

Unit 1: Exploring One-Variable Data and Collecting Data

Unit 1 is the biggest unit in AP Statistics and the foundation for everything after it. It covers how data are collected, how one variable is displayed and described, and how studies are designed so their conclusions mean something. Every later unit assumes you can describe a distribution, choose the right summary numbers, and tell a good study from a bad one.

AP Statistics

Exploring One-Variable Data and Collecting Data

About 14 minutes to read

How to use this guide

Read it in order the first time because the unit builds in two arcs. The first arc is data itself: what a variable is, how to display one variable, how to describe its distribution, and which numbers summarize it. The second arc is collecting data: how studies are designed, how samples are chosen, and where bias creeps in. Exam questions almost always pair these arcs, asking you to describe data and then judge whether the study behind it supports the claim being made.

After the first read, use the trap boxes and the tables to review the distinctions the exam tests most often. Finish with the practice questions, then complete the recall check on the last page out loud and note any items you cannot explain yet.

What this unit is worth. Exploring one-variable data and collecting data is roughly 15 to 23 percent of the AP Statistics exam, and it is tested in every free-response section because study design and distribution descriptions appear inside later questions too. The vocabulary here is also the language of the whole course, so learning it precisely now pays off in every unit that follows.

1.1 What Statistics Studies

A **statistical study** collects data from a sample to answer an **investigative question** about a larger population. The **population** is every item or individual of interest, with total size N . The **sample** is the subset the data actually come from, with size n . A single piece of information is a **datum**, the collection of them is a **data set**, and the item or individual a datum is collected from is the **observational unit**.

An investigative question should have a defined purpose, be phrased so the needed data can actually be collected, and make clear whether the goal is estimating a parameter or testing a claim. At the end of a study, answers must be given **in context**, meaning each number is tied back to the real-world thing it describes. A mean of 24.3 means nothing until you say it is the mean number of text messages per day for the students surveyed.

Trap. Conclusions without context lose credit. If a question asks what the median tells you, "the median is 42" is an incomplete answer. Say the median is 42 *what*: 42 minutes of screen time per day for the students in the sample, for example.

1.2 Variables: Categorical and Quantitative

A **variable** is a characteristic that can change from one observational unit to another. A **categorical variable** takes category names or group labels, like eye color or preferred lunch. A **quantitative variable** takes numerical values for something measured or counted, usually with units, like height in centimeters or number of siblings.

Quantitative variables split further. A **discrete quantitative variable** takes a countable number of values, like the number of cars in a household. A **continuous quantitative variable** can take any value in an interval, like the exact time a runner finishes a race. On the exam, identifying the variable type is the first step, because it decides which displays and summaries are allowed.

A **parameter** is a numerical summary of a population, and a **statistic** is the corresponding summary of a sample. Parameters are what we want to know and almost never do. Statistics are what we compute from the data in hand, and they estimate the parameters.

Trap. Parameter and statistic are not interchangeable. A parameter describes the whole population and is usually unknown. A statistic describes the sample you actually measured. Mixing them up is the same as confusing the thing you want with the thing you have.

1.3–1.4 Displaying One Categorical Variable

Start with a table. A **frequency table** shows the count in each category, and a **relative frequency table** shows the proportion in each category. The **relative frequency** of a category is its count divided by the total, expressible as a proportion or a percentage.

Then draw the picture. A **bar chart** displays frequencies or relative frequencies for the categories of a single categorical variable, with a separate bar per category. A **pie chart** shows the same information as slices of a circle, where each slice's area is the fraction of the whole that category represents. Both can compare two or more data sets on the same categorical variable, for example the lunch preferences of freshmen versus seniors side by side.

Choose relative frequencies when the groups being compared have different totals. If 60 of 100 freshmen and 90 of 200 seniors prefer pizza, the raw counts make seniors look like bigger pizza fans, but the relative frequencies tell the opposite story: 60 percent of freshmen versus 45 percent of seniors.

Trap. Comparing raw counts across groups of different sizes is misleading. Convert to relative frequencies first. A bar chart of counts makes the bigger group look bigger in every category, which is an artifact of group size, not of preference.

1.5 Displaying One Quantitative Variable

Quantitative data need displays that show the pattern of values. A **dotplot** puts one dot per value above a number line, so you see every observation and any clumping or gaps. A **stem-and-leaf plot** splits each value into leading digits (the stem) and a trailing digit (the leaf), ordering both, which keeps the actual values visible while showing the shape. A back-to-back stem-and-leaf plot puts two distributions on either side of a shared stem column for easy comparison.

A **histogram** groups values into ordered intervals called bins along the horizontal axis, with each bar's height showing how many values fall in that bin. Unlike a bar chart, the bars touch, because the horizontal axis is a continuous number line. The **distribution** is the pattern the display reveals: which values occur and how often.

Trap. Bar charts and histograms are not the same display. Bar charts are for categorical variables and have separated bars, one per category. Histograms are for quantitative variables and have touching bars, because the bins are intervals on a number line. Using the wrong one signals that the variable type was misidentified.

1.6 Describing a Distribution: SOCS

Every distribution description on the exam should cover four things, remembered as SOCS: **shape**, **outliers**, **center**, and **spread** (variability). **Shape** is the overall pattern: whether it is skewed or symmetric and how many peaks it has. **Center** is a typical or middle value. **Variability**, also called spread, is how spread out the values are. **Outliers** are points unusually small or large relative to the rest.

For shape, learn the vocabulary precisely. **Skewed right** (positively skewed) means the right tail toward larger values stretches longer than the left tail. **Skewed left** (negatively skewed) is the mirror image: the left tail toward smaller values is longer. **Approximately symmetric** means the left half nearly mirrors the right half. **Unimodal** means one main peak, **bimodal** means two prominent peaks, and **approximately uniform** means the frequencies are nearly the same everywhere with no real peaks. Also watch for **gaps**, regions with no data, and **clusters**, concentrations of values separated by gaps.

Trap. Skew is named for the direction of the long tail, not the side where most data pile up. A distribution with most values bunched on the left and a long tail stretching right is skewed right, even though the bulk of the data sits left. Read the tail, not the pile.

1.7 Summary Statistics: Center

The **mean**, $\bar{x}$ for a sample, is the sum of all values divided by the number of values. The **median** is the middle value when the data are ordered from smallest to largest. Both measure center, but they react differently to extreme values, which is why the exam keeps asking you to choose between them.

The mean is **nonresistant**: outliers pull it toward themselves, so in a right-skewed distribution the mean is usually larger than the median, and in a left-skewed distribution it is usually smaller. The median is **resistant**: extreme values barely move it. When a distribution is skewed or has outliers, report the median. When it is roughly symmetric with no outliers, the mean is fine, and the two will be close to each other.

Trap. The most tested sentence in this unit is the mean-median relationship to shape. Skewed right means mean greater than median. Skewed left means mean less than median. Symmetric means they are close. Do not reverse it, and do not claim it for bimodal or uniform distributions, where the rule does not apply the same way.

1.7 Summary Statistics: Spread

Spread has several measures. The **range** is the maximum minus the minimum: quick but nonresistant, since one outlier changes it completely. The quartiles divide ordered data into quarters. The **first quartile (Q1)** is the median of the lower half, about 25 percent of values at or below it. The **third quartile (Q3)** is the median of the upper half, about 75 percent at or below it. The **second quartile (Q2)** is just the median again. A **percentile** generalizes this: the value with p percent of the data at or below it.

The **interquartile range (IQR)** is $Q3 - Q1$, the spread of the middle 50 percent of the data. Like the median, it is resistant. The **standard deviation**, s for a sample, is a typical distance of values from their mean: the square root of the sum of squared deviations divided by $n - 1$. Its square is the **sample variance**, s^2 . The standard deviation is nonresistant, because squaring makes extreme deviations count heavily.

Outliers can be flagged two ways. The **$1.5 \times \text{IQR}$ rule** calls a value an outlier if it sits more than $1.5 \times \text{IQR}$ below $Q1$ or above $Q3$. The **2-standard-deviation rule** flags values more than 2 standard deviations from the mean. The IQR rule is the one boxplots use, and it pairs with the resistant summaries.

PAIR	RESISTANT	NONRESISTANT
Center	Median	Mean
Spread	IQR	Range, standard deviation

Trap. Match the summaries to the shape. Skewed data or data with outliers get the median and IQR. Symmetric data with no outliers get the mean and standard deviation. Reporting the mean for heavily skewed data, or the standard deviation when outliers dominate, is the wrong toolkit for the data.

1.8 Boxplots and the Five-Number Summary

The **five-number summary** is the minimum, Q1, the median, Q3, and the maximum. A **boxplot** draws it: a box from Q1 to Q3 with a line at the median, and **whiskers** extending to the minimum and maximum. When outliers exist by the $1.5 \times \text{IQR}$ rule, the whiskers stop at the most extreme non-outlier values and the outliers are plotted as individual points beyond them.

Read skew from a boxplot by comparing the halves. If the median sits left of center in the box and the right whisker stretches longer, the distribution is skewed right. If the median sits right of center with a longer left whisker, it is skewed left. A roughly centered median with even whiskers suggests symmetry. Boxplots hide bimodality and gaps, so they are summaries, not complete pictures.

Changing units works the way you would expect. Adding a constant to every value shifts the center measures (mean, median, quartiles) by that constant but leaves spread measures (range, IQR, standard deviation) unchanged. Multiplying every value by a constant scales both center and spread by that constant.

Trap. A boxplot cannot show everything. Two distributions can share a five-number summary while having different shapes, because the boxplot discards bimodality, gaps, and clusters. If the question asks about shape details, the boxplot alone is not enough.

1.9 Comparing Distributions and z-Scores

To compare two distributions of the same quantitative variable, describe each with SOCS and then contrast them directly: which center is larger, which spread is wider, how the shapes differ, and whether one has outliers the other lacks. Back-to-back stem-and-leaf plots and side-by-side boxplots are built for exactly this.

A **z-score**, also called a **standardized score**, measures how many standard deviations a value falls above or below the mean: $z = (x - \bar{x}) / s$, using the population mean and standard deviation when those are known. Positive z means above the mean, negative means below. Because z-scores strip away units, they let you compare values from different distributions, like a test score of 82 in a class with mean 75 against a score of 88 in a class with mean 90.

Trap. Raw scores from different distributions are not directly comparable. A higher raw score can be the weaker performance if its distribution has a higher mean. Convert to z-scores before comparing across distributions, and keep the sign: a z-score of -1.5 is below average no matter how big the raw number looks.

1.10 Studies: Observational or Experiment

Studies divide by whether the researcher imposes conditions. In an **observational study**, treatments are not imposed; the researcher records the values of the variables of interest as they are. A **survey** is an observational study that collects data from people with a standard set of questions. A **prospective study** selects units now and follows them into the future, while a **retrospective study** selects units now and looks back at past data. A **census** records information from every member of the population rather than a sample.

In an **experiment**, the researcher assigns treatments to units and measures the response. The **experimental units** are the individuals or items receiving treatments (people are called **subjects** or participants). The **explanatory variable**, also called a factor, is the variable whose levels are imposed. Each level or combination of levels is a **treatment**. The **response variable** is the outcome measured after treatment.

This distinction decides what conclusions are allowed. A **cause-and-effect conclusion**, attributing differences in the response to the explanatory variable, is valid only from a well-designed experiment with random assignment. In an observational study, a **confounding variable** can always offer an alternative explanation for the observed relationship, so the study can show association but not causation.

Trap. The single most tested idea in study design: observational studies cannot establish cause and effect. If a survey finds that coffee drinkers live longer, a confounding variable like income or exercise habits could explain it. Only random assignment in an experiment breaks the link between the treatment and the confounders.

1.11 How Samples Are Chosen

A **random sample** selects units using a random mechanism, often a **random number generator**. The gold standard is the **simple random sample (SRS)**: every possible sample of size n has the same chance of being chosen. **Sampling without replacement** selects each unit at most once; **sampling with replacement** returns each unit before the next draw so it could be chosen again. Real studies almost always sample without replacement.

Three other random methods handle populations with structure. A **stratified random sample** divides the population into non-overlapping homogeneous groups called **strata**, then takes a random sample within each stratum. A **cluster random sample** divides the population into groups called **clusters**, randomly selects some clusters, and collects data from everyone in the chosen clusters. A **systematic random sample** picks a random starting point and then every k th individual after that.

METHOD	HOW IT WORKS	WHEN IT FITS
Simple random sample	Every sample of size n equally likely.	The default when the population has no useful structure.
Stratified	Divide into homogeneous strata, sample within each.	You want guaranteed representation of each subgroup, like each grade level.
Cluster	Randomly pick whole clusters, survey everyone inside.	The population is already grouped, like classrooms, and listing everyone is impractical.
Systematic	Random start, then every k th person.	You have an ordered list and want an even spread through it.

Trap. Stratified and cluster sampling are easy to confuse because both divide the population into groups. The difference is what happens next. Stratified samples *within* every group. Cluster samples *whole groups* at random. Also remember that random selection lets results generalize to the population, which is a separate question from whether a treatment caused the effect.

1.12 What Can Go Wrong with Sampling

Bias in a sampling method is a systematic error that makes the statistic consistently too high or consistently too low relative to the parameter. Bias is about direction, not size: a biased method misses the same way every time, while random sampling error bounces around the truth.

Learn each bias by its mechanism. **Voluntary response bias** comes from samples of pure volunteers, who tend to hold strong opinions. **Undercoverage bias** happens when part of the population is left out of the sampling frame or is less likely to be selected. **Nonresponse bias** happens when chosen individuals do not respond and differ from those who do. **Response bias** is answers that systematically differ from the truth, and **question wording bias** is response bias caused by confusing or leading questions. **Nonrandom sampling methods** like convenience sampling invite all of these because chance never gets a say in who is selected.

Trap. Bias is systematic, not random. A small sample has more variability but is not automatically biased, and a large sample does not fix bias. If the method systematically excludes a group, taking more data just reproduces the same error with more confidence.

1.13 Designing Experiments

A **well-designed experiment** compares at least two groups getting different conditions, assigns those conditions by **random assignment**, and uses enough experimental units per treatment (**replication**) to see real differences through the noise. Random assignment is what makes the treatment groups similar in every way except the treatment, so differences in the response can be attributed to the explanatory variable rather than to a **confounding variable**.

The **control group** provides the comparison baseline. In drug trials it often receives a **placebo**, an inactive treatment, because the **placebo effect** means patients can improve just from believing they were treated. **Single-blind** experiments hide the treatment from participants; **double-blind** experiments hide it from the interacting researchers too, so expectations cannot leak into the measurements.

Extraneous variables are other factors that affect the response. Two tools handle them. **Direct control** holds a factor constant across all units. **Blocking** groups units by a **blocking variable** first, then randomizes within each **block**, which separates that variable's effect from the treatment effect. A **completely randomized design** skips blocking and assigns units to treatments entirely at random. A **matched pairs design** is a randomized block design with two treatments, pairing similar units (or using the same unit twice) before randomizing within each pair.

Trap. Blocking is not the same as stratifying, though both group units first. Blocking happens in experiments before random assignment of treatments, and its purpose is to reduce variation from a known extraneous variable. Stratifying happens in sampling before random selection, and its purpose is guaranteed representation. One serves experiments, the other serves surveys.

Confusions That Cost Points

PAIR	HOW TO KEEP THEM STRAIGHT
Parameter vs statistic	The parameter describes the population and is usually unknown. The statistic describes the sample and is computed from data.
Bar chart vs histogram	Bar charts display categorical variables with separated bars. Histograms display quantitative variables with touching bars on a number line.
Skewed right vs skewed left	Name the long tail. Right tail longer means skewed right. Left tail longer means skewed left.
Mean vs median	Mean is nonresistant and pulled toward skew. Median is resistant. Skewed right: mean above median. Skewed left: mean below median.
Standard deviation vs IQR	Standard deviation pairs with the mean for symmetric data. IQR pairs with the median for skewed data or data with outliers.
Stratified vs cluster sampling	Stratified samples within every group. Cluster randomly selects whole groups and surveys everyone inside.
Observational study vs experiment	Observational studies record without imposing treatments and can show only association. Experiments assign treatments with random assignment and can support cause and effect.
Random sampling vs random assignment	Random sampling (selection) lets results generalize to the population. Random assignment (of treatments) lets results support causation. They answer different questions.
Bias vs random error	Bias is systematic and misses the same direction every time. Random error bounces around the truth and shrinks with larger samples. Bigger samples do not fix bias.
Blocking vs stratifying	Blocking groups units in an experiment before random assignment to control an extraneous variable. Stratifying groups a population in sampling to guarantee subgroup representation.

Practice Questions

Original questions written for this guide in the style of the AP exam. Answers and explanations are on the next page, so complete the questions before checking them.

- 1.** A histogram of household incomes in a town is strongly skewed right with a few very high incomes. Which pair of summaries best describes the center and spread of this distribution?
 - A. Mean and standard deviation
 - B. Mean and interquartile range
 - C. Median and interquartile range
 - D. Median and standard deviation

- 2.** A data set has $Q1 = 20$, $Q3 = 44$, and one value of 85. Using the $1.5 \times \text{IQR}$ rule, is 85 an outlier?
 - A. Yes, because 85 is greater than $Q3$
 - B. Yes, because 85 exceeds $Q3 + 1.5 \times \text{IQR}$
 - C. No, because 85 is less than $Q3 + 2 \times \text{IQR}$
 - D. No, because the $1.5 \times \text{IQR}$ rule only applies to values below $Q1$

- 3.** A school district wants to survey parents about a new schedule. It divides parents by elementary, middle, and high school level, then randomly selects 50 parents within each level to survey. This is an example of
 - A. a simple random sample
 - B. a stratified random sample
 - C. a cluster random sample
 - D. a systematic random sample

- 4.** Researchers randomly assign 200 volunteers with high blood pressure to either a new exercise program or a control group that keeps its usual routine. After six months they compare average blood pressure. This study can support a cause-and-effect conclusion because
 - A. the sample was randomly selected from the population
 - B. treatments were randomly assigned to the volunteers
 - C. the study used a large sample size
 - D. blood pressure is a quantitative variable

Answer Key

- 1. C.** Strong right skew means outliers pull the mean upward, so the resistant summaries are the right toolkit: median for center, IQR for spread. A pairs the mean with the standard deviation, which is correct only for roughly symmetric data. B mixes a nonresistant center with a resistant spread. D mixes a resistant center with a nonresistant spread.
- 2. B.** The IQR is $44 - 20 = 24$, so the upper fence is $Q3 + 1.5 \times \text{IQR} = 44 + 36 = 80$. Since 85 exceeds 80, it is an outlier. A is wrong because merely exceeding Q3 does not make a value an outlier; the fence is what matters. C applies the wrong rule and the wrong fence. D misstates the rule, which checks both tails.
- 3. B.** The district divided parents into homogeneous groups (school levels) and sampled within each group, which is stratified random sampling. A is wrong because not every sample of 150 parents was equally likely; the level quotas were fixed. C would require randomly selecting whole levels and surveying everyone in the chosen ones. D would require a random start and a fixed interval through a list.
- 4. B.** Random assignment balances extraneous variables across the treatment groups, so the only systematic difference between them is the exercise program. That is what licenses a cause-and-effect conclusion. A confuses random selection (which supports generalization) with random assignment (which supports causation). C is wrong because size reduces variability but does not create comparability. D is irrelevant; the variable type has nothing to do with causal conclusions.

One-Page Recall Check

- Explain the difference between a population and a sample, and between a parameter and a statistic.
- Classify a variable as categorical or quantitative, and quantitative variables as discrete or continuous.
- Explain when to use relative frequencies instead of counts when comparing groups.
- State the difference between a bar chart and a histogram and when each is appropriate.
- Describe a distribution using SOCS: shape, outliers, center, and spread.
- Define skewed right and skewed left in terms of the tails.
- State the mean-median relationship to shape for symmetric, right-skewed, and left-skewed distributions.
- Explain why the median and IQR are resistant while the mean and standard deviation are not.
- Apply the $1.5 \times \text{IQR}$ rule to flag outliers.
- Read skew from a boxplot using the median position and whisker lengths.
- Compute a z-score and explain what its sign tells you.
- Explain why an observational study cannot establish cause and effect.
- Distinguish stratified, cluster, systematic, and simple random sampling.
- Name the five sampling biases and the mechanism behind each.
- Explain what random assignment accomplishes in an experiment.
- Distinguish blocking from stratifying by purpose and setting.

Where to go next. Turn every missed item above into flashcards and drill them spaced out over several days rather than in one sitting. In Rycal, open the Exploring One-Variable Data deck under AP Statistics. The deck covers the terms in this guide, and its practice questions target the same traps named here. If you have a test date, add it in the Test Planner. You can also start your next review with a Brain Dump, then check what you missed against this guide. [Study this unit in Rycal.](#)

Key terms for this unit

Statistical study, Investigative question, Datum, Data set, Population, Sample, In context, Observational unit, Variable, Parameter, Statistic, Categorical variable, Quantitative variable, Discrete quantitative variable, Continuous quantitative variable, Frequency table, Relative frequency table, Relative frequency, Bar chart, Pie chart, Comparing multiple categorical data sets, Histogram, Stem-and-leaf plot, Dotplot, Distribution, Shape (of a distribution), Center (of a distribution), Variability (spread), Skewed right (positively skewed), Skewed left (negatively skewed), Approximately symmetric, Unimodal, Bimodal, Approximately uniform, Outlier, Gap, Cluster (of values), Mean, Median, Minimum value, Maximum value, First quartile (Q1), Third quartile (Q3), Second quartile (Q2), Percentile, Range, Interquartile range (IQR), Standard deviation, Sample variance, Resistant (robust), Nonresistant (non-robust), $1.5 \times \text{IQR}$ rule for outliers, 2-standard-deviation rule for outliers, Changing units of measurement, Five-number summary, Boxplot, Whiskers, Mean-median relationship to shape, Back-to-back stem-and-leaf plot, Standardized score, z-score, Components of an investigative question, Investigative question for a hypothesis test, Investigative question for a confidence interval, Cause-and-effect conclusion, Census, Experiment, Experimental unit, Subject (participant), Explanatory variable (factor), Treatment (level), Response variable (experiment), Observational study, Prospective study, Retrospective study, Survey, Confounding variable (observational study), Random sample, Scope of generalization, Sampling without replacement, Sampling with replacement, Simple random sample (SRS), Stratified random sample, Stratum (strata), Cluster random sample, Cluster (sampling group), Systematic random sample, Random number generator, Justifying an appropriate sampling method, Bias (in a sampling method), Voluntary response bias, Undercoverage bias, Nonresponse bias, Response bias, Question wording bias, Nonrandom sampling methods, Well-designed experiment, Control group, Placebo, Placebo effect, Single-blind experiment, Double-blind experiment, Extraneous variable, Random assignment, Confounding variable