The Simple AP® Statistics Unit 1 Cheat Sheet

1.1–1.2 Data, Variables & Questions

Investigative Questions

Investigative question: names the population + variable; answered w/ data. e.g., "What is a typical commute time for students here?"

Population (size N) = whole group of interest. **Sample** (size n) = part we actually measure.

Individuals = people/objects described. **Variable** = trait that varies across them (e.g., height).

Parameter vs. Statistic

Parameter: number describing a population (μ, σ, p). **Statistic**: number from a sample ($\bar{x}$, s, $\hat{p}$); estimates the parameter (often unknown).

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).

Discrete: gaps between values (counts, e.g., # of pets). **Continuous**: infinitely many values in an interval (measurements, e.g., time).

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

✓ 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.

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.

Graphs

Bar chart: 1 bar per category, height = freq or rel. freq. Bars separated (categories ≠ number line). Order of bars is arbitrary (so no "shape").

Pie chart: slices = parts of 1 whole; must sum to 100%. Hard to compare similar slices → bar chart usually better (lengths beat angles).

Mode = most common category. No mean/median/shape for categorical data.

Segmented bar: 1 bar = 100%, split by category.

Reading & Misleading Graphs

Describe categorical data w/ %s + context: "37.5% of the 120 surveyed students own a dog."

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

✓ Comparing groups of unequal size? Always compare %s, never raw counts.

1.5 Graphing One Quantitative Variable

Graph Types

Graph	Best for	Keeps data?
Dotplot	Small n	Yes
Stemplot	Small n	Yes, w/ key
Histogram	Large n	No, bins
Boxplot	Compare	No

Dotplot: 1 dot per value on a number line. Shows gaps, clusters, outliers clearly at a glance.

Stemplot: stem = leading digit(s), leaf = last digit. Always add a **key** ("3 | 7 = 37 min").

Back-to-back stemplot: shared stems, leaves on both sides → compare 2 groups.

Histograms

Histogram: bars touch (x-axis is a number line). Equal-width bins; each value goes in exactly 1 bin. Height = count or rel. freq. (proportion)

Bin width alters the shape you see: too wide hides detail, too narrow looks noisy. Rel. freq. histogram lets you compare groups w/ diff n.

Boundary rule: a value on a bin edge goes in the bin to its right (20 goes in 20 to <30).

Find % from a histogram: add bar heights in the range ÷ total n, e.g., 12 of 40 = 30%.

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

✓ Label both axes w/ variable + units and give a scale. Missing labels cost FRQ points.

1.6 Describing a Distribution

The 4 Features (SOCS)

Shape, **O**utliers/unusual features, **C**enter, **S**pread. All 4, every time, in context of the variable.

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., 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 in your answer.

Writing It Up

Template: "The distribution of *commute time* for these 40 students is skewed right w/ a possible outlier at 75 min. Center ≈ 18 min (median); values range 4–75 min (IQR ≈ 12 min)."

Which center/spread? Skewed/outliers → median + IQR. Symmetric → mean + SD.

⚠ No context = no credit. "It is skewed right" loses points; say WHAT is skewed and use units.

✓ Use "approximately symmetric": real data is never perfectly symmetric in shape.

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.

Median: middle of ordered data. n odd → middle value; n even → avg of 2 middle values. Position = (n+1)/2 (e.g., n = 9 → the 5th value).

Mean vs. median: 1 huge value (e.g., CEO salary) → mean jumps, median barely moves.

Measures of Position

Q1 = median of lower half, **Q3** = median of upper half. Q1 to Q3 holds the middle 50% of data.

pth percentile: p% of values are at or below it. Q1 = 25th, median = 50th, Q3 = 75th percentile.

Calculator: 1-Var Stats gives $\bar{x}$, Sx, n, min, Q1, med, Q3, max. Write the values, not "calc."

Percentile ≠ % correct: 90th percentile = scored at or above 90% of test takers.

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 does nothing to s).

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

✓ Skewed or outliers → report median + IQR. Roughly symmetric → mean + SD.

1.7 Spread, Outliers & Resistance

Measures of Variability

Range = max − min. Uses only 2 values → very sensitive to outliers; report it w/ IQR or SD.

IQR = Q3 − Q1: spread of the middle 50%. Resistant to outliers → use w/ skewed data.

Standard deviation (s): typical distance from the mean. Bigger s → more spread. **Variance** = s^2.

Interpret s in context: "Commute times typically vary by about 6.2 min from the mean of 18.4 min."

$s \geq 0$; s = 0 only if all values equal.

Units: s, IQR, range use the data's units; variance has squared units (e.g., min^2, hard to read).

Outlier Rules

1.5×IQR rule: outlier if $x < Q1 - 1.5 \cdot IQR$ or $x > Q3 + 1.5 \cdot IQR$. Show both fences w/ numbers!

2 SD rule: outlier if more than 2 SDs from the mean: $x < \bar{x} - 2s$ or $x > \bar{x} + 2s$ (symmetric data).

Example: Q1 = 12, Q3 = 24, IQR = 12. Fences: 12 − 18 = −6, 24 + 18 = 42. A 75 min commute → outlier.

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 (median, IQR barely do).

⚠ Show work for outliers: compute both fences and compare. "75 looks far away" earns no credit on that part of the FRQ answer.

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.

Modified boxplot: whiskers go to the most extreme NON-outliers; outliers = separate dots.

Each section ≈ 25% of data. Longer section = values more spread, NOT more values.

Boxplot hides n, gaps, clusters, and bimodal shape. Use a dotplot or histogram to see those.

Read skew from a boxplot: long right whisker + median near Q1 → skewed right.

No mean or SD from a boxplot; IQR & range: yes.

Comparing Distributions

Compare all 4 features (shape, outliers, center, spread) in context w/ explicit 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

Template (names both groups + units): "Median wait at B (25 min) is greater than at A (18 min); A's waits are less variable (IQR 12 vs. 20 min)."

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

✓ FRQ: compare center w/ median if either group is skewed or has outliers (resistant).

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.

Interpret: z = −1.5 → "Ana's time is 1.5 SDs below the mean time for her class." Give direction

Compare across groups: higher z = better relative position (if higher is better). e.g., SAT z = 1.2 vs. ACT z = 1.5 → ACT is relatively better.

Normal Distribution

Normal model N(μ, σ): symmetric, single-peaked, bell-shaped curve. Mean = median. Use only if data are roughly normal (check w/ a graph).

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

Calculator Moves

Find %: normalcdf(lower, upper, μ, σ). **Find value**: invNorm(area left, μ, σ) on the TI-84 or Desmos.

Show work: sketch curve, shade region, label μ, σ, boundary. Label calculator inputs (lower =, upper =, μ =, σ =); bare calc syntax can lose credit.

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 (score).

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