CHEAT SHEETS HERO

Statistics Cheat Sheet

A quick reference guide covering fundamental statistical concepts, formulas, and distributions. This cheat sheet provides a concise overview for students, researchers, and data analysts.

Descriptive Statistics

Measures of Central Tendency

Mean	Average of all values: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Median	Middle value when data is ordered. If n is even, average of the two middle values.
Mode	Most frequent value. A dataset can have multiple modes or no mode.
Weighted Mean	Average where each data point contributes unequally: $\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Measures of Dispersion

Range	Difference between the maximum and minimum values: Range = max(x_i) - min(x_i)
Variance	Average squared difference from the mean: Sample Variance: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ Population Variance: $\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$
Standard Deviation	Square root of the variance: Sample Standard Deviation: $s = \sqrt{s^2}$ Population Standard Deviation: $\sigma = \sqrt{\sigma^2}$
Coefficient of Variation	Relative measure of dispersion: $CV = \frac{\sigma}{\mu}$ (for population), $CV = \frac{s}{\bar{x}}$ (for sample)
Interquartile Range (IQR)	The difference between the 75th percentile (Q3) and the 25th percentile (Q1): IQR = Q3 - Q1

Measures of Shape

Skewness	Measure of asymmetry of the distribution. Positive skew (right-skewed) indicates a longer tail on the right side. Negative skew (left-skewed) indicates a longer tail on the left side. $\text{Skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{n s^3}$
Kurtosis	Measure of the 'tailedness' of the distribution. High kurtosis indicates heavy tails (more outliers). Low kurtosis indicates light tails. $\text{Kurtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n s^4} - 3$

Probability

Basic Probability Concepts

Probability of an Event	$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$
Complement Rule	P(A') = 1 - P(A)
Addition Rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional Probability	$P(A\|B) = \frac{P(A \cap B)}{P(B)}$
Multiplication Rule	$P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$
Independent Events	If A and B are independent: $P(A \cap B) = P(A)P(B)$, and P(A|B) = P(A)

Discrete Probability Distributions

Bernoulli Distribution	Probability of success (p) or failure (1-p) in a single trial. $P(X=x) = p^x (1-p)^{(1-x)}$, where x = 0 or 1
Binomial Distribution	Number of successes in n independent trials. $P(X=k) = \binom{n}{k} p^k (1-p)^{(n-k)}$
Poisson Distribution	Number of events in a fixed interval of time or space. $P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$
Geometric Distribution	Number of trials until the first success. $P(X=k) = (1-p)^{(k-1)} p$

Continuous Probability Distributions

Uniform Distribution	Probability is constant over a given interval [a, b]. $f(x) = \frac{1}{b-a}$ for $a \le x \le b$
Normal Distribution	Bell-shaped curve, defined by mean (μ) and standard deviation (σ). $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Exponential Distribution	Time until an event occurs. $f(x) = \lambda e^{-\lambda x}$ for $x \ge 0$

Inferential Statistics

Confidence Intervals

General Form	Estimate \pm (Critical Value * Standard Error)
CI for Population Mean (\mu) with known \sigma	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
CI for Population Mean (\mu) with unknown \sigma	$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$
CI for Population Proportion (p)	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Hypothesis Testing

Null Hypothesis (H_0)	Statement being tested.
Alternative Hypothesis (H_1)	Statement to be supported if H_0 is rejected.
Test Statistic	Value calculated from sample data to test the hypothesis.
P-value	Probability of observing a test statistic as extreme as, or more extreme than, the one computed, assuming H_0 is true.
Significance Level (\alpha)	Probability of rejecting H_0 when it is true (Type I error).
Decision Rule	If p-value $\le \alpha$, reject H_0. Otherwise, fail to reject H_0.

Common Hypothesis Tests

Z-test	Testing population mean with known \sigma or large sample size. $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
t-test	Testing population mean with unknown \sigma and small sample size. $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Chi-Square Test	Testing association between categorical variables. $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$

Regression Analysis

Simple Linear Regression

Regression Equation	$y = \beta_0 + \beta_1 x + \epsilon$
Estimating Coefficients	$\hat{\beta_1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ $\hat{\beta_0} = \bar{y} - \hat{\beta_1} \bar{x}$
Coefficient of Determination (R^2)	Proportion of variance in dependent variable explained by the independent variable. $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
SSR, SSE, SST	Sum of Squares Regression (SSR), Sum of Squares Error (SSE), Total Sum of Squares (SST) $SST = \sum (y_i - \bar{y})^2$ $SSE = \sum (y_i - \hat{y_i})^2$ $SSR = \sum (\hat{y_i} - \bar{y})^2$

Multiple Linear Regression

Regression Equation	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$
Adjusted R^2	Adjusts R^2 for the number of predictors in the model. $R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$

Assumptions of Linear Regression

1. **Linearity:** The relationship between the independent and dependent variables is linear.
2. **Independence:** The errors are independent of each other.
3. **Homoscedasticity:** The errors have constant variance.
4. **Normality:** The errors are normally distributed.