Statistics Exam Cheat Sheet

Student-Ready Reference Guide

---

1. DESCRIPTIVE STATISTICS

Measures of Central Tendency

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample Mean – Use for continuous data, most common

Median (Q2) – Middle value; use for skewed distributions

- Odd n: position $\frac{n+1}{2}$
- Even n: average of positions $\frac{n}{2}$ and $\frac{n+2}{2}$

Mode – Most frequently occurring value

Measures of Dispersion

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Sample Standard Deviation – Typical distance from mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample Variance – s^2

$$\mathrm{IQR} = Q_3 - Q_1$$

Interquartile Range – Middle 50% of data

$$V = \frac{s}{\bar{x}}$$

Coefficient of Variation – Standardized spread, allows comparison across different datasets

Position Measures

Quartiles: Q_1 (25th), Q_2 (50th), Q_3 (75th), Q_4 (100th)

Five-Number Summary: Min, Q_1, Q_2, Q_3, Max

Correlation & Covariance

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right) \left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

Pearson's Correlation Coefficient
- Range: -1 to +1
- Interpretation: 0 (none), 0.2-0.4 (weak), 0.4-0.6 (moderate), 0.6-0.8 (strong), 0.8-1.0 (very strong)

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Covariance – Direction of linear relationship

$$r^2$$

Coefficient of Determination – Proportion of variance explained

---

2. PROBABILITY DISTRIBUTIONS

Discrete Distributions

Bernoulli Distribution (Single Trial)

$$f(x) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

- **Mean**: $\mu = p$
- **Variance**: $\sigma^2 = p(1-p)$

Binomial Distribution

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- **Mean**: $\mu = np$
- **Variance**: $\sigma^2 = np(1-p)$
- **Use**: Number of successes in n independent trials

Poisson Distribution

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- **Mean**: $\mu = \lambda$
- **Variance**: $\sigma^2 = \lambda$
- **Use**: Count of rare events in fixed interval

Discrete Uniform Distribution

$$f(x) = \frac{1}{m}, \quad x \in \{a, a+1, \ldots, b\}$$

- **Mean**: $\mu = \frac{a+b}{2}$
- **Variance**: $\sigma^2 = \frac{(b-a+1)^2-1}{12}$

Continuous Distributions

Normal Distribution (Gaussian)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- **Mean**: μ
- **Variance**: σ^2
- **Notation**: $X \sim N(\mu, \sigma^2)$

Standard Normal Distribution

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

- **Use Z-tables** for probabilities
- **Empirical Rule**: 68-95-99.7%

Uniform Distribution

$$f(x) = \frac{1}{b-a}, \quad x \in [a,b]$$

- **Mean**: $\mu = \frac{a+b}{2}$
- **Variance**: $\sigma^2 = \frac{(b-a)^2}{12}$

Exponential Distribution

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

- **Mean**: $\mu = \frac{1}{\lambda}$
- **Variance**: $\sigma^2 = \frac{1}{\lambda^2}$

Chi-Square Distribution

$$X^2 = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$$

- **Mean**: $E(X) = k$ (degrees of freedom)
- **Variance**: $\text{Var}(X) = 2k$
- **Use**: Goodness of fit, independence tests

t-Distribution

$$T = \frac{Z}{\sqrt{Y/\nu}} \sim t_\nu$$

- **Mean**: $\mu = 0$ (for $\nu > 1$)
- **Variance**: $\sigma^2 = \frac{\nu}{\nu-2}$ (for $\nu > 2$)
- **Use**: Inference with unknown population SD
- **Approaches normal as ν increases**

F-Distribution

$$X = \frac{U/m}{V/n} \sim F_{m,n}$$

- Where $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ independent
- **Use**: ANOVA, comparing variances

---

3. SAMPLING & SAMPLING DISTRIBUTIONS

Central Limit Theorem

For sample size $n \geq 30$ (or $n \geq 15$ for symmetric, $n \geq 50$ for skewed):

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Sampling Distribution of Sample Mean

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(Standard Error)

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Finite Population Correction

When $\frac{n}{N} > 0.05$ and sampling without replacement:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

---

4. ESTIMATION & CONFIDENCE INTERVALS

Point Estimation

- **Sample Mean** ($\bar{x}$) estimates μ
- **Sample Proportion** ($\hat{p}$) estimates p
- **Sample Variance** (s^2) estimates σ^2

Confidence Interval General Form

$$\text{Point Estimate} \pm (\text{Critical Value}) \times (\text{Standard Error})$$

CI for Population Mean

Known σ:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Unknown σ (use s):

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

CI for Population Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Conditions: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

CI for Population Variance

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2}}\right)$$

Margin of Error (MOE)

For means (unknown σ):

$$\text{MOE} = t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

For proportions:

$$\text{MOE} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Sample Size Determination

For estimating mean:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{\text{MOE}}\right)^2$$

For estimating proportion:

$$n = \frac{z_{\alpha/2}^2 \cdot p(1-p)}{(\text{MOE})^2}$$

Conservative: Use $p = 0.5$ if unknown

Common Critical Values

Confidence Level	α	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.576

---

5. HYPOTHESIS TESTING

Six-Step Process

Step 1: State Hypotheses

- H_0: Null hypothesis (status quo, equality, no effect)
- H_1 or H_a: Alternative hypothesis

Step 2: Choose Significance Level

- $\alpha = 0.05$ (standard), 0.01 (conservative), 0.10 (liberal)

Step 3: Select Test Statistic

- Mean (known σ): Z-test
- Mean (unknown σ): t-test
- Proportion: Z-test

Step 4: Determine Critical Region

- Based on α and test type (one-tailed vs two-tailed)

Step 5: Calculate Test Statistic

- Use observed sample data

Step 6: Make Decision & Conclusion

- If p-value $< \alpha$ → Reject H_0
- Otherwise → Fail to reject H_0

Hypothesis Test Types

Two-Tailed (Testing for difference):

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

- Reject if $|t_{\text{obs}}| > t_{\alpha/2}$

Right-Tailed (Testing if greater):

$$H_0 : \mu \leq \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

- Reject if $t_{\text{obs}} > t_{\alpha}$

Left-Tailed (Testing if less):

$$H_0 : \mu \geq \mu_0 \text{ vs } H_1 : \mu < \mu_0$$

- Reject if $t_{\text{obs}} < -t_{\alpha}$

Test Statistics

One-Sample t-test (Unknown σ):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad df = n - 1$$

One-Sample Z-test (Known σ):

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Proportion Z-test:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Conditions: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$

Two-Sample t-test (Welch):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$$

Pooled Two-Sample t-test:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}, \quad df = n_1 + n_2 - 2$$

P-Values

Definition: Probability of observing test statistic as extreme or more extreme than observed, assuming H_0 is true.

Decision Rule:

- If p-value $< \alpha$ → Reject H_0 (significant)
- If p-value $\geq \alpha$ → Fail to reject H_0 (not significant)

Interpretation Guide:

- p-value < 0.01: Very strong evidence against H_0
- $0.01 \leq$ p-value < 0.05: Strong evidence
- $0.05 \leq$ p-value < 0.10: Moderate evidence
- p-value ≥ 0.10: Weak/no evidence

Type I and Type II Errors

Decision	H_0 True	H_0 False
Reject H_0	**Type I Error** (α)	Correct (Power)
Fail to Reject H_0	Correct	**Type II Error** (β)

Power = $1 - \beta$ = Probability of correctly rejecting false H_0

---

6. LINEAR REGRESSION

Simple Linear Regression

Regression Equation:

$$\hat{y} = b_0 + b_1 x$$

Slope:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = r \cdot \frac{s_y}{s_x}$$

Intercept:

$$b_0 = \bar{y} - b_1\bar{x}$$

Correlation & Regression Relationship

$$b_1 = r \cdot \frac{s_y}{s_x}$$

$$r = b_1 \cdot \frac{s_x}{s_y}$$

Sum of Squares Decomposition

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

Total Sum of Squares: SST = total variability

Regression Sum of Squares: SSR = explained variability

Error Sum of Squares: SSE = unexplained variability

Coefficient of Determination

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = r^2$$

Interpretation: Proportion of variance in y explained by x

Standard Error of Estimate

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

Interpretation: Average distance data points fall from regression line

Hypothesis Test for Slope

$$H_0 : \beta_1 = 0 \text{ (no linear relationship)}$$

$$t = \frac{b_1 - 0}{SE(b_1)} = \frac{b_1}{s_e/\sqrt{\sum(x_i - \bar{x})^2}}, \quad df = n - 2$$

Reject H_0 if $|t| > t_{\alpha/2,n-2}$ or p-value $< \alpha$

Confidence Interval for Slope

$$b_1 \pm t_{\alpha/2,n-2} \cdot SE(b_1)$$

Prediction vs Confidence Intervals

Confidence Interval for Mean Response (narrower):

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Prediction Interval for Individual (wider):

$$\hat{y} \pm t_{\alpha/2,n-2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Regression Assumptions (LINE)

1. **Linearity**: Relationship is linear
2. **Independence**: Observations are independent
3. **Normality**: Residuals are normally distributed
4. **Equal Variance** (Homoscedasticity): Variance constant across x values

---

7. COMPARING MEANS

Two-Sample Tests

Independent Samples (Welch t-test):

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Confidence Interval for Difference:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Paired t-test

$$t = \frac{\bar{d} - \delta_0}{s_d/\sqrt{n}}, \quad df = n - 1$$

Where $d_i = x_{1i} - x_{2i}$

One-Way ANOVA (F-test)

ANOVA Decomposition:

$$\mathrm{SST} = \mathrm{SS(Tr)} + \mathrm{SSE}$$

Mean Squares:

$$\mathrm{MS(Tr)} = \frac{\mathrm{SS(Tr)}}{k-1}$$

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-k}$$

F-statistic:

$$F = \frac{\mathrm{MS(Tr)}}{\mathrm{MSE}} \sim F_{k-1,n-k}$$

Decision: Reject H_0 if $F > F_{\alpha,k-1,n-k}$

Post-Hoc Comparisons (Pairwise)

$$\bar{y}_i - \bar{y}_j \pm t_{\alpha/2}\sqrt{\mathrm{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

With Bonferroni correction: Use $\alpha_{\text{adj}} = \alpha/M$ where $M = k(k-1)/2$

---

8. CHI-SQUARE TESTS

Chi-Square Test for Goodness of Fit

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1}$$

- O_i = observed frequency
- E_i = expected frequency
- Reject H_0 if $\chi^2 > \chi^2_{\alpha,k-1}$

Chi-Square Test for Independence

$$\chi^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$df = (r-1)(c-1)$$

- Reject H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

Conditions: All $E_{ij} \geq 5$

---

9. INFERENCE FOR PROPORTIONS

One-Sample Proportion Test

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Conditions: $np_0 \geq 10$ and $n(1 - p_0) \geq 10$

Two-Sample Proportion Test

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}$$

Where $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ (pooled)

---

10. KEY FORMULAS SUMMARY

Concept	Formula	Use
Standard Error	$SE = \frac{s}{\sqrt{n}}$	Precision of estimate
t-test statistic	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	Test means, unknown σ
Confidence Interval	$\bar{x} \pm t_{\alpha/2} \cdot SE$	Range for population parameter
Regression slope	$b_1 = r(s_y/s_x)$	Change in y per unit x
R^2	$1 - \frac{\text{SSE}}{\text{SST}}$	Fit of regression model
Chi-square	$\sum \frac{(O-E)^2}{E}$	Categorical data tests
F-statistic	$\frac{\text{MS(Tr)}}{\text{MSE}}$	Compare group means

---

11. DECISION TREES & QUICK REFERENCE

Which Test to Use?

Comparing One Mean to Known Value:

- Known σ → Z-test
- Unknown σ → t-test

Comparing Two Means:

- Independent samples → Welch t-test (or pooled if equal variance)

- Paired data → Paired t-test

Comparing Multiple Means (3+):

- One-way ANOVA (F-test)

Categorical Data:

- Goodness of fit → χ^2 test
- Independence → χ^2 test

Relationships:

- Linear correlation/regression → Pearson r or regression

Interpreting Statistical Output

If p-value < 0.05: Result is statistically significant at 5% level

- Reject null hypothesis
- Evidence supports alternative hypothesis

If p-value ≥ 0.05: Result not statistically significant

- Fail to reject null hypothesis
- Insufficient evidence

Confidence Interval Contains 0: Not significant
Confidence Interval Excludes 0: Significant

---

12. EXAM TIPS & REMINDERS

✓ **Always check assumptions** before applying tests

✓ **Show work clearly** – Label steps, formulas, calculations

✓ **Report in context** – Use variable names, units, practical interpretation

✓ **Distinguish between** parameters (μ, σ, p) and statistics ($\bar{x}, s, \hat{p}$)

✓ **One-tailed vs two-tailed**: Affects critical values and p-value calculations

✓ **When in doubt**: Use conservative approach (larger sample, higher confidence)

✓ **Degrees of freedom**: Crucial for t and χ^2 tests

✓ **Standard error ≠ Standard deviation**: SE accounts for sample size

✓ **Correlation ≠ Causation**: Regression shows association only

✓ **Read carefully**: Identify what's being asked – point estimate, CI, test, prediction?

Last Updated: November 2025
Ready for Exam – Use strategically during study time!