Statistics Basics Cheat Sheet

Essential statistics concepts and formulas for descriptive and inferential analysis.

Measures of Center

Mean: Sum of all values / Number of values

Median: Middle value when data is ordered

Mode: Most frequently occurring value

Weighted mean: Sum of (value x weight) / Sum of weights

Measures of Spread

Range: Maximum - Minimum

IQR: Q3 - Q1 (interquartile range)

Variance: $(\text{std dev})^2$

Population std dev: $s = (\text{Sum of } (x - \text{mean})^2 / N)$

Sample std dev: $s = (\text{Sum of } (x - \text{mean})^2 / (n - 1))$

Z-Score

Formula: $z = (x - \text{mean}) / \text{std dev}$

Meaning: How many std devs a value is from the mean

z = 0: Value equals the mean

z > 0: Value is above the mean

z < 0: Value is below the mean

Five-Number Summary

Minimum: Smallest value in the data set

Q1: 25th percentile (median of lower half)

Median: 50th percentile (middle value)

Q3: 75th percentile (median of upper half)

Maximum: Largest value in the data set

Box Plot Interpretation

Box: Spans from Q1 to Q3 (contains middle 50% of data)

Line inside box: The median

Whiskers: Extend to min and max (or 1.5 x IQR)

Outliers: Points beyond 1.5 x IQR from Q1 or Q3

Skewed right: Right whisker longer, median closer to Q1

Skewed left: Left whisker longer, median closer to Q3

Normal Distribution

The 68-95-99.7 Rule (Empirical Rule)

68%: of data falls within 1 std dev of the mean

95%: of data falls within 2 std devs of the mean

99.7%: of data falls within 3 std devs of the mean

Normal Distribution Properties

Shape: Symmetric, bell-shaped curve

Center: Mean = Median = Mode

Tails: Curve approaches but never touches x-axis

Total area: Area under the curve = 1 (100%)

Descriptive vs Inferential

Descriptive: Summarizes data (mean, graphs, tables)

Inferential: Makes predictions about a population from a sample

Population: Entire group being studied

Sample: Subset selected from the population

Parameter: Measure describing a population

Statistic: Measure describing a sample

Always check whether a problem asks about a population or a sample. This determines which formula (N vs n - 1) to use for standard deviation.