax, §12.5, Example 12.11, pp. 635–636. Access for free at openstax.org.

Being inside the range is helpful but insufficient. A line based on ordinary paper may not describe thick card. A fair-weather travel model may not describe an ice storm. Check the measurement definitions, group, and operating conditions as well as the smallest and largest observed inputs. An unfamiliar condition can undermine a forecast even when its numerical input is in range.

A local application is the USGS streamgage on the Rogue River at Grants Pass. River height, called stage, and water flow, called discharge, are different quantities. Field measurements support a rating curve that converts stage into estimated discharge. USGS checks these curves because changing channels can change the relationship. A useful model needs continuing evidence, not just an equation.

Consider a hypothetical classroom curve used above every river height in its supporting measurements. Its flow estimate is an extrapolation. Explain that limitation and seek measurements or other defensible evidence near the new conditions. This is a classroom scenario, not a report of an actual Grants Pass flood or a claim about the current USGS curve.

UNIT 2 / DAY 10 • PREDICTION AND ROBUSTNESS / T5 • CONTINUED

2

Where does the model have support?

Does the requested distance lie inside the recorded range?

Requested trip	Matching range	Distance check
3.5 miles, ordinary	2–4 miles	Inside
2.5 miles, heavy	3–5 miles	Below group range
9 miles, heavy	3–5 miles	Above group range

The complete log spans 2–5 miles. For the 2.5-mile heavy trip, the pooled range and the heavy-traffic range give different answers.

A nine-mile request

The six-trip shuttle log covers 2–4 miles in ordinary traffic and 3–5 miles in heavy traffic. The pooled line is predicted minutes = 1.91 + 3.45d. The heavy line is 9.67 + 2d.

1. Nine miles exceeds the largest observed distance of five miles, so all fits would extrapolate.

2. The pooled line gives 1.91 + 3.45(9) = 32.96 minutes. The heavy line gives 9.67 + 2(9) = 27.67 minutes.

3. The two forecasts differ by about 5.3 minutes. Neither has been compared with observed nine-mile times.

Before setting a nine-mile heavy-traffic schedule, record several comparable eight- to ten-mile trips under heavy traffic. The current figures illustrate model assumptions and leave the reliability of that schedule unresolved.

KEY IDEA Check the range and conditions for the specific group the forecast describes. Identify which new observations could test the extension.

WATCH FOR Interpolation does not guarantee accuracy. Check the relevant group and conditions, and remember that a typical prediction is not a promise for every case.

Sources and adaptation. Interpolation and extrapolation adapted from OpenStax, Introductory Statistics 2e, §12.5, Example 12.11, printed pp. 635–636. The exam illustration changes the extrapolated third-exam score from 90 to a possible score of 80. Rogue River station and rating-curve facts: USGS station 14361500 and USGS "How is a rating curve used to convert gage height into streamflow?" The high-water scenario and shuttle exercises are SMS classroom cases.

UNIT 2 / DAY 11 • PREDICTION AND ROBUSTNESS / T4–T5

Writing a model recommendation

Defend an in-range model decision with a prediction, a checked error, and a specific limit.

Use the worked example for feedback and revision after the individual checkpoint or memo.

A useful recommendation answers a particular decision for a reader who has not seen the question. Identify the job or group, the quantity being predicted, the model, and the decision threshold. Use the same units for the prediction and threshold. "It works" leaves the reader unable to check what was decided or why.

First show the calculation that matters. Then interpret it: "For a routine six-book pickup, this model predicts 6.8 minutes, which is 0.2 minute below the seven-minute slot." The comparison supports accepting the job under a rule based on predicted time. It does not establish a maximum time for every pickup.

Next use actual observations to check the claim. If one six-book pickup took eight minutes, the residual is +1.2 minutes and that pickup missed the seven-minute slot. One missed deadline disproves a claim that every recorded job met it. One successful job, however, cannot establish that all future jobs will meet it.

Finally, state a specific limit and a useful next step. Is the input inside the appropriate group's range? Are the conditions comparable? Does a reasonable alternative model change the decision? Repeated new observations can show how often a proposed deadline is missed; a single residual cannot establish that rate.

Keep the amount of writing appropriate to the task. A calculation needs clear numbers and units. A recommendation needs enough context, evidence, and limitation to stand on its own. You can be concise and still state exactly what the evidence supports.

Two different checks on the six-book recommendation

Question	Evidence	Meaning
Is the input supported?	Six books within 2–8	In-range for routine orders
Does every job meet seven minutes?	One six-book order took eight	At least one observed miss

A range check and an observed miss answer different questions, even when the predicted time is below the deadline.

UNIT 2 / DAY 11 ● PREDICTION AND ROBUSTNESS / T4–T5 ● CONTINUED

2

Writing a model recommendation

Review after the individual memo

Illustrative library pickup orders contained 2–8 books. For routine orders the supplied line predicts minutes = 2 + 0.8b, where b is books. One six-book order took 8 minutes. Staff consider a seven-minute slot.

1. For six books the prediction is 2 + 0.8(6) = 6.8 minutes. The count of six falls inside the recorded range.

2. The observed residual is 8 − 6.8 = +1.2 minutes. This pickup took 1.2 minutes longer than the line predicted.

3. A seven-minute slot exceeds the typical prediction by only 0.2 minute. The observed six-book order exceeded the slot.

For routine six-book library pickups, the model predicts 6.8 minutes, below the seven-minute slot. However, an observed six-book pickup took eight minutes, so the evidence does not support a seven-minute maximum for every pickup. Record repeated comparable pickups and check the share that miss the deadline before making a stronger scheduling promise.

KEY IDEA	A defensible memo links a specific recommendation to a quantity, a decision rule, and the cases the evidence can describe.
WATCH FOR	An input inside the observed range removes one concern. It does not by itself establish that the model is suitable or that an individual job will meet a deadline.

Sources and adaptation. Prediction and residual interpretation build on OpenStax, Introductory Statistics 2e, §§12.3 and 12.5. The library example, decision-memo structure, and follow-up exercises are SMS instructional material.

UNIT 3 / DAY 12 • ESTIMATION AND CONFIDENCE / T6

One sample, one estimate

Turn individual yes/no outcomes into a sample proportion and explain what one dot in a plot of samples represents.

A parameter describes a population or a probability model. A statistic is calculated from the observations in a sample. Let p be the population proportion with a specified characteristic. If x of n sampled cases have that characteristic, the sample proportion is $\hat{p} = x/n$. For example, seven dark beads in ten draws give $\hat{p} = 7/10 = 0.70$, or 70%. Define what counts as a success before collecting data; "success" is simply the category being counted.

The bag below has a known proportion, $p = 12/20 = 0.60$. Drawing a bead at random, replacing it, and thoroughly remixing keeps the chance of a dark bead at 0.60 on every draw. Replacement alone is not enough: choosing by sight or consistently reaching into an unmixed layer can spoil that model. In a survey, random selection serves a similar purpose by connecting the observed sample to a specified population.

In an OpenStax survey example, 421 of 500 randomly selected city adults report owning a smartphone. The statistic is 421/500 = 84.2%; the target is the percentage of all adults in that city who own one. Drawing another 500 adults from the same population can change the estimate even while the population percentage stays fixed. This is sampling variability.

Adapted from OpenStax, §8.3, Example 8.10, pp. 421–422. Access for free at openstax.org.

Make one dot for each complete sample of the same size. Its horizontal position is that sample's proportion. The sampling distribution describes the proportions from all possible random samples of that size and how often each proportion occurs. A plot of many independent samples approximates that distribution. Pooling all successes and draws instead produces a single proportion; the individual samples' spread disappears.

UNIT 3 / DAY 12 • ESTIMATION AND CONFIDENCE / T6 • CONTINUED

One sample, one estimate

Six example teams sample the same bag

3



Example data: each dot is one team's ten draws. Two dots at 0.50 mean two teams each drew five dark beads.

Three teams, one bag

A bag contains 12 dark and 8 light beads. Three teams each make ten blind draws, replacing and remixing after every draw. The dark counts are 7, 4, and 7.

1. The fixed bag proportion is 12/20 = 0.60.

2. The sample proportions are 0.70, 0.40, and 0.70. Stack two dots at 0.70 and one at 0.40.

3. Across all 30 draws the dark share is 18/30 = 0.60. That pooled result happens to match the bag.

Each team estimated the same 60% dark share. A different set of 30 draws could have a different pooled result.

KEY IDEA One point in a plot of sample proportions stands for one entire sample.

WATCH FOR Write the denominator beside each proportion. Ten draws per team and three team estimates answer different questions.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §1.1, Definitions of Statistics, Probability, and Key Terms; §1.2, Data, Sampling, and Variation in Data and Sampling; §8.3, A Population Proportion, pp. 420–421. Definitions and repeated-sampling reasoning adapted; local bag, sprouting, and sign cases and numerical data are original classroom examples. Textbook example adapted from OpenStax, §8.3, Example 8.10, pp. 421–422. Study design and 421/500 retained; task addresses sampling variation and deliberately omits the source's interval calculation.

UNIT 3 / DAY 13 • ESTIMATION AND CONFIDENCE / T6

Why sample size changes spread

Compare variability at two sample sizes and explain why increasing the sample size cannot correct selection bias.

A sample proportion averages observations coded 1 for success and 0 for failure. One extra success changes a proportion based on ten trials by 1/10, or ten percentage points. It changes a proportion based on 100 trials by only one percentage point. This helps explain why larger independent random samples tend to give more stable estimates.

For repeated independent trials with the same success probability p, the sample proportions are centered at p. Their standard deviation is $\sqrt{[p(1 - p)/n]}$, where n is the size of each sample. This standard deviation describes variation between sample proportions, not variation between individual outcomes. You may use supplied values in this course; the important relationship is that spread is proportional to $1/\sqrt{n}$. Quadrupling n halves this spread.

A larger sample is not guaranteed to be closer to the population value on every attempt. The claim concerns the pattern over many random samples. One small sample can land exactly on the target while one larger sample misses it. Use several samples to compare distributions, rather than declaring a winner from a single pair.

Increasing sample size reduces random sampling variation, but it does not remove selection bias. A thousand volunteers reached through one club may describe the school less fairly than a smaller random sample from the full student roster. Repeated measurements from the same person may also be dependent; calling them separate trials does not automatically create independent evidence. Check how observations were selected and how they relate before treating sample size as precision.

Theoretical spread in the mark model

Items in one sample	Center of estimates	SD of estimates
25	40%	9.8 percentage points
100	40%	4.9 percentage points

Model summaries for repeated independent samples. The SD describes sample proportions, not individual items.

UNIT 3 / DAY 13 • ESTIMATION AND CONFIDENCE / T6 • CONTINUED

Why sample size changes spread

Two sizes in a stable model

3

In a simulated process each item independently has a 40% chance of showing a mark. The underlying model spread of sample proportions is about 0.098 for n = 25 and 0.049 for n = 100.

1. Both sampling distributions are centered at 0.40.

2. The supplied SDs follow $\sqrt{[p(1-p)/n]}$: $\sqrt{(0.40 \times 0.60 / 25)} \approx 0.098$ and $\sqrt{(0.40 \times 0.60 / 100)} \approx 0.049$. Multiply by 100 to express these as 9.8 and 4.9 percentage points.

3. The spread fell from about 9.8 to 4.9 percentage points when sample size quadrupled. The units describe how far sample estimates typically vary.

4. A single n = 100 estimate could still lie farther from 0.40 than one particular n = 25 estimate.

Repeated 100-item samples estimate the mark rate more consistently under the same selection process.

KEY IDEA	Increasing the size of a random sample generally reduces sampling variation. Check who could enter the sample before using its estimate for the population named in the question.
WATCH FOR	A claim about the spread across samples gives no guarantee about which of two single estimates will be closer.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §8.3, A Population Proportion, pp. 420–421: sampling proportion and its standard deviation; §1.2: sampling variability and bias. Exposition adapted; mark and bead data are original classroom examples. The square-root formula is shown to explain the sample-size relationship, not as a required derivation.

UNIT 3 / DAY 14 • ESTIMATION AND CONFIDENCE / T6

What 95% confidence means

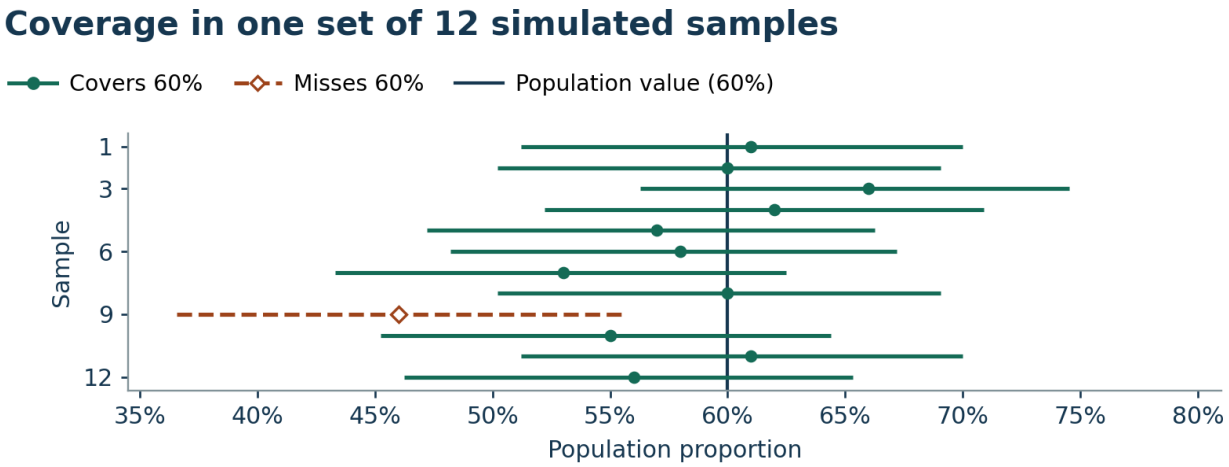
Read a repeated-interval display and explain coverage as a property of the sampling and interval procedure.

A point estimate gives one sample-based value for an unknown population quantity. A confidence interval adds a range of estimates to express sampling uncertainty. For a proportion, the lower and upper endpoints refer to the population proportion with the stated characteristic. They are not a range containing individual yes/no responses or a claim that most people lie between two percentages.

The confidence level describes the interval-producing method. Imagine repeatedly taking fresh random samples in the same way and constructing an interval from every sample. A 95% confidence method aims to capture the fixed population value in about 95% of those intervals over the long run, when its assumptions hold. The sample and interval change from repetition to repetition; the population value does not.

In the simulation below, we know that p = 0.60, so we can check every interval. An interval captures 0.60 when its lower endpoint is at or below 0.60 and its upper endpoint is at or above 0.60. Include endpoints. Count captured intervals and divide by the total number of intervals. Twelve intervals are a small batch, so their capture percentage need not be close to 95%.

In an actual study, p is usually unknown. After an interval is calculated, it either contains p or it does not; we cannot discover which merely by quoting 95%. Our confidence comes from the method's repeated-sampling performance. The supplied intervals here use the Wilson score method. Its coverage is approximate and can vary with the sample size and true proportion. You will interpret the output, not derive the formula.



One segment represents one sample of 100 and one 95% Wilson score interval. The vertical line marks the fixed population rate of 60%.

UNIT 3 / DAY 14 • ESTIMATION AND CONFIDENCE / T6 • CONTINUED

What 95% confidence means

3

When an interval misses

A sensor simulation uses a fixed 60% probability of a positive reading. One simulated sample of 100 has 46 positive readings, with a supplied 95% interval of 36.6%–55.7%. Another sample of 100 has 60 positives and an interval of 50.2%–69.1%.

1. For the first sample, 60% lies above the upper endpoint of 55.7%, so it misses.

2. For the second sample, 50.2% ≤ 60% ≤ 69.1%, so it captures the rate.

3. The two intervals used the same procedure and sample size. Different outcomes changed the estimates and endpoints.

A miss is possible in a sound simulation. In a real study the true population rate is usually unavailable for this kind of check.

KEY IDEA	The parameter stays fixed while new samples generate new intervals. Long-run coverage belongs to the procedure.
WATCH FOR	A 95% confidence interval for a population proportion does not enclose 95% of individual yes/no observations.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §8.1, confidence-level explanation, p. 407, and §8.3, Examples 8.10–8.11, pp. 421–423. Confidence interpretation adapted; coverage counts and exercise data are original classroom examples. Wilson score endpoints follow NIST/SEMATECH e-Handbook §7.2.4.1, https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm. Supplied Wilson intervals differ from OpenStax's unadjusted normal and plus-four intervals; no derivation is required.

UNIT 3 / DAY 15 • ESTIMATION AND CONFIDENCE / T6

Using an interval to make a decision

Interpret both confidence-interval endpoints for a population and apply a decision rule chosen in advance.

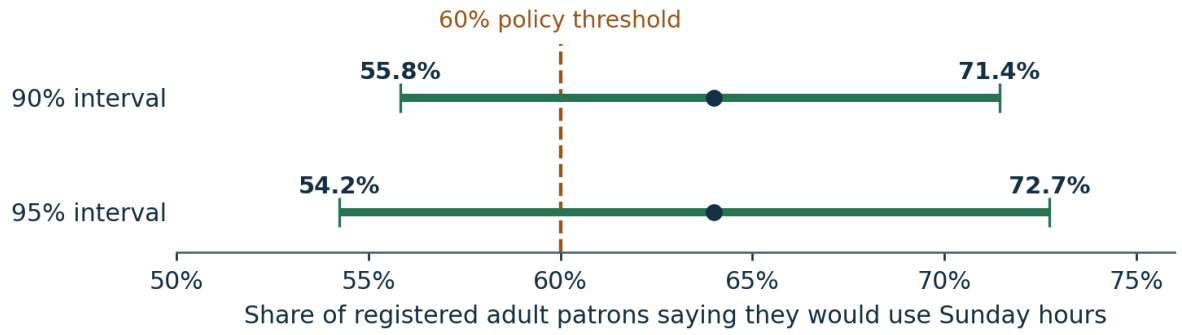
To interpret an interval, name the population, the characteristic being counted, the confidence level, and both endpoints. For example, "We are 95% confident that between 54.2% and 72.7% of registered adult library patrons would say they would use Sunday library hours." This sentence makes the target and uncertainty clear even when the reader has not seen the question.

An interval's credibility depends on the data collection. A random sample from the relevant list supports a claim about that list's population if selected people respond accurately. Independent trials are another common model. Sampling without replacement from a much larger population can be treated as approximately independent; in this course, use a sample no larger than 10% of the population for that approximation. A volunteer poll or a heavily incomplete response does not acquire representative sampling merely because a calculator supplies an interval.

For a decision, compare the whole interval with a benchmark chosen in advance. If a policy requires the lower endpoint to exceed 60%, a sample proportion above 60% is insufficient when the interval still includes values below 60%. That finding does not prove the policy should never be adopted. It says this evidence does not meet the stated rule.

Using the same data and interval method, higher confidence requires a wider interval. Width equals the upper endpoint minus the lower endpoint, in percentage points when endpoints are percentages. Do not change the confidence level after seeing which interval produces a preferred answer. Also distinguish a measured response from the action it might predict: stated willingness to attend is not guaranteed future attendance.

Does the whole interval clear 60%?



Both Wilson intervals use the same 64 yes responses in 100. From the displayed endpoints, the 90% interval spans 15.6 percentage points and the 95% interval spans 18.5.

UNIT 3 / DAY 15 • ESTIMATION AND CONFIDENCE / T6 • CONTINUED

Using an interval to make a decision

3

Should the library add Sunday hours?

Suppose a simple random sample of 100 adult library patrons is selected from a large register. All respond, and 64 say they would use Sunday hours. The supplied 95% interval is 54.2%–72.7%. A trial requires the entire interval to exceed 60%.

1. The sample estimate for stated willingness is 64/100 = 64%.

2. The interval estimates the share of registered adult patrons who would say yes. The 60% threshold lies between 54.2% and 72.7%.

3. The lower endpoint does not exceed the rule's threshold.

We are 95% confident that between 54.2% and 72.7% of registered adult library patrons would say they would use Sunday hours. The survey does not meet the council's trial rule because the lower endpoint is below 60%. Actual Sunday attendance is a different outcome.

KEY IDEA Compare the endpoint specified in the decision rule with its threshold, and state what the study measured.

WATCH FOR An interval crossing a threshold leaves that threshold plausible. This alone cannot establish that the true value equals the threshold.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §8.1, Example 8.4 and confidence-level comparison, pp. 413–414; §8.3, Examples 8.10–8.11, pp. 421–423. Interpretation and confidence-width principles adapted; local surveys and decision rules are original classroom examples. Wilson outputs use the method cited for Day 14.

UNIT 3 / DAY 16 • ESTIMATION AND CONFIDENCE / T6 AND T1

How far can an estimate reach?

Use the sampling design and confidence interval together to defend a population claim, then revise a specific error.

Use the worked example for feedback and revision after the individual checkpoint or memo.

A sampling frame is the list or procedure used to select cases. Start an inference by asking who could have been selected. A random sample from a permit register can support a claim about people on that register. Residents without permits have no chance of selection, so that design does not directly estimate the views of all town residents.

Keep three descriptions separate: the intended population, the population reached by the frame, and the people who actually supplied data. Missing groups create undercoverage. If selected people do not respond and their views differ from respondents' views, nonresponse can also distort the estimate. Random selection from a frame does not repair a frame that leaves out relevant people.

A confidence interval describes sampling uncertainty under a model. Its width does not include every possible problem in the study. Misleading wording, inaccurate answers, repeated dependent observations, and selection bias can affect an estimate even when its calculated interval is narrow. A larger sample alone does not resolve these problems.

Write conclusions at the scale the evidence supports. Identify the group, time or setting, and measured characteristic. A weekday trail-user sample need not describe weekend users. A survey of willingness does not establish actual attendance. If a claim goes beyond the frame, either narrow the claim or explain what additional data are needed. This is a reasoned limitation, not a reason to discard all information from the sample.

Two registers answer two questions

Sampling frame	Random sample	Defensible target
Permit holders	50 permit holders	People listed on the permit register
All town residents	A random sample from a complete resident register	Residents on that register

The two registers cover different groups. If the question concerns all residents, a permit-holder sample alone leaves residents without permits outside the study.

UNIT 3 / DAY 16 • ESTIMATION AND CONFIDENCE / T6 AND T1 • CONTINUED

How far can an estimate reach?

3

Who had a chance to answer?

A parking office draws 50 names at random from a large complete register of permit holders. All respond. Of the 50, 32 favor adding a covered bicycle rack; a supplied 95% interval is 50.1%–75.9%. The headline says, 'The town wants a new rack.'

1. The sampling frame is registered permit holders. Other town residents were never sampled.

2. The outcome is stated support for this proposal. The sample share is 32/50 = 64%.

3. The 95% interval starts at 50.1%, just above one-half. That supports a narrow majority claim for permit holders on the register. Changing even one "yes" to "no" would lower the sample share from 64% to 62%; the evidence is close to the cutoff.

Among registered permit holders, 32 of 50 sampled favored the rack. We are 95% confident that 50.1% to 75.9% of permit holders favor it. The evidence does not extend to the entire town.

KEY IDEA Name the population that the selection procedure actually represents. Keep that population in the estimate and in the final claim.

WATCH FOR Random selection from a narrow list supports a claim about that list. A larger sample from the same list still leaves excluded people unmeasured.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §1.2, Data, Sampling, and Variation in Data and Sampling, and §8.3, A Population Proportion, pp. 420–423. Sampling-scope and inference principles adapted; local registers and survey results are original classroom examples. Wilson outputs use the method cited for Day 14.

UNIT 4 / DAY 17 • SIGNIFICANCE AND DECISIONS / T7

Simulate a benchmark

I can estimate an upper-tail p-value and explain what one simulated batch represents.

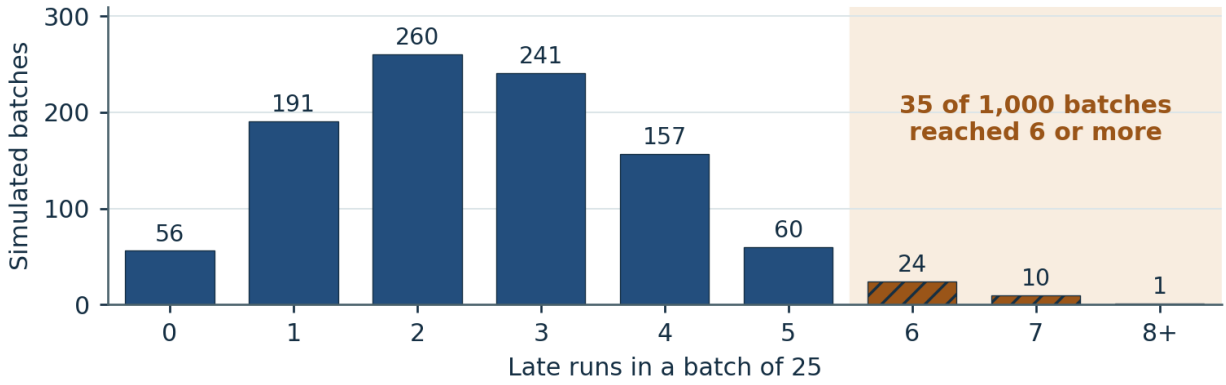
A significance test asks how unusual the observed result would be under a specified benchmark model. Let p_0 denote the benchmark probability of the outcome being counted. This is the null value. The alternative states the direction we are investigating, such as a late-run rate greater than 10%. Choose that direction before examining results; an increase requires an upper-tail comparison.

The model includes more than a percentage. In the examples here, each batch has a fixed number n of trials, each trial has two possible outcomes, trials are independent, and every trial has the same success probability p_0. A rush-hour cluster or a changing process may violate those assumptions. "Random" should describe an actual selection or generation procedure, not merely an unpredictable result.

To simulate a benchmark, generate one complete batch of n trials at p_0 and count its successes. Repeat for many independent batches. For a test of an increase, count simulated batches with a success count equal to or greater than the observed count. Divide that count by all simulated batches. This estimates the p-value: the probability, under the null model, of a result at least as far in the specified direction as the observed result.

Include ties, and use batches rather than individual trials as the denominator. If 35 of 1,000 batches reach the observed count or higher, the estimated p-value is 0.035. A new simulation will usually give a somewhat different estimate. More simulated batches improve the estimate of this probability; they do not add observations to the original study or make its sampling design better.

Which simulated shuttle batches enter the upper tail?



Each batch has 25 independent trials with late probability 0.10. Of 1,000 simulated batches, 905 had 0–4 late runs, 60 had five, and 35 had six or more.

UNIT 4 / DAY 17 ● SIGNIFICANCE AND DECISIONS / T7 ● CONTINUED

Simulate a benchmark

When is a parcel count unusual?

4

Suppose 20% of parcels normally arrive after closing. A depot checks 20 comparable, independently selected parcels and finds 7 late arrivals. The question, chosen in advance, is whether the late rate has increased.

1. Assign 20 independent trials to each simulated batch and give every parcel a 0.20 chance of being late. Record one late count per batch.

2. In an illustrative simulation of 1,000 batches, 52 had exactly 7 late arrivals, 14 had exactly 8, and 10 had 9 or more. The upper tail contains 52 + 14 + 10 = 76 batches.

3. Divide 76 by 1,000 to estimate the tail probability as 0.076. Dividing 7 by 20 gives the observed late fraction, 0.35.

At a modeled 20% late rate, 76 of 1,000 simulated groups of 20 parcels had seven or more late arrivals. The estimated upper-tail probability is 7.6%. Another simulation could give a different estimate.

KEY IDEA	A null simulation holds the proposed rate and sample size fixed. The late count can change from batch to batch.
WATCH FOR	The observed fraction describes the sampled parcels. The p-value describes repeated samples under the proposed model.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §4.3, Binomial Distribution; §9.1, Null and Alternative Hypotheses, pp. 462–463; §9.4, Rare Events, the Sample, Decision and Conclusion, pp. 467–469. Null-model and tail-probability reasoning adapted; cases and simulation counts are original classroom examples. Inference is taught using simulation and supplied exact binomial output, not the normal test in OpenStax §9.3.

UNIT 4 / DAY 18 • SIGNIFICANCE AND DECISIONS / T7

How strong and how large?

I can compare a p-value with a preselected cutoff and report the size and practical importance of the observed change.

The significance level, written α (alpha), is a cutoff chosen before looking at the data. If the p-value is at most α, call the result statistically significant at that level. A larger p-value falls short of the cutoff. Notice the distinction: p is a population proportion, while a p-value is a probability calculated assuming the benchmark model. A test decision does not prove either hypothesis.

Always report the observed size of the difference separately. A sample rate of 24% compared with a 10% benchmark has an observed excess of 14 percentage points. The p-value answers a different question: how unusual is a result at least this far above the benchmark when the benchmark model holds? It is not the probability that the observed increase happened "by chance," and 1 minus the p-value is not the probability that an increase is real.

Two samples can have the same observed excess but different p-values. Larger independent samples generally have less variation in their sample proportions, so a fixed difference can be less compatible with the benchmark. Conversely, an imprecise small sample can show a substantial observed difference without meeting the cutoff. A result just above 0.05 and one just below it should not be described as entirely different kinds of evidence.

Practical importance asks what the size of a difference would mean for people, costs, safety, or service. Report the sample, observed magnitude, p-value and cutoff, then name the consequence relevant to the decision. Check the study design before generalizing. A threshold can help set a consistent action rule, but the threshold alone does not establish that a change is large, useful, or worth the cost of a response.

Same observed excess in two inspections

Defective items	Observed defect rate	Upper-tail p-value
4 of 16	25%	0.210
20 of 80	25%	0.013

Against a 15% defect benchmark, both samples show an excess of 10 percentage points. Only the 80-item result meets a 0.05 cutoff for an increase.

How strong and how large?

Should dispatch investigate lateness?

Six of 25 independently sampled comparable runs were more than five minutes late. The office specified an upper-tail test against a 10% benchmark and chose a 5% significance level.

4

1. Compute the observed rate, 6/25 = 24%. Its difference from the benchmark is 24% − 10% = 14 percentage points.

2. The benchmark model gives a probability of about 0.033 for six or more late runs. The Day 17 shuttle simulation estimated the same probability as 35/1,000 = 0.035. Compare 0.033 with the preselected 0.05 cutoff.

3. Report evidence of a higher rate among comparable runs. Ask whether riders missed transfers or lost substantial time before deciding how much the change matters.

Among the 25 sampled shuttle runs, 6 were more than five minutes late, an observed rate of 24%, or 14 percentage points above the 10% benchmark. Under the independent-trial model, the upper-tail p-value is about 0.033, which supports a higher late rate for comparable runs at the preselected 5% level. Dispatch should also consider missed transfers and delay costs when choosing a response.

KEY IDEA	Report the observed change, the benchmark, the p-value and the cutoff. Then address the practical decision in context.
WATCH FOR	A p-value of 0.033 gives no 96.7% probability that the shuttle rate increased. Its calculation assumes the benchmark.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §9.4, Decision and Conclusion, pp. 468–469. Benchmark decisions and contextual conclusions adapted; local cases are original. This course consistently uses p-value ≤ α for the cutoff, including equality. American Statistical Association, Statement on Statistical Significance and P-Values (2016), principles 2, 3, and 5; https://doi.org/10.1080/00031305.2016.1154108. The contextual distinction between evidence, effect size, and practical decisions follows these principles.

UNIT 4 / DAY 19 ● SIGNIFICANCE AND DECISIONS / T8

Two ways a decision can go wrong

I can identify a false alarm and a missed change and explain a concrete consequence of each.

A statistical decision has two separate parts: the unknown truth about the population and the action chosen from the sample. For the shuttle, the benchmark is a true late rate of 10%; the alternative is a higher rate. The office either flags a possible increase for review or does not flag it. A sample determines the action, while the true rate determines whether that action is correct.

A Type I error is a false alarm: flagging an increase when the benchmark is true. A Type II error is a missed increase: not flagging when a specified higher rate is true. The other two truth-and-action combinations are correct decisions. Place the true conditions in rows and the possible actions in columns to keep these four combinations distinct.

In an OpenStax example, the starting claim is that a climber's equipment is safe. A false alarm would sideline sound gear. A miss would leave faulty gear in use, exposing the climber to injury. In the shuttle case, an unnecessary review costs staff time, while a missed increase in late runs could cost riders a connection. The consequences depend on the action and whom it affects.

Adapted from OpenStax, §9.2, Example 9.5, p. 464. Access for free at openstax.org.

We usually cannot tell from one sample whether an error occurred, because the true population rate is unknown. A flag can be either a correct detection or a false alarm. No flag can mean a correct decision or a missed change. "Do not reject the benchmark" therefore means insufficient evidence for the specified alternative, not proof that the population rate equals the benchmark. Classify an error only when the question supplies both the truth and the action.

Actual route condition and dispatcher action

True late rate	Flag for review	Leave schedule alone
10% benchmark	False alarm (Type I)	Correct decision
25% alternative	Correct detection	Missed change (Type II)

Each row assumes a different true rate. A sample selects an action; the process condition determines whether that action is an error.

Two ways a decision can go wrong

Whose time does the error cost?

Dispatch flags a route for review when at least six of 25 trips are late. Compare a route truly late 10% of the time with a route truly late 25% of the time.

1. At the 10% benchmark, flagging the route is a false alarm. Staff spend time investigating a change that did not occur.

2. At a true late rate of 25%, leaving the schedule alone misses deterioration. Riders may continue to miss transfers.

3. If six of 25 trips are late, the action is to flag. The count alone does not reveal whether that action was a false alarm.

A false alarm burdens staff with an unnecessary shuttle review. A missed increase can keep riders on a schedule that no longer meets its reliability target.

KEY IDEA	Name both the underlying truth and the action. A consequence should identify who bears a cost and what that cost is.
WATCH FOR	A decision to leave the schedule alone does not establish that the route's rate is exactly 10%.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §9.2, Outcomes and the Type I and Type II Errors, Table 9.2 and Example 9.5, pp. 464–465. Truth-by-decision framework and comparison of consequences adapted; shuttle, purifier, bulb, and cafeteria cases are original classroom examples. Textbook example adapted from OpenStax, §9.2, Example 9.5, p. 464. Climbing-equipment scenario retained; operational consequences restated and compared with the classroom shuttle.

UNIT 4 / DAY 20 ● SIGNIFICANCE AND DECISIONS / T8

Power at a stated alternative

I can read a decision rule under two true rates and explain the tradeoff between false alarms and power.

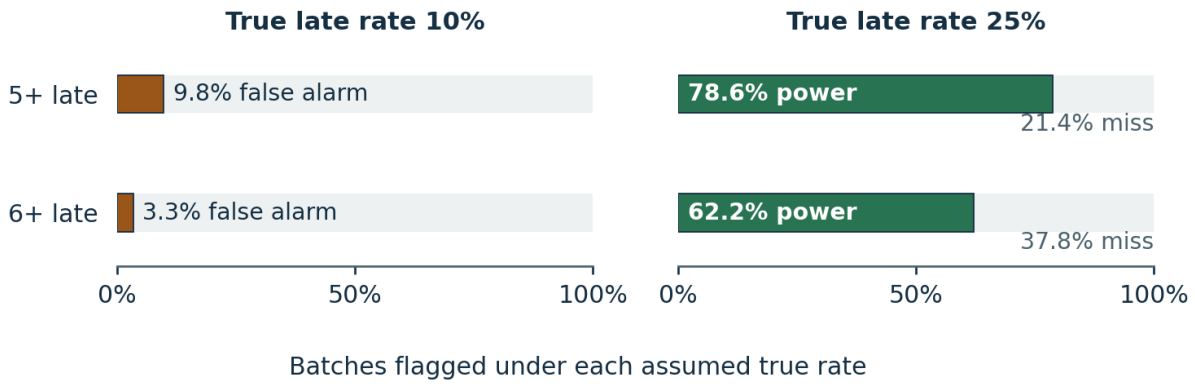
Power is the probability that a decision rule detects a change when a particular alternative is true. Always name that alternative. A rule can have different power at a true late rate of 15% and at a true late rate of 25%. At the same alternative, power plus the probability of a Type II error equals 1: every batch either triggers the rule or misses the increase.

Read false-alarm rates and power from different populations of repeated batches. To measure false alarms, generate batches at the benchmark. To measure power at 25%, generate a separate set at a true rate of 25%. For example, 62.2% power means about 622 detections per 1,000 batches generated at that alternative. It does not mean that 62.2% of flagged batches came from a changed process.

To compare cutoffs, hold sample size and both assumed true rates fixed. Lowering the required number of late runs from six to five flags every previously flagged outcome plus the outcome of exactly five. This raises the chance of flagging under the benchmark and under the alternative. Thus it raises both false alarms and power, while reducing missed changes at the specified alternative.

A preselected 5% significance level is a limit, not a promise that exactly 5% of benchmark batches will be flagged. With whole-number counts, available rules jump in steps: the six-of-25 shuttle rule has about 3.3% false alarms; the five-of-25 rule has about 9.8%. A larger well-designed sample can often improve power while keeping false alarms controlled. Collecting more data requires time or money, and a changed sample size requires a recalibrated rule.

How often will each shuttle rule flag?



For each cutoff, the left panel shows a false alarm at the 10% benchmark. The right panel shows power and the miss probability at the 25% alternative.

Power at a stated alternative

Should the library flag five late returns or six?

A library checks batches of 20 independent returns. The benchmark late rate is 15%. The library wants to detect an increase to 35% and can investigate at most 8 false alarms per 100 benchmark batches on average.

1. If five or more late returns trigger a flag, the model gives 17.0% false alarms at the benchmark and 88.2% power at 35%. This rule exceeds the 8% false-alarm limit.

2. Requiring six or more gives 6.7% false alarms at the benchmark and 75.5% power at 35%. Only this rule meets the limit.

3. At the 35% alternative, its miss rate is 100% − 75.5% = 24.5%. Among 1,000 batches from the changed process, expect about 245 misses. To count false alarms, use batches from the 15% process instead.

The library should flag six or more late returns under its stated rule. Compared with flagging at five, this saves investigations when returns are on schedule but detects fewer increases when the true late rate is 35%.

KEY IDEA	Power and miss probability add to one at the same specified alternative. The false-alarm rate uses batches generated at the benchmark.
WATCH FOR	Power is a detection probability given a true change. It does not give the probability of a change given that an alarm sounded.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §9.2, p. 464, and Chapter 9 Review, p. 487: Type II error, power, and sample size. Definitions adapted; library, shuttle, orchard, and dispatch decision rules are original. Supplied probabilities are exact binomial upper tails, rounded to one decimal percent.

UNIT 4 / DAY 21 ● SIGNIFICANCE AND DECISIONS / T7 AND T8

Defend a decision rule

I can recommend a rule using its stated risk limit and explain what the actual sample contributes to the decision.

Use the worked example for feedback and revision after the individual checkpoint or memo.

A defensible decision rule is designed before its result is known. Specify the population or process, benchmark, sample size, direction of concern, and a change large enough to matter. State the maximum acceptable false-alarm rate and what a flag will trigger. These choices connect the statistical model to the actual decision.

First eliminate rules whose false-alarm rates exceed the stated cap. Among the remaining rules, compare power at the same alternative. If detecting that change is the priority and other costs are equal, prefer the eligible rule with greater power. If neither rule meets the cap, redesign the study or reconsider the policy openly; do not quietly change the cap after observing a result.

Next apply the chosen rule to the observed sample and report the size of the observed difference. A p-value describes the observed result under the benchmark. Power describes a rule's detection probability under a specified alternative, before any particular sample is seen. They answer different questions and should not be substituted for one another.

Finish with the action, a consequence of a possible error, and a relevant limitation. A flag may justify inspection without establishing the cause of a problem. A no-flag result may leave a costly change undetected. Neither a small p-value nor high power repairs biased selection or dependent trials. A reader should be able to identify the population, evidence, rule and recommendation from your memo without having the original question beside it.

Warehouse review options

Flag in 25 crates	False alarm at 20%	Power at 40%	Fits a 3% cap?
9 or more	4.7%	72.6%	No
10 or more	1.7%	57.5%	Yes

The client's cap applies under the benchmark. The power column applies when deterioration to 40% is real.

UNIT 4 / DAY 21 ● SIGNIFICANCE AND DECISIONS / T7 AND T8 ● CONTINUED

Defend a decision rule

A supplier review

4

A warehouse reviews whether the re-inspection rate exceeds 20%. It samples 25 independent crates and must keep false alarms under 3%. A 9-crate cutoff gives a 4.7% false-alarm rate and 72.6% power at a true 40% rate. A 10-crate cutoff gives 1.7% false alarms and 57.5% power at 40%.

1. Only the 10-crate cutoff meets the client's 3% false-alarm limit. Under the 40% alternative its miss probability is 100% − 57.5% = 42.5%.

2. The warehouse then finds 10 of 25 crates need re-inspection. The observed rate is 40%, or 20 percentage points above the benchmark. The upper-tail p-value for ten or more under the benchmark is about 0.017.

3. The sample triggers a review under the chosen rule. The 57.5% power is a repeated-sample detection rate at a specified alternative, rather than a probability assigned to this particular supplier.

Review the supplier under the 10-crate rule. This rule fits the warehouse's false-alarm cap, though it would miss about 42.5% of batches at a true 40% re-inspection rate.

KEY IDEA	A defensible decision names the risk limit, the specific alternative, the observed evidence, and a consequence for someone affected.
WATCH FOR	Choose a rule using the stated policy before you see the sample. Picking a cutoff after seeing the count can change its false-alarm behavior.

Sources and adaptation. OpenStax, Introductory Statistics 2e, §§9.2 and 9.4, pp. 464–469: errors, power, and contextual decisions. Original synthesis of these ideas using hypothetical warehouse, device, and reservation cases; supplied probabilities are exact binomial upper tails.

UNIT 5 / DAY 22 ● SYNTHESIS AND REVISION / T1–T8

Repair a statistical explanation

Find the first unsupported step in a claim and show how the correction changes its meaning.

A row and column can share a numerator but give different percentages. In an OpenStax fictional driver study, 25 of the 70 drivers with a speeding violation used a phone while driving (35.7%). Among the 305 drivers who used a phone while driving, 25 had a violation (8.2%). The word "among" identifies the denominator. A cell counts both conditions; the row and column totals count different groups.

Adapted from OpenStax, §3.4, Example 3.20, Table 3.3, pp. 183–184. Access for free at openstax.org.

For an alert system, distinguish detection among cases that truly have the problem from the share of alerts that actually identify a problem. Detection uses all true problem cases in its denominator. The second rate uses all alerted cases. To find the false-alarm share among sound cases, divide false alerts by all sound cases. Switching the order of the descriptions usually changes the denominator and the question.

Use the same habit with a model or an interval. A residual is observed minus predicted: a trip lasting 12 minutes against a 10-minute prediction takes two minutes longer than predicted. For an interval, check which endpoint controls the claim. An interval from 6% to 17% includes 10%, but it also includes rates below 10%. It cannot support a claim that the rate is at least 10%.

Keep valid work when revising. Identify the earliest incorrect choice, explain why it is wrong, and correct the statement in context. Then use fresh data to check whether you can make that choice independently. A narrower, supported claim is more useful than a precise calculation attached to an unsupported conclusion.

Alerts on 80 observed trips

Actual trip	Flagged	Unflagged	Total
Late	10	10	20
On time	6	54	60
Total	16	64	80

Read across a row or down a column according to the question. The cells describe observed trips.

UNIT 5 / DAY 22 ● SYNTHESIS AND REVISION / T1–T8 ● CONTINUED

Repair a statistical explanation

Which trips supply the denominator?

In this supplied practice record, a dispatcher flagged 16 of 80 observed trips as possibly late. Ten flagged trips were actually late, and 20 trips were late in total. Compare "late among flagged" with "flagged among late."

5

1. For the share of flagged trips that were late, restrict attention to the 16 flagged trips. The fraction is 10/16 = 62.5%.

2. For the share of late trips that were flagged, start with the 20 late trips. The fraction is 10/20 = 50%. The same ten trips appear in both numerators.

3. Check the table: 6 flagged trips were on time, 10 late trips were unflagged, and 54 on-time trips were unflagged. The four cells sum to 80.

Of the 20 late trips observed, the dispatcher flagged half. Among flagged trips, 62.5% were late. The two rates describe different groups.

KEY IDEA Before dividing, finish the phrase 'among the …' and count that group.

WATCH FOR A large sample can estimate the wrong group quite precisely.

Sources and adaptation. Two-way-table and conditional-rate reasoning adapted from OpenStax, Introductory Statistics 2e, §3.4, especially Example 3.20 (pp. 183–184). Residual and interval review draws on §§12.3 and 8.3. Alert and scanner data are original classroom examples. Textbook example adapted from OpenStax, §3.4, Example 3.20, Table 3.3, pp. 183–184. Original fictional cell and margins retained; reversal added to compare the denominators.

UNIT 5 / DAY 23 ● SYNTHESIS AND REVISION / T1–T8

Write a decision memo from a new case

Combine study design, numerical evidence, and a stated decision rule in a short recommendation.

A decision memo should answer the client's actual question. Begin with the action under consideration, the target population, and the outcome measured. A survey about willingness to attend measures stated intention; it does not count attendance. A sample percentage can describe responses accurately even when it cannot justify a claim about everyone the client hopes to serve.

Under a suitable sampling design, the sample proportion estimates the population share, and a confidence interval expresses uncertainty about that share. Interpret its endpoints in context. The confidence level belongs to the interval-producing method: over repeated comparable random samples, about that percentage of intervals would contain the fixed population proportion. It does not mean that the population proportion changes from sample to sample or that every value in the interval is equally likely.

A decision rule states which evidence will justify an action. If the rule requires the entire interval to exceed a threshold, inspect the lower endpoint: it must be greater than the threshold. A point estimate above the threshold is not enough. If the interval crosses the threshold, the rule is not met; this does not prove that the population share is below the threshold. Different clients can reasonably choose different rules when costs or consequences differ.

Write the recommendation as a connected argument: report the relevant evidence, compare it with the rule, and state the resulting action and its scope. Add the limitation that matters for this decision, rather than a generic warning. A partner reading without the prompt should be able to identify the population, outcome, evidence, and reason for the recommendation.

Decision, evidence, and limit

What to name	Tool-library example
Who was studied	Registered members, sampled at random
What was measured	Stated intention to use evening hours
What controls the rule	The 48% lower endpoint falls below 50%
What remains unknown	How many people would actually attend

A numerical point can be accurate even if the proposed action exceeds the evidence.

UNIT 5 / DAY 23 ● SYNTHESIS AND REVISION / T1–T8 ● CONTINUED

Write a decision memo from a new case

Would a tool library extend its hours?

In a fictional tool-library study, all 100 randomly selected registered members respond; 58 say they would use an evening session. A supplied 95% interval for this stated-interest proportion is 48%–67%. The board starts a trial only if the entire interval exceeds 50%.

1. The observed share is 58/100 = 58%. The interval describes stated interest among registered members.

2. The lower endpoint, 48%, is below the board's threshold. An estimate greater than 50% alone does not meet the rule.

We are 95% confident that 48% to 67% of registered tool-library members would say they would use evening hours. The 48% lower endpoint falls below the board's 50% threshold, so the survey does not support a trial under its rule.

KEY IDEA Link the feature of the evidence that controls the decision to the client's stated rule.

WATCH FOR A survey about intended use measures a response; attendance is a different outcome.

Sources and adaptation. Population-proportion interval interpretation adapted from OpenStax, Introductory Statistics 2e, §8.3 (pp. 420–422); sampling limits from §1.2. Prediction and residual review draws on §12.3 (pp. 623–627). Tool-library, trail-program, pickup, and garden-club cases and decision rules are original practice examples.

UNIT 5 / DAY 24 ● SYNTHESIS AND REVISION / T1–T8

Independent performance task

Use supplied statistical evidence to defend a conclusion independently in an unfamiliar setting.

Review the types of evidence before the individual task. Follow its directions about available materials.

An assessment may supply several kinds of evidence. Begin by identifying what the question asks about: a distribution of measurements, a prediction, a population proportion, or a decision under uncertainty. Choose the relevant output before calculating. A fitted line predicts a typical response at a specified predictor value; a confidence interval estimates a population quantity. Neither guarantees the outcome for one individual.

Check what each number measures. A residual is observed minus predicted and has the response variable's units. A standardized score divides a signed difference from the group mean by that group's SD. A difference between percentages can be reported in percentage points. For a conditional rate, name the group in the denominator before dividing.

A null simulation represents repeated samples under a stated benchmark model. The share of simulated statistics at least as extreme as the observation, in the direction being tested, estimates how surprising the data would be under that benchmark. It is not the probability that the benchmark is true. Statistical evidence against a benchmark and practical importance are separate questions: an observed change also needs an effect size and a consequence.

When recommending a rule, check both errors. A false alarm acts against a true benchmark; a missed change fails to detect a specified departure. Power describes the chance of detecting that specified departure. Use the client's stated risk limit and cost priority, then report a relevant assumption or uncertainty. Afterward, reread the conclusion alone: the reader should know what you recommend, which evidence supports it, and to whom or what it applies.

What the supplied output can answer

Output	Reasonable use
Graph of individual heights	Describe center, variation and unusual observations
Fitted response line	Predict within the studied range and check residuals
Interval for a proportion	Estimate a fixed population share
Null simulation	Evaluate evidence against a benchmark

Use this as preparation before the sample task. Each question specifies which tools are appropriate.

UNIT 5 / DAY 24 ● SYNTHESIS AND REVISION / T1–T8 ● CONTINUED

Independent performance task

Three questions about a tree nursery

A nursery has three summaries. Sapling heights at planting had medians of 34 cm in Plot A and 42 cm in Plot B. For trees aged 2–5 years, a line predicts height as 16 + 5 × age in years, in centimeters. A 95% interval for first-season survival among last year's plantings is 72%–84%.

1. Which plot had taller saplings at planting? The median in B was 42 − 34 = 8 cm higher. This compares the centers of the two recorded groups.

2. What height does the model predict at age 4? Four years is within the fitted range. The prediction is 16 + 5(4) = 36 cm. Individual four-year-old trees can be taller or shorter.

3. Did more than 80% survive their first season? The interval includes rates below and above 80%, so it does not establish survival above 80% for last year's plantings.

The two medians describe height at planting, the line predicts height at a specified age, and the interval estimates a share. Identify the quantity and its units before choosing a number.

KEY IDEA	A complete answer names what the numbers measure and explains how they bear on the question.
WATCH FOR	A confident prediction or policy decision still needs an appropriate study and an explicit limit.

Sources and adaptation. Synthesis of OpenStax, Introductory Statistics 2e, §2.7 (standardized distance), §3.4 (conditional rates), §8.3 (proportion intervals), §§9.1–9.3 (hypotheses, errors, and probability evidence), and §12.3 (regression). The sapling illustration and assessment contexts are original classroom examples; this lesson adds no new procedure.

UNIT 5 / DAY 25 ● SYNTHESIS AND REVISION / T1–T8

Revision and independent transfer

Repair an earlier answer and use the same reasoning in a fresh problem.

Use this page during the revision workshop. The independent transfer task has its own directions.

Revision begins with the earliest decision that needs to change. If the arithmetic is wrong, recompute it and update later work affected by the error. If the number is correct but the interpretation reverses the sign, retain the calculation and repair the language. If the population is wrong, a narrower conclusion or a better design is needed; repeating the arithmetic will not repair that problem.

Use the rubric to locate the issue. Context names the cases and measured outcome. Evidence and method supplies appropriate calculations or observations. Reasoning explains how the evidence bears on the claim. The Conclusion and scope row asks whether the answer addresses the question within the study's limits. A revision should improve the affected part without discarding valid work elsewhere.

Explain why the correction changes the answer. "I changed −4 to +4" reports an edit; "Residual means observed minus predicted, so this longer-than-predicted wait needs a positive residual" states the principle. If your original work is correct, change one condition, such as volunteer selection instead of random selection, and explain whether the original conclusion still follows.

Then attempt a fresh case independently, with the marked solution put away. New numbers and a different setting require you to choose the relevant group, comparison, or rule again. A copied correction shows what the feedback said; successful transfer shows that you can use the idea yourself. Keep one short reminder of the question you should ask before accepting a statistical claim outside class.

Match the repair to the evidence

Observed issue	Check in the revision
Correct result, wrong population	Who was selected and what was measured?
Residual sign reversed	Which value is observed and which is predicted?
Interval treated as certainty	What quantity is fixed across repeated sampling?
Power reported without an alternative	Under what specified true rate was detection measured?

Choose an actual error from your marked work. Give a reason even when the new number is correct.

UNIT 5 / DAY 25 ● SYNTHESIS AND REVISION / T1–T8 ● CONTINUED

Revision and independent transfer

Repair the interpretation

A fitted model predicts a 20-minute wait; the observed wait is 16 minutes. A response gives a residual of −4 minutes and says the wait was longer than predicted.

1. The calculation 16 − 20 = −4 is correct. The sign means the observation lies below the prediction.

2. A new check: if another pickup took 18 minutes against a 15-minute prediction, its residual would be +3 minutes.

5

The first wait was four minutes shorter than predicted. The correction changes the interpretation and preserves the valid calculation.

KEY IDEA	Name the error, change the reasoning it affects, and demonstrate the idea once with fresh evidence.
WATCH FOR	A copied correction can look complete. A new question reveals whether the underlying decision is understood.

Sources and adaptation. Revision examples synthesize OpenStax, Introductory Statistics 2e, §1.2 (population scope), §8.3 (interval interpretation), §9.2 (errors and power), and §12.3 (observed minus predicted residuals). The revision workflow, rubric, and numerical wait-time examples are original course materials.

QUICK REFERENCE

Choosing a defensible claim

Start with the study question and identify what was measured. The same numbers may support different conclusions when the sampling frame or study design changes.

Question	Evidence to use	Limit to check
Which group is typically faster or more consistent?	Compare centers and spreads in the same units. Inspect shape and unusual values.	Identify the groups and whether the observations represent a wider population.
Which observation is more unusual in its group?	Compute each signed z-score from its own reference mean and positive standard deviation.	A relative-standing comparison does not imply a particular percentile.
Can this fit support a prediction?	Use a supplied fit and inspect residuals. Recompute using another reasonable specification.	Name the observed predictor range. Associations alone do not establish a causal effect.
What share of the population meets a condition?	Use the sampled count and denominator, then interpret the endpoints of a supplied interval.	Identify who could enter the sample and why repeated samples would vary.
Is the outcome unusual under the benchmark?	Prespecify the direction. In a null simulation, count outcomes at least as extreme as the observation.	Report effect size, the preselected threshold, and a consequence or cost.
Which decision rule fits a stated risk limit?	Compare false-alarm rates under the benchmark and power under a specified alternative.	Compare rates with matching denominators. Name whose risk changes when the rule changes.

Choosing a procedure from the question

Suppose a school sampled 120 lunch orders and found that 78 used a reusable container. An estimate reports the sample share, 65%, and uses an interval to express uncertainty about the population share. To ask whether the evidence supports an increase over an earlier benchmark, first state the benchmark and the direction, then examine outcomes under that benchmark.

To recommend changing a rule, identify the concrete cost of a false alarm and of a missed change. Two decision makers who agree on the data can reasonably prefer different rules if they face different costs.

> **USE BEFORE SUBMITTING**
>
> Name the people or objects, measured variable, units, and relevant comparison. Tie each claim to a number or a feature of a display. Write the final decision at the scope the data support.

QUICK REFERENCE

Notation and calculations

Quantity	Calculation or meaning	Illustrative result
Sample proportion, $\hat{p}$	Count meeting the condition divided by the total sampled.	18 of 30 trips were on time. $\hat{p}$ = 18/30 = 0.60.
Interquartile range, IQR	Third quartile minus first quartile. Same units as the observations.	Quartiles 8 and 14 minutes give IQR = 6 minutes.
Standardized score, z	(Value − reference mean) / reference SD. Requires a positive SD.	A 13-minute run in a group with mean 17 and SD 2 is 2 SDs below the mean: z = −2.
Residual	Observed response minus predicted response at the same predictor value.	A run took 16 min and the prediction was 14 min. The residual is +2 min.
Interval width	Upper endpoint minus lower endpoint.	A 54%–69% interval spans 15 percentage points.
Simulated p-value	Extreme simulated samples divided by all simulated samples under the null model.	21 of 500 null batches are at least as extreme in the prespecified direction. Estimated p-value = 0.042.
Power	Probability a rule flags a change when a specified alternative is true.	At the specified alternative, 74% flagged means 74% power and 26% missed.

Percentage points and relative change

If a rate rises from 20% to 25%, its absolute increase is 5 percentage points. Its relative increase is 5/20 = 25% of the original rate. Give the original rate whenever the comparison could be ambiguous.

Read a conditional rate aloud

True condition	Flag	No flag	Total
Benchmark actually holds	6	94	100
Specified change actually exists	72	28	100

In these illustrative repeated samples the false-alarm rate is 6/100 and power at the specified change is 72/100. Of the 78 total flags, 72 came from the second row. The fraction 72/78 answers a different question because its denominator is all flagged samples.

Reference group: Always name the group or condition that supplies the denominator. For a false alarm, condition on the benchmark being true. For a miss, condition on a specified alternative.