Statistics Quick Reference

2

1. Sample Mean: $\bar{x} = \frac{\sum x_i}{n} = \frac{x_1+x_2+\cdots+x_n}{n}$

What it does: Calculates the average value of a dataset by summing all values and dividing by count.

Use: Measures central tendency (center point of data).

Variables: $\bar{x}$ = sample mean (average), x_i = individual data values, n = total number of observations, $\sum$ = sum all values.

2. Sample Variance: $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$

What it does: Measures how spread out data is from the mean by averaging squared deviations.

Use: Quantifies variability/dispersion in dataset.

Variables: s^2 = variance, x_i = each data value, $\bar{x}$ = sample mean, n = sample size, $(n-1)$ = degrees of freedom (corrects bias for sample vs. population).

3. Sample Standard Deviation: $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} = \sqrt{s^2}$

What it does: Square root of variance to return to original units.

Use: Shows typical spread/deviation from mean in data's original units.

Variables: s = standard deviation, s^2 = variance.

4. Empirical Rule (68-95-99.7): 68% within $\bar{x} \pm s$, 95% within $\bar{x} \pm 2s$, 99.7% within $\bar{x} \pm 3s$.

What it does: Approximates data distribution for normal data.

Use: Quick estimation of where data falls.

Variables: $\bar{x}$ = mean, s = std dev.

5. Complement Rule: $P(A') = 1 - P(A)$

What it does: Finds probability that event A does NOT occur.

Use: When calculating "not A" is easier than finding "A" directly.

Variables: $P(A')$ = probability of not A (complement), $P(A)$ = probability of event A occurring.

6. Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

What it does: Finds probability of either event occurring.

Use: Combines probabilities, avoids double-counting overlap.

Variables: $P(A \cup B)$ = prob. A or B, $P(A \cap B)$ = prob. both A and B.

6a. Independent Events: $P(A \cap B) = P(A) \cdot P(B)$

What it does: Calculates probability of both events occurring when they're independent.

Use: Test for independence: if this equation holds, events are independent; outcome of one doesn't affect the other.

Variables: $P(A \cap B)$ = prob. both A and B occur, $P(A)$ = prob. of A, $P(B)$ = prob. of B.

6b. Mutually Exclusive (Disjoint) Events: $P(A \cap B) = 0$

What it does: States that both events cannot occur simultaneously.

Use: Test for mutual exclusivity: if events can't happen together, their intersection is zero. When disjoint: $P(A \cup B) = P(A) + P(B)$.

Variables: $P(A \cap B)$ = prob. both occur (zero for disjoint events).

7. Expected Value: $E(X) = \mu = \sum x \cdot p(x) = x_1p(x_1) + x_2p(x_2) + \cdots + x_np(x_n)$

What it does: Calculates theoretical mean/average of a probability distribution by weighting each value by its probability.

Use: Predicts long-run average value over many trials.

Variables: $E(X)$ = expected value, μ = population mean, x = possible outcome values, $p(x)$ = probability of each outcome.

8. Variance of Random Variable: $V(X) = \sigma^2 = \sum(x - E(X))^2 \cdot p(x)$

What it does: Measures spread of probability distribution by averaging squared deviations from expected value.

Use: Quantifies uncertainty/variability in random variable.

Variables: $V(X)$ or σ^2 = population variance, x = outcome values, $E(X)$ = expected value/mean, $p(x)$ = probabilities.

9. Standard Deviation of RV: $SD(X) = \sigma = \sqrt{V(X)}$ where $(SD_X)^2 = V(X)$

What it does: Square root of variance.

Use: Shows spread in original units.

Variables: $SD(X)$ or σ = population standard deviation, $V(X)$ or σ^2 = variance.

10. Central Limit Theorem: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ as $n \to \infty$

What it does: States that sample means form a normal distribution regardless of original population shape.

Use: Justifies using normal distribution for statistical inference with large samples.

Variables: $\bar{X}$ = sample mean distribution, μ = population mean, σ = population std dev, n = sample size, $\frac{\sigma}{\sqrt{n}}$ = standard error (spread of sample means).

10a. Sampling Distribution of a Proportion: $\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$

What it does: Describes distribution of sample proportions.

Use: Finding probabilities about proportions in surveys/sampling. Mean = p, Standard Error = $\sqrt{\frac{pq}{n}}$.

Variables: $\hat{p}$ = sample proportion, p = population proportion, $q = 1 - p$, n = sample size.

10b. Sampling Distribution of the Sum: Sum $\sim N(n\mu, \sqrt{n}\sigma)$

What it does: Describes distribution of total/sum of n independent observations.

Use: Finding probabilities about totals rather than averages. Mean of sum = $n\mu$, Std dev of sum = $\sqrt{n}\sigma$.

Variables: Sum = total of n values, n = number of observations, μ = population mean, σ = population std dev.

11. Binomial Distribution: $X \sim \text{Bin}(n, p)$ where $P(X = x) = \binom{n}{x} p^x q^{n-x}$, $q = 1 - p$

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$ Moments: $E(X) = np$, $V(X) = npq$, $SD(X) = \sqrt{npq}$

What it does: Calculates probability of exactly x successes in n independent trials with constant success probability.

Use: Fixed number of trials, two outcomes (success/failure), constant probability, independent trials.

Variables: n = number of trials, x = number of successes, p = success probability, $q = 1 - p$ = failure probability, $\binom{n}{x}$ = "n choose x" combinations.

12. Poisson Distribution: $X \sim \text{Pois}(\mu)$ where $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$

Moments: $E(X) = \mu$, $V(X) = \mu$, $SD(X) = \sqrt{\mu}$

What it does: Calculates probability of exactly x events occurring in a fixed interval when events happen at a known average rate.

Use: Rare events, known average rate, events independent.

Variables: μ = average rate/expected count per interval, x = number of occurrences, e = Euler's number (2.718), $x!$ = factorial of x.

13. Normal Distribution: $X \sim N(\mu, \sigma)$ with Z-score: $z = \frac{x-\mu}{\sigma}$

What it does: Standardizes any normal variable to standard normal distribution.

Use: Converts to z-scores to use standard normal tables for probability calculations.

Variables: X = original normal variable, μ = mean of distribution, σ = standard deviation, z = standard score (number of std devs from mean), Result: $Z \sim N(0, 1)$.

Sampling Methods:

Random Sampling:

What: Every member has equal selection chance.

Stratified Sampling:

What: Divide population into homogeneous groups (strata), random sample from each stratum.

Cluster Sampling (1-stage):

What: Divide into clusters, randomly select some clusters, survey ALL members in selected clusters.

Cluster Sampling (2-stage):

What: Divide into clusters, randomly select clusters, then random sample within selected clusters.