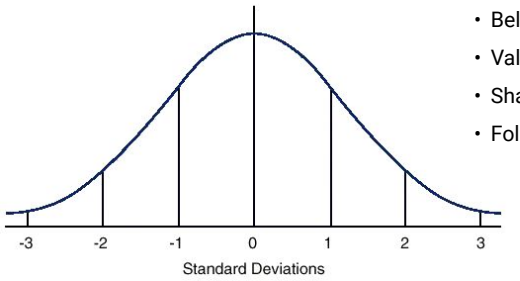
Data Analytics and Visualization Boot Camp Statistics Cheat Sheet

Selecting an Appropriate Statistical Test

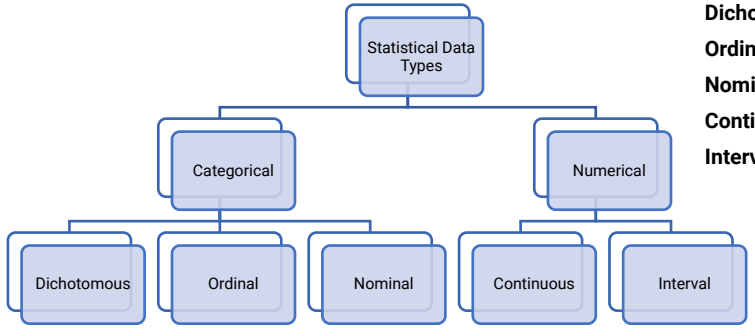
Statistical Test	Input Variable Type				Analytical Question
	Independent		Dependent		
	# of Variables	Data Type	# of Variables	Data Type	
One-Sample t-Test	1	Dichotomous (Population or Sample)	1	Continuous	Is there a statistical difference between the mean of the sample distribution and the mean of the population distribution?
Two-Sample t-Test	1	Dichotomous (Sample A vs. Sample B)	1	Continuous	Is there a statistical difference between the distribution means from two samples?
ANOVA	1+	Categorical	1	Continuous	Is there a statistical difference between the distribution means from multiple samples?
Simple Linear Regression	1	Continuous	1	Continuous	Can we predict values for a dependent variable using a linear model and values from the independent variable?
Multiple Linear Regression	2+	Continuous	1	Continuous	How much variance in the dependent variable is accounted for in a linear combination of independent variables?
Chi-Squared Test	1	Categorical	1+	Categorical	Is there a difference in categorical frequencies between groups?

What Is Normal Data?



- Bell curve distribution
- Values closer to the mean occur more frequently than values away from mean
- Shapiro-Wilk test p-value approximately greater than 0.05
- Follows the 68-95-99.7 rule
  - 68% of all data falls within 1 standard deviation from mean
  - 95.54% of all data falls within 2 standard deviations
  - 99.73% of all data falls within 3 standard deviations

Identifying Data Types



Dichotomous—one of two categories
Ordinal—ranked order, has a sequence
Nominal—labels and names
Continuous—can be subdivided infinitely
Interval—spaced out evenly on a scale

Selecting a Significance Level

Importance of Findings	Significance Level	Probability of Being Wrong
Low	0.1	1 in 10
Normal	**0.05**	**5 in 100**
High	0.01	1 in 100
Very High	0.001	1 in 1,000
Extreme	0.0001	1 in 10,000

Types of Analytical Errors

Type I

- False positive error
- Reject the null hypothesis when true
- Can be limited by making significance smaller

Type II

- False negative error
- Fail to reject the null hypothesis when false
- Can be limited by adding measurements to analysis

Equation of a Line

$$y = mx + b$$

- y: Dependent variable
- m: Slope
- x: Independent variable
- b: y intercept

Pearson's Correlation

Absolute Value of r	Strength of Correlation
$r < 0.3$	None or very weak
$0.3 \leq r < 0.5$	Weak
$0.5 \leq r < 0.7$	Moderate
$r \geq 0.7$	Strong

A/B Testing Criteria

- If the success metric is **numerical** and the **sample size is small**, use a **z-score summary statistic.**
- If the success metric is **numerical** and the **sample size is large**, use a **two-sample t-test.**
- If the success metric is **categorical**, use a **chi-squared test.**