Quiz 2 Study Guide [redacted]

EDA: Exploratory Data Analysis (thinking process)

1. Generate questions about your data
2. Search for answers by visualizing, transforming, and modelling your data.
3. Use what you learned to refine questions/generate new questions (figure out focus point).
    - **Bar charts** (similar to histogram), **histogram** (use when figuring out distribution), **box plot** (use when trying to compare 2 groups of predictor variables--often categorical predictor variables--with some outcome variable), **scatter plot** (can see if there are any trends in the data)

Probability Distributions:

- Bernoulli: single trial (sample) can only end in one of two outcomes: success (p) or fail (1-p)
- Binomial: multiple bernoulli trials/samples strung together
- Geometric Distribution: number of failures you would need before a success when you are stringing together bernoulli trials
- **Normal Distribution (AKA gaussian distributions--v common):**
    - Bell curve; use when you have *continuous random variables*; with a large enough sample size, you will usually see a normal distribution
- Poisson Distribution: use when interested in events over (space and) time
    - number of events over time period
- Exponential Distribution: deal with time between events

Hypothesis Testing:

Come up with a hypothesis, then use some sort of statistical test to see whether your hypothesis holds true. More for quantitative research/testing.

1. Come up w a research question
2. turn research question into a hypothesis
    - null hypothesis — H_0
        - some baseline assumption that variables are independent (default hypothesis--not statistically significant)
    - alternative hypothesis (something you hope/suspect might be true) — H_A
        - Diff in means
3. Collect samples/data (can collect before hypothesis formation)
4. EDA
5. Determine a significance level
6. Run your statistic test
7. Examine your output/make sure output (results) seem reasonable
8. Report your key findings and statistics

Z and T Tests: common statistical tests

Purpose: You have 2 samples. you want to compare the means.

- null hypothesis states that the means of these 2 samples are NOT statistically different
- alt hypothesis states that the means of these 2 samples ARE statistically different

Z and T tests will tell you whether or not the difference is STATISTICALLY SIGNIFICANT (null or alt). But, they will not tell you HOW DIFFERENT.

Knowing the population's standard deviation and the sample size (</>30)will determine which test to use.

Terms:

- one-sample test (only need 1 sample)
    - When you're comparing a sample's mean to a predetermined mean (ex. hypothesized mean); 1 continuous outcome variable and one categorical predictor
- paired test -- comparing the means of 2 samples from the SAME population
    - Ex. before+after -- different measurement points

- two-sample test -- comparing the means of 2 samples from DIFFERENT populations
    - two continuous outcome variables (from two samples) with the same categorical predictor (the level doesn't have to be the same)

T-Test (more common)

- Assumptions of a one-sample t-test--continuous, not categorical variables; data follows normal distribution; doing random sampling of the population
- Assumptions of a paired t-test--continuous variables; the difference between the 2 samples (matched sample pair) is normally distributed; random sampling of the population
- Assumptions of a two-sample t-test--continuous variables/data; both samples are normally distributed; samples are independent (well, technically by def); variances of the 2 samples are the same (or at least very close to the same); random sampling from both populations

Z-Test

- Assumptions of a one-sample z-test--continuous data; data is approx normally distributed; random sampling; population standard deviation is known (true for all z-tests; by def)
- Assumptions of a paired z-test--continuous data/variables; difference between each pair of data points (of the 2 samples) follows a normal distribution; random sampling; population standard deviation (of the difference between each pair of data points) is known
- Assumptions of a two-sample z-test--continuous variables/data; random sampling from both populations; samples are independent; variance of the distributions of both populations is known

ANOVA: statistical test; use when you want to compare the means of 3+ samples/populations

1. One-way ANOVA — if one variable differs across the populations

one categorical predictor variable with 2+ levels and one continuous outcome variable.

Assumptions:

- normality (the samples are taken from normally distributed population s)
- independence of samples
- variance across the groups (samples or populations) is the same
- outcome variable (dependant variable) is continuous

null: there is no difference between the groups; means are equal or close enough to equal

alternative: means across the groups differ (at least one mean is different from the other means)

2. Two-way ANOVA — how two categorical factors differ across the populations AND these categorical factors affect each other

Multiple categorical predictor variables (with different levels in each) and a continuous outcome variable. Want to measure the impact of each categorical variable's levels on the outcome and *also* the potential interaction between the two categorical variables.

- Outcome variable is continuous
- groups/predictor (independant) variables should be categorical and independant
- samples are independent
- variance across groups is approx the same
- normality (samples were taken from normally distributed populations)

P-Value and Confidence Interval -- You get a p-value+ a confidence interval when you run a statistical test.

Significance Level (α) — how much tolerance we have for error due to chance

Commonly, $\alpha=0.05$, but if you want to be super super confident or exact, you can set α as a lower number/percent.

P-Value: probability that your result is due to chance

- $p<0.05$ ($p<\alpha$) → you are reasonably confident that your result is NOT due to chance (good!) → reject the null hypothesis
- $p\geq0.05$ ($p\geq\alpha$) → fail to reject the null hypothesis (no relationship between variables)

Confidence Interval: provides a range of means that you think the actual population mean is within/in-between.

$CI=1-\alpha$

- Ex. $\alpha=0.05$ → $CI=0.95$ or 95% confident
    - means that if you run the CI test many times, 95% of the CI will cover/include the actual population mean

Range of means example: (1.02, 3.5)

If interval includes 0 → no relationship between variables