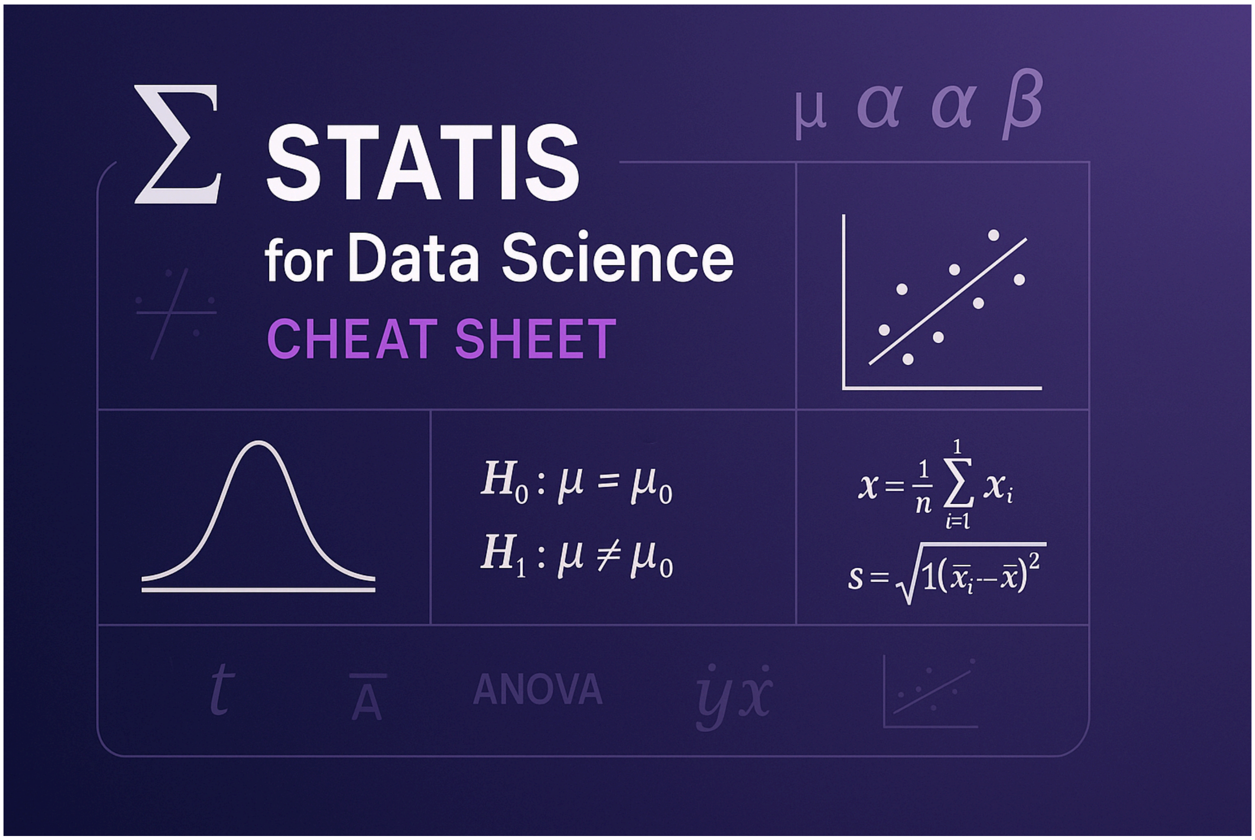
Statistics for Data Science - Comprehensive Cheat Sheet



Descriptive Statistics

Measures of Central Tendency

```
# Mean (Average)
μ = Σx / n              # Population mean
x̄ = Σx / n              # Sample mean

# Median
- Middle value when data is ordered
- Less sensitive to outliers than mean
```

```
- Better for skewed distributions

# Mode
- Most frequently occurring value
- Can have multiple modes (bimodal, multimodal)
- Useful for categorical data

# Relationship
- Normal distribution: Mean = Median = Mode
- Right skew: Mean > Median > Mode
- Left skew: Mode > Median > Mean
```

Measures of Variability

```
# Range
Range = Max - Min

# Variance
σ² = Σ(x - μ)² / N          # Population variance
s² = Σ(x - x̄)² / (n-1)      # Sample variance (Bessel's correction)

# Standard Deviation
σ = √(σ²)                   # Population std dev
s = √(s²)                   # Sample std dev

# Coefficient of Variation
CV = (σ / μ) × 100%         # Relative variability

# Interquartile Range (IQR)
IQR = Q3 - Q1               # Middle 50% of data
Outliers: < Q1 - 1.5×IQR or > Q3 + 1.5×IQR
```

Measures of Shape

```
# Skewness
- Measures asymmetry of distribution
- Skewness = 0: Symmetric
- Skewness > 0: Right-tailed (positive skew)
- Skewness < 0: Left-tailed (negative skew)

# Kurtosis
- Measures tail heaviness
- Kurtosis = 3: Normal distribution (mesokurtic)
- Kurtosis > 3: Heavy tails (leptokurtic)
- Kurtosis < 3: Light tails (platykurtic)
```

Percentiles & Quartiles

```
# Percentiles
P_k = Value below which k% of data falls

# Quartiles
Q1 = 25th percentile (First quartile)
Q2 = 50th percentile (Median)
Q3 = 75th percentile (Third quartile)

# Five-number summary
Min, Q1, Median, Q3, Max
```

Probability Distributions

Discrete Distributions

Bernoulli Distribution

```
# Single trial with two outcomes (success/failure)
P(X = 1) = p            # Probability of success
P(X = 0) = 1 - p        # Probability of failure
```

```
Mean = p
Variance = p(1-p)
```

Binomial Distribution

```
# n independent Bernoulli trials
P(X = k) = C(n,k) × p^k × (1-p)^(n-k)

Mean = np
Variance = np(1-p)
Standard Deviation = √(np(1-p))

# Use when:
- Fixed number of trials
- Each trial is independent
- Constant probability of success
- Two possible outcomes
```

Poisson Distribution

```
# Number of events in fixed interval
P(X = k) = (λ^k × e^(-λ)) / k!

Mean = λ
Variance = λ
Standard Deviation = √λ

# Use when:
- Events occur independently
- Average rate is constant
- Rare events over time/space
```

Continuous Distributions

Normal Distribution

```
# Bell-shaped, symmetric distribution
X ~ N(μ, σ²)

# Standard Normal Distribution
Z = (X - μ) / σ          # Z-score transformation
Z ~ N(0, 1)

# Properties:
- 68% within 1 standard deviation
- 95% within 2 standard deviations
- 99.7% within 3 standard deviations

# Central Limit Theorem
Sample means approach normal distribution as n increases
```

Student's t-Distribution

```
# Similar to normal but heavier tails
# Used when:
- Small sample sizes (n < 30)
- Population standard deviation unknown
- Degrees of freedom = n - 1

# As df increases, approaches normal distribution
```

Chi-Square Distribution

```
# Right-skewed distribution
# Used for:
- Goodness of fit tests
- Test of independence
- Variance testing
```

```
χ² = Σ((Observed - Expected)² / Expected)
df = (rows - 1) × (columns - 1)  # For contingency tables
```

F-Distribution

```
# Ratio of two chi-square distributions
# Used for:
- ANOVA (Analysis of Variance)
- Comparing variances
- Regression analysis

F = (s₁² / σ₁²) / (s₂² / σ₂²)
df₁ = numerator degrees of freedom
df₂ = denominator degrees of freedom
```

Sampling and Estimation

Sampling Methods

```
# Simple Random Sampling
- Every element has equal probability
- Use random number generation

# Stratified Sampling
- Divide population into strata
- Sample from each stratum
- Ensures representation

# Cluster Sampling
- Divide into clusters
- Randomly select clusters
- Sample all elements in selected clusters

# Systematic Sampling
```

```
- Select every kth element
- k = N/n (population size / sample size)
```

Central Limit Theorem

```
# For sample means:
μ_x̄ = μ                    # Mean of sample means
σ_x̄ = σ / √n               # Standard error of mean

# Conditions:
- Sample size n ≥ 30 (or population is normal)
- Samples are independent
- 10% condition: n < 0.1N (for sampling without replacement)
```

Confidence Intervals

```
# For population mean (σ known)
CI = x̄ ± z_(α/2) × (σ/√n)

# For population mean (σ unknown)
CI = x̄ ± t_(α/2,df) × (s/√n)

# For population proportion
CI = p̂ ± z_(α/2) × √(p̂(1-p̂)/n)

# Margin of Error
ME = Critical Value × Standard Error

# Common Confidence Levels:
90%: z = 1.645
95%: z = 1.96
99%: z = 2.576
```

Hypothesis Testing

Hypothesis Testing Framework

```
# Step 1: State hypotheses
H₀: Null hypothesis (status quo)
H₁: Alternative hypothesis (what we want to prove)

# Step 2: Choose significance level
α = 0.05 (common choice)

# Step 3: Calculate test statistic
# Step 4: Find p-value or critical value
# Step 5: Make decision
# Step 6: State conclusion in context
```

Types of Errors

```
# Type I Error (α)
- Reject true null hypothesis
- False positive
- P(Type I Error) = α

# Type II Error (β)
- Fail to reject false null hypothesis
- False negative
- P(Type II Error) = β

# Power of Test
Power = 1 - β
- Probability of correctly rejecting false H₀
```

One-Sample Tests

One-Sample t-Test

```
# Test population mean when σ unknown
H₀: μ = μ₀
H₁: μ ≠ μ₀ (two-tailed)
H₁: μ > μ₀ (right-tailed)
H₁: μ < μ₀ (left-tailed)

t = (x̄ - μ₀) / (s/√n)
df = n - 1

# Assumptions:
- Random sample
- Normal distribution or n ≥ 30
- Independent observations
```

One-Sample Proportion Test

```
# Test population proportion
H₀: p = p₀
H₁: p ≠ p₀

z = (p̂ - p₀) / √(p₀(1-p₀)/n)

# Assumptions:
- Random sample
- np₀ ≥ 10 and n(1-p₀) ≥ 10
- Independent observations
```

Two-Sample Tests

Two-Sample t-Test

```
# Independent samples (equal variances)
t = (x̄₁ - x̄₂) / (s_p × √(1/n₁ + 1/n₂))
```

```
s_p = √(((n₁-1)s₁² + (n₂-1)s₂²) / (n₁+n₂-2))  # Pooled std dev
df = n₁ + n₂ - 2

# Independent samples (unequal variances - Welch's t-test)
t = (x̄₁ - x̄₂) / √(s₁²/n₁ + s₂²/n₂)

# Paired samples
t = (d̄ - μ_d) / (s_d/√n)
df = n - 1
```

Two-Sample Proportion Test

```
# Compare two proportions
z = (p̂₁ - p̂₂) / √(p̂(1-p̂)(1/n₁ + 1/n₂))

p̂ = (x₁ + x₂) / (n₁ + n₂)  # Pooled proportion

# Assumptions:
- Independent samples
- Large sample sizes
```

ANOVA (Analysis of Variance)

```
# Compare means of 3+ groups
H₀: μ₁ = μ₂ = ... = μₖ
H₁: At least one mean is different

F = MSB / MSW

MSB = SSB / (k-1)          # Mean Square Between
MSW = SSW / (N-k)          # Mean Square Within

# Where:
k = number of groups
N = total sample size
```

```
# Post-hoc tests (if F significant):
- Tukey's HSD
- Bonferroni correction
- Scheffé test
```

Chi-Square Tests

Goodness of Fit Test

```
# Test if data follows expected distribution
H_0: Data follows specified distribution
H_1: Data does not follow specified distribution

χ² = Σ((O_i - E_i)² / E_i)
df = categories - 1

# Assumptions:
- Expected frequency ≥ 5 for each category
- Independent observations
```

Test of Independence

```
# Test relationship between two categorical variables
H_0: Variables are independent
H_1: Variables are dependent

χ² = Σ((O_ij - E_ij)² / E_ij)
df = (rows - 1) × (columns - 1)

E_ij = (row total × column total) / grand total

# Effect size: Cramér's V
V = √(χ² / (n × min(r-1, c-1)))
```

Correlation and Regression

Correlation Analysis

```
# Pearson Correlation Coefficient
r = Σ((x_i - x̄)(y_i - ȳ)) / √(Σ(x_i - x̄)² × Σ(y_i - ȳ)²)

# Interpretation:
r = 1:   Perfect positive correlation
r = 0:   No linear correlation
r = -1:  Perfect negative correlation

|r| ≥ 0.7: Strong correlation
0.3 ≤ |r| < 0.7: Moderate correlation
|r| < 0.3: Weak correlation

# Spearman Rank Correlation
- Non-parametric alternative
- Based on ranks, not raw values
- Detects monotonic relationships
```

Simple Linear Regression

```
# Model: y = β₀ + β₁x + ε

# Least Squares Estimates:
β₁ = Σ((x_i - x̄)(y_i - ȳ)) / Σ(x_i - x̄)²  # Slope
β₀ = ȳ - β₁x̄                              # Intercept

# Coefficient of Determination
R² = SSR / SST = 1 - SSE / SST

SST = Σ(y_i - ȳ)²    # Total Sum of Squares
SSR = Σ(ŷ_i - ȳ)²    # Regression Sum of Squares
SSE = Σ(y_i - ŷ_i)²  # Error Sum of Squares
```

```
# Standard Error of Estimate
s_e = √(SSE / (n-2))

# Assumptions:
- Linearity
- Independence
- Homoscedasticity (constant variance)
- Normality of residuals
```

Multiple Linear Regression

```
# Model: y = β₀ + β₁x₁ + β₂x₂ + ... + βₖxₖ + ε

# Adjusted R²
R²_adj = 1 - ((1-R²)(n-1)) / (n-k-1)

# F-test for overall significance
F = (R²/k) / ((1-R²)/(n-k-1))

# t-test for individual coefficients
t = β_j / SE(β_j)

# Multicollinearity detection:
- Variance Inflation Factor (VIF)
- VIF > 10 indicates multicollinearity
```

Non-Parametric Tests

When to Use Non-Parametric Tests

- Data is not normally distributed
- Ordinal data
- Small sample sizes

- Presence of outliers
- Violated assumptions of parametric tests

Common Non-Parametric Tests

```
# Mann-Whitney U Test (Wilcoxon Rank-Sum)
- Alternative to two-sample t-test
- Compares medians of two independent groups

# Wilcoxon Signed-Rank Test
- Alternative to paired t-test
- Compares medians of paired samples

# Kruskal-Wallis Test
- Alternative to one-way ANOVA
- Compares medians of 3+ independent groups

# Friedman Test
- Alternative to repeated measures ANOVA
- Compares medians of 3+ related groups

# Sign Test
- Tests median of single population
- Uses only direction of differences
```

Effect Size and Power Analysis

Effect Size Measures

```
# Cohen's d (standardized mean difference)
d = (μ₁ - μ₂) / σ_pooled

# Interpretation:
d = 0.2: Small effect
d = 0.5: Medium effect
```

```
d = 0.8: Large effect

# Eta-squared (η²) for ANOVA
η² = SSB / SST

# Cramér's V for Chi-square
V = √(χ² / (n × min(r-1, c-1)))

# R² for regression
- Proportion of variance explained
```

Power Analysis

```
# Power = P(Reject H₀ | H₀ is false)
# Power = 1 - β

# Factors affecting power:
- Effect size (larger = more power)
- Sample size (larger = more power)
- Significance level (α) (larger = more power)
- Population variance (smaller = more power)

# Sample size calculation:
n = (z_α/2 + z_β)² × σ² / (μ₁ - μ₀)²
```

Bayesian Statistics Basics

Bayes' Theorem

```
P(A|B) = P(B|A) × P(A) / P(B)

# Where:
P(A|B) = Posterior probability
P(B|A) = Likelihood
```

```
P(A) = Prior probability
P(B) = Marginal probability

# Bayesian vs Frequentist:
Frequentist: Parameters are fixed, data is random
Bayesian: Parameters are random, data is observed
```

Bayesian Inference

```
# Prior beliefs + Data → Posterior beliefs
Posterior ∝ Likelihood × Prior

# Credible Intervals
- Bayesian equivalent of confidence intervals
- Probability that parameter lies in interval

# Bayesian Hypothesis Testing
- Bayes Factor
- Posterior probability of hypotheses
```

Time Series Analysis

Components of Time Series

```
# Trend: Long-term movement
# Seasonality: Regular periodic patterns
# Cyclical: Long-term fluctuations (non-regular)
# Irregular/Random: Unpredictable fluctuations

# Decomposition Models:
Additive: Y(t) = Trend + Seasonal + Irregular
Multiplicative: Y(t) = Trend × Seasonal × Irregular
```

Time Series Tests

```
# Stationarity Tests:
- Augmented Dickey-Fuller (ADF) Test
- Phillips-Perron Test
- KPSS Test

# Autocorrelation Function (ACF)
- Measures correlation between observations at different lags

# Partial Autocorrelation Function (PACF)
- Correlation between observations k periods apart, controlling for intermediat
e observations
```

Experimental Design

Principles of Experimental Design

```
# Randomization
- Random assignment to treatments
- Controls for confounding variables

# Replication
- Multiple observations per treatment
- Increases precision and power

# Blocking
- Group similar experimental units
- Controls for known sources of variation

# Factorial Design
- Multiple factors studied simultaneously
- Can detect interactions between factors
```

A/B Testing

```
# Steps:
1. Define hypothesis and metrics
2. Determine sample size
3. Randomize users to treatments
4. Collect data
5. Analyze results
6. Draw conclusions

# Key Considerations:
- Statistical significance vs practical significance
- Multiple testing corrections
- Minimum detectable effect
- Statistical power
```

Common Statistical Mistakes

Interpretation Errors

```
# Correlation ≠ Causation
- Correlation does not imply causation
- Consider confounding variables
- Use experimental design for causal inference

# p-hacking
- Multiple testing without correction
- Cherry-picking significant results
- Use Bonferroni or FDR corrections

# Survivorship Bias
- Only analyzing successful cases
- Ignoring failures or dropouts

# Simpson's Paradox
```

```
- Trend reverses when data is aggregated
- Consider lurking variables
```

Assumption Violations

```
# Check assumptions before analysis:
- Normality (Q-Q plots, Shapiro-Wilk test)
- Independence (residual plots)
- Homoscedasticity (Levene's test)
- Linearity (scatterplots)

# Solutions:
- Data transformations
- Non-parametric alternatives
- Robust statistical methods
```

Statistical Software Commands

Python (scipy.stats)

```
import scipy.stats as stats
import numpy as np

# Descriptive statistics
np.mean(data)
np.median(data)
np.std(data, ddof=1)  # Sample std dev

# Hypothesis tests
stats.ttest_1samp(data, popmean)
stats.ttest_ind(group1, group2)
stats.chi2_contingency(contingency_table)
stats.f_oneway(group1, group2, group3)
```

```
# Distributions
stats.norm.pdf(x, loc=mu, scale=sigma)
stats.norm.cdf(x, loc=mu, scale=sigma)
stats.norm.ppf(q, loc=mu, scale=sigma)
```

R Commands

```
# Descriptive statistics
mean(data)
median(data)
sd(data)
summary(data)

# Hypothesis tests
t.test(data, mu=mu0)
t.test(group1, group2)
chisq.test(contingency_table)
aov(response ~ factor)

# Distributions
dnorm(x, mean, sd)  # Density
pnorm(x, mean, sd)  # CDF
qnorm(p, mean, sd)  # Quantile
```