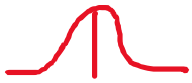
Quiz 2 Notes

Overview: Basics of EDA, probability, and statistics

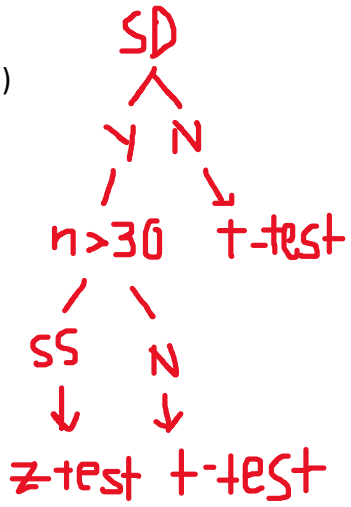
EDA 

- Patterns or trends in data to figure out which tests are most appropriate
- Numerical data EDA: scatterplot (x=predictor, y=outcome), boxplot, histograms (helpful to determine distribution of data),

Probability Distributions (helpful with which statistical tests are valid, what to expect, what data looks like):

- Bernoulli – single trial/ two possible outcomes
- Binomial – outcome of repeated Bernoulli trials
- Normal – common distribution, continuous outcome 
- Poisson – probability of certain number of events occurring over interval of time or space

Hypothesis testing



- **1)** come up with question
- **2)** turn question into hypothesis (more of quantitative research)
- Ho= null
- Ha= alternative
- **3)** collect samples/data
- **4)** EDA
- **5)** significance level
- **6)** run stat test
- **7)** examine output/ seems reasonable
- **8)** report key stats/results

Statistical Tests **(z & t tests)**

- Z Test requirements -> Standard deviation, n>30, sample size
- T test requirements -> No standard deviation, no sample size
- Ho: 2 means are not stat significantly difference
- Ha: significant difference between two means
- One sample T/Z test: comparing a sample into a predetermined mean (hypothesized mean)
- Paired t-test: comparing means of 2 samples from same population
- 2 sample tests: comparing means of 2 samples, independent samples from two diff pops.

Anova Tests

- 3 or more populations to compare means across
- One-way assumptions: 1) normality, 2) independence of samples, 3) various groups same, 4) outcome is continuous (Ho: no difference between groups and means are equal, Ha: means of groups are diff)
- Two-way assumptions: 1) outcome variable is continuous, 2) independent groups should be categorical, 3) samples are independent, 4) variance across groups the same, 5) normality

P Values

- Significance level: 0.05
- P-value: probability that result is due to random chance
- P< 5% (.05) (reject null hypothesis)
- P> 5% (.05) (fail to reject null)

Confidence Interval

- 95% of confidence intervals will cover the true population mean
- Provides a range of means
- True population mean is within