The Ultimate AP® Statistics Unit 1 Cheat Sheet

1.1–1.2 Data, Variables & Questions

Investigative Questions

Statistics = learning from data w/ **variation**. Start w/ a question, collect data, analyze, conclude in context.
Investigative question: names the population + variable(s); answered by data, not 1 fact. e.g., "What is the typical commute time, in min, for students at this school?"
Population (size N) = whole group of interest. **Sample** (size n) = part we actually measure to learn about the population.
Individuals = objects described (people, cars, days).
Variable = trait that varies across individuals.

Parameter vs. Statistic

Parameter: describes **p**opulation (μ, σ, p). **Statistic**: describes **s**ample ($\bar{x}$, s, $\hat{p}$). Statistic estimates parameter.
Sampling variability: diff samples → diff statistics. Bigger n → statistics vary less around the true parameter.
Memory trick: **P**arameter ↔ **P**opulation; **S**tatistic ↔ **S**ample.

Types of Variables

Type	Values	Example
Categorical	Groups/labels	Eye color, ZIP
Quant. discrete	Countable #s	# of siblings
Quant. contin.	Any # in range	Height, time

Test: does avg make sense? Yes → **quantitative**. No → **categorical**. ZIP code, jersey #, area code = categorical (numbers as labels) → avg is meaningless.
Discrete: gaps between values (counts). **Continuous**: infinitely many values in an interval (measurements).
Quantitative → categorical: binning ages into "under 18 / 18+" makes a categorical variable (age group).
Univariate data = 1 variable (this unit). **Bivariate** = 2 variables, e.g., height vs. weight, shown in a scatterplot.

⚠ A variable isn't quantitative just b/c it's written as a number (ZIP, jersey #). Ask what the number MEANS.

✓ FRQ: name the variable AND the individuals in context ("heights, in cm, of 30 7th graders").

1.3–1.4 One Categorical Variable

Tables

Frequency table: count per category. **Relative frequency** = count ÷ total → proportion or %. Rel. freqs sum to 1 (100%), ± rounding (e.g., 0.33 + 0.33 + 0.33 = 0.99, not 1).

Pet	Count	Rel. freq.
Dog	45	45/120 = 0.375
Cat	39	0.325
None	36	0.300
Total	120	1.000

Use rel. freq. to compare groups of diff sizes. 30 of 50 (60%) > 40 of 100 (40%), though 40 > 30.
Proportion ↔ %: 0.375 = 37.5%. Report the denominator: "45 of 120 (37.5%)" so readers see the sample size.

Graphs

Bar chart: 1 bar per category, height = freq or rel. freq. Bars separated (categories ≠ number line).
Pie chart: slices = parts of 1 whole; must sum to 100%. Hard to compare similar slices → bar chart usually better.
Mode = most common category. No mean/median/shape for categorical data (use counts, %s & the mode).
Segmented (stacked) bar: 1 bar = 100%, split into categories. Good for comparing distributions across groups.
Y-axis label: "Count" vs. "Proportion"; check which before reading values (same bars, different scale).

Reading & Misleading Graphs

Truncated axis (not starting at 0) exaggerates diffs.
Pictographs: area grows faster than height → misleads.
Describe categorical data w/ proportions in context: "37.5% of the 120 surveyed students own a dog."

⚠ Never describe the "shape" or "skew" of a bar chart for categorical data: the order of the bars is arbitrary.

✓ Comparing groups of unequal size? Always compare %s, never raw counts b/c group sizes differ.

1.5 Graphing One Quantitative Variable

Graph Types

Graph	Best for	Keeps data?
Dotplot	Small n	Yes, each value
Stemplot	Small n, 2 grps	Yes, w/ key
Histogram	Large n	No, only bins
Boxplot	Compare grps	No, 5-# summ.

Dotplot: 1 dot per value on a number line. Shows gaps, clusters, outliers clearly that a histogram can hide.
Stemplot: stem = leading digit(s), leaf = last digit. Always include a **key** ("3 | 7 = 37 min"). **Split stems** (0–4, 5–9) spread out crowded data when leaves pile up.
Back-to-back stemplot: shared stems, leaves on both sides → compare 2 groups (e.g., boys vs. girls).
All 4 show distribution: what values occur + how often. Dotplots and stemplots keep every value in order.

Histograms

Histogram: bars touch (number line). Equal-width bins; each value goes in exactly 1 bin. Height = count or rel. freq.
Bin width alters look: too wide hides detail, too narrow looks noisy. Rel. freq. histogram lets you compare diff n.
Reading bins: you can find # of values in a bin, not exact values → can't compute exact mean/median.
Boundary rule: a value on a bin edge goes in the bin to its right, e.g., 20 goes in 20–<30, not in 10–<20.
Find % from a histogram: add bar heights in range ÷ total n.

Cumulative Graphs

Cumulative rel. freq. plot: y = proportion of values ≤ x. Read percentiles: x where y = 0.25 → Q1; 0.75 → Q3.
Median = x where cumulative rel. freq. = 0.50.

⚠ Histogram ≠ bar chart. Histograms show quantitative data (bars touch); bar charts show categories (gaps).

✓ Label both axes w/ variable + units and give a scale (start, step size). Missing labels cost FRQ points.

1.6 Describing a Distribution

The 4 Features (SOCS)

Shape, **O**utliers/unusual features, **C**enter, Spread. All 4, every time, in context of the variable & its units, e.g., min.

Shape	Looks like	Mean vs. med.
Symmetric	Mirror halves	Mean ≈ median
Skewed right	Long right tail	Mean > median
Skewed left	Long left tail	Mean < median
Uniform	Flat, no peak	Mean ≈ median

Skew direction = tail direction, not where the peak is. Income, wait times → usually skewed right.
Modes: unimodal (1 peak), **bimodal** (2 peaks → maybe 2 groups mixed, e.g., heights of boys + girls).

Unusual Features

Outliers (far from rest; check w/ a rule), **gaps** (no data in a region), **clusters** (data bunch in groups).
Outlier ≠ error: never delete an outlier unless it's a recording mistake; report it; analyze w/ & w/o it if unsure.
Gap example: no commutes 40–70 min, then 1 value at 75 → outlier + gap → describe both in your write-up.

Writing It Up

Template: "The distribution of *commute time* for these 40 students is skewed right w/ a possible outlier at 75 min. Median ≈ 18 min; IQR ≈ 12 min; range 4–75 min."
Justify a claim: point to the graph ("most values cluster 10–20 min, few above 40") as specific evidence.
Which center/spread? Skewed/outliers → median + IQR. Symmetric → mean + SD (resistant vs. not).
Vocabulary: "typical" = center, "variability" = spread, "consistent" = small spread; "unusual" = outlier.

⚠ No context = no credit. "It is skewed right" loses points; say WHAT is skewed and use units (e.g., min).

✓ Use "approximately symmetric": real data is never perfectly symmetric, so never say "perfectly".

1.7 Center, Position & Unit Changes

Measures of Center

Mean $\bar{x} = \Sigma x/n$: balance point. Population mean = μ. Pulled toward tail and outliers (so it's not resistant).
Median: middle of ordered data. n odd → middle value; n even → avg of 2 middle values. Position = (n+1)/2.
Mean vs. median: 1 huge value (e.g., CEO salary) → mean jumps, median barely moves → median for incomes.
Mean of combined groups: weight by n: $(n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1 + n_2)$, not avg of means unless $n_1 = n_2$ (equal sizes).

Measures of Position

Q1 = median of lower half, **Q3** = median of upper half (skip the median if n odd). Q1–Q3 = middle 50%.
pth percentile: p% of values are at or below it. Q1 = 25th, median = 50th, Q3 = 75th percentile of the data.
Calculator: 1-Var Stats gives $\bar{x}$, Sx, n, min, Q1, med, Q3, max. Write each value w/ its label, not just "calc."
Percentile ≠ % correct: 90th percentile score = scored ≥ 90% of test takers, not 90% on the test.

Changing Units (Linear Transformations)

Do to data	Center/position	Spread
Add/subtr. a	± a	No change
Multiply by b	× b	× |b|
y = a + bx	a + b(center)	|b|(spread)

Example: °F = 32 + 1.8(°C). Mean 20°C → 68°F. s = 5°C → s = 9°F (the +32 shifts center only; s & IQR stay ×1.8).
Shape never changes under a linear transformation (if b > 0).
Order: apply "× b" to everything; apply "+ a" only to center + position (mean, median, Q1, Q3, min, max).

⚠ Adding a constant shifts center but NOT spread. Most-missed MCQ idea in this topic on the exam.

✓ Skewed or outliers → report median + IQR (both resistant). Roughly symmetric → mean + SD.

1.7 Spread, Outliers & Resistance

Measures of Variability

Range = max − min; uses only 2 values → not resistant.
IQR = Q3 − Q1: spread of the middle 50%. Resistant.
Std. deviation $s = \sqrt{[\Sigma(x - \bar{x})^2/(n-1)]}$: typical distance of values from the mean. **Variance** = s^2.
Interpret s: "Commute times typically vary by ≈ 6.2 min from the mean of 18.4 min." s = 0 only if all values equal.
Why n − 1: corrects for $\bar{x}$ sitting closer to the data than μ.
Bigger s = values more spread. Compare: s = 2 min (consistent) vs. s = 9 min (much more variable).
Units: s, IQR, range have the variable's units; variance has squared units (e.g., min^2, hard to interpret).

Outlier Rules

1.5×IQR rule (AP default): outlier if x < Q1 − 1.5·IQR or x > Q3 + 1.5·IQR. Show both fences w/ numbers!
2 SD rule: outlier if more than 2 SDs from the mean: $x < \bar{x} - 2s$ or $x > \bar{x} + 2s$ (use only for roughly symmetric data).
Example: Q1 = 12, Q3 = 24, IQR = 12. Fences: 12 − 18 = −6, 24 + 18 = 42. A 75 min commute → outlier.
2 rules can disagree. Say which rule you used. 1.5×IQR is safer for skewed data (its fences are resistant).

Resistance

Resistant	NOT resistant
Median	Mean
IQR, Q1, Q3	SD, variance, range

Resistant = barely moves when an extreme value is added or altered. Remove an outlier → mean and s change a lot.
Effect of removing a high outlier: mean ↓, SD ↓, range ↓; median and IQR ≈ same (both are resistant).

⚠ Show work for outliers: compute both fences and compare. "75 looks far away" earns no credit.

1.8–1.9 Boxplots & Comparing

Five-Number Summary & Boxplots

5-number summary: min, Q1, median, Q3, max. Box = Q1 to Q3 (middle 50%), line at median inside the box.
Modified boxplot: whiskers go to the most extreme NON-outliers; outliers plotted as separate dots.
Each section ≈ 25% of data. Longer section = values more spread, NOT more values in that quarter of the data.
Boxplot hides n, gaps, clusters, and bimodal shape. Use a dotplot or histogram to see those features clearly.
Read skew from a boxplot: longer right whisker/box section + median near Q1 → skewed right (reverse for left).
Can't find mean or SD from a boxplot; can find range, IQR, median plus Q1, Q3, min, max & any outliers.
Outlier on boxplot: report its value + that it falls beyond the 1.5×IQR fence (show the fence calculation).

Comparing Distributions

Compare all 4 (shape, outliers, center, spread) w/ comparison words: "greater than," "less variable than."

Feature	Weak	Strong
Center	A: 18, B: 25	Med. of B > A by 7
Spread	A: IQR 12	A less variable
Shape	A skewed	Both skewed right

Graphs for 2+ groups: parallel boxplots, back-to-back stemplot, stacked dotplots, same-scale histograms.
Same scale! Comparing graphs on diff axes misleads.
Template: "Median wait at B (25 min) is greater than at A (18 min); A's waits are less variable (IQR 12 vs. 20 min)."
Context + units in every sentence; name both groups.
Overlap: "Q3 of A is below the median of B → at least 75% of A's values fall below B's median." = clear separation

⚠ Listing two sets of stats side by side is NOT a comparison. Use explicit words: higher, lower, more spread.

✓ FRQ: compare center w/ median if either group is skewed or has outliers; if both roughly symmetric, mean is OK.

1.9 z-Scores & the Normal Model

z-Scores (Relative Position)

z = (x − mean) / SD: # of SDs a value is above (+) or below (−) the mean. Unitless; compares values on diff scales.
Interpret: z = −1.5 → "Ana's time is 1.5 SDs below the mean time for her class." Give direction + context
Compare across groups: higher z = better relative position. SAT z = 1.2 vs. ACT z = 1.5 → ACT is relatively better.
Standardize all data → new mean 0, new SD 1, same shape.
Solve for x: x = mean + z·SD. z = 2, mean 50, SD 8 → x = 66.
|z| > 2 → unusual (matches 2 SD outlier rule).

Normal Distribution

Normal model N(μ, σ): symmetric, unimodal, bell-shaped. Mean = median. Only use if data are approx. normal.

Within	Empirical rule
μ ± 1σ	≈ 68% of values
μ ± 2σ	≈ 95% of values
μ ± 3σ	≈ 99.7% of values

Std. normal N(0, 1): z-scores from a normal dist. Inflection points of curve at μ ± σ (where concavity flips).
Area = proportion: total area under curve = 1. P(X = exact value) = 0 for continuous → P(X ≤ a) = P(X < a).

Calculator Moves

Find %: normalcdf(lower, upper, μ, σ). **Find value**: invNorm(area left, μ, σ) on TI-84; Desmos works too.
Show work: sketch curve, shade region, label μ, σ, boundary. Write "normalcdf(lower = 60, upper = 1000, μ = 50, σ = 8)"; bare calc syntax may lose credit on any FRQ.
Example: N(50, 8). P(X > 60) = normalcdf(60, 10^{99}, 50, 8) ≈ 0.106. Top 10% cutoff = invNorm(0.90, 50, 8) ≈ 60.3.

⚠ Empirical rule and normalcdf only work for normal data. For skewed data, use the 1.5×IQR rule, not ±2σ %s.