$\text{final}_o utline_c of c$

Vinny Paris

November 2026

There will be three major topics tested over (t-tests, linear regression, and experimentation). With respect to linear regression I will test over the the different regressions we went over. You will need to make a prediction, calculate a residual, test the assumptions by looking at residual graphs, interpret the parameters, etc... Experimentation will be mostly identifying the parts of an experiment. The t-test will work similar to the z-test of exam 1.

If I bold something in the slides it means it's important. Expect a mix of short answer, multiple choice & true/false, and questions on regressions. There will be graphs. I would advise against memorizing the formulas (see the bottom section for my comments on the math you need to know) and focusing on the concepts.

There are useful practice problems (and solutions to some) in the textbook (link is in the syllabus) which I advise you leverage. They have both t-tests for 1 mean and linear regression. I would also advise checking which questions have solutions and focusing on working on them first.

1 t-tests

- Z-test is used when the population variance is known (rare)
- t-test is used when we have to estimate the pop. variance (by using our sample variance)
  - degrees of freedom = n-1 for a single mean
  - The t-distribution is wider than a normal distribution because we have to estimate two things instead of one (mean and var vs just the mean)
- Assumptions for the sampling distribution
  - Random: Our data was collected randomly if it's a survey or the treatments were randomly assigned for an experiment
  - Independent and Identically Distributed:
    - Independent: Knowing something about one observation doesn't tell you anything about the next observation
      - Genetically related subjects violate this
      - Measuring the same things over time violate this

        * Identically Distributed
            · No a priori belief that one observation will be different than another observation (outside of random noise)
            · Basically there is no lurking variable that might be messing up our analysis
            · Eg Testing high jump distances given different shoes and break participants into two groups....but one group is half made up of geriatrics while the other is the local college's basketball team
    - Population is normal or n is large
        * Don't know population's shape but we do know the sample's shape so use that instead
        * How large n needs to be depends on the shape of the sample distribution, see slide 8 of z-test notes
        * Large n invokes the Central Limit Theorem which says a sampling distribution will approach normality if the sample size is large

- Understand how strong a p-value is classified by
    - See slides of z-tests for table of p-value strengths
- Using a p-value be comfortable writing a decision
    - You *will* lose points if you say reject/fail to reject
    - Strength of evidence
- Define a p-value
    - Probability we'd see a sample statistic as or more extreme than what we saw, given the null hypothesis is true
- Understand the limits of a p-value
    - Doesn't give the probability the null/alt hypothesis is true
    - A large p-value doesn't mean the null is true
- Build a confidence interval and interpret it using a t-distribution
    - Same formula except $\hat{\sigma^2}$ replaces σ^2 and t^* for z^*
    - Same interpretation as from z-tests

2 Experiments

- What makes something an experiment vs an observational study (assignment of treatments!!!!!!!)
- Identify an experiment's...
  - Factor(s)
  - Levels of that/those factor(s)
  - Treatment: a combinations of all factors
  - Experimental Unit: Thing a treatment is applied to
  - Observational Unit: Thing that we measure and record in our notebooks/spreadsheets
- Understand that the number of experimental units acts as an upper limit on how much information is in the experiment
  - Imagine four cows split into two diet treatment groups
  - We can weigh the cows every hour for a month
  - Observational unit: a cow at a given time point (eg cow 2 weighed 350lbs on Nov 25th at 2pm)
  - Experimental unit: a cow (eg cow 2 is assigned diet A)
  - At the end of the day...we have 4 cows and that's all we have; not a lot of information
- Argue for or against causality in experiments. You do not need to agree with my position so long as you make a rational, well thought out argument

3 Simple Linear Regression

- What are the two main goals of simple linear regression?
- Be comfortable talking about what an explanatory variable and response variable are
- Difference between the regression equation and an estimated regression equation
- Difference between $\hat{y}$ and y; similar between $\hat{\beta}_1$ and β_1 and again $\hat{\beta}_0$ and β_0
- Use a best-fit-line to make a prediction
- Use said prediction and the actual value to calculate the residual
- Interpret your coefficients
- 3 Big Assumptions. Be comfortable checking the ones you can with a residual vs predicted plot
  - Data is randomly collected for observational studies or the treatments were randomly assigned for experiments.
  - Independent and Identically distributed residuals (not data points but the residuals of our model)

        - Any given grouping of residuals looks like another grouping of residuals
        - Homoskedasticity (constant spread)
        - X and Y's relationship is correctly identified (here linear)
        - No noticeable pattern in the y-direction
    - Normality
        - Equal spread of points above and below the line
        - More points near 0
        - No bad outliers in the y-direction

- Extrapolation and why that does not bring joy
- R^2
    - Interpretation
    - Relationship to Pearson's correlation coefficient r
    - Appreciation for the fact (and resulting limitations that) it boils down the entire scatterplot/linear regression to a single number

4 Regression with Categorical Predictors

- What's a good (best?) value for guessing a prediction for a group without outliers nor is it skewed? (hint: it's the mean)
- Indicator variables
    - What are they?
    - Why do we use them?
    - How to make them
- Make a prediction using a model with a categorical predictor, and calculate a residual
- Interpretations of it's parameters (HINT: I'll only be using the format with β's, the same as R)
    - What's a baseline category? (The default category associated with β_0, and the category the other categories are compared to when estimating β_1, β_2, etc..)

5 Odds and Ends

- A large p-value is not considered a bad thing. It suggests that the evidence against the null hypothesis is lacking.
- Know what an *error* in hypothesis testing it
    - You did NOT do something wrong
    - Your sample was unlucky

  - So you came to the wrong conclusion
  - Type 1 error: Incorrectly claim strong evidence against the null hypothesis even though the null is true
  - Type 2 error: Incorrectly claim you have little/no evidence against the null even though the null is false

- The p-value is uniformly distributed between 0-1
  - So probability of getting a p-value less than .1 (weak evidence) if the null hypothesis is true is 10%
  - Similarly .05 (moderate evidence) would have a 5% probability of randomly happening if the null is true

- Effect size vs p-value
  - An effect size is how strong a change in the response we expect if we change our explanatory variable
    * Increase it by one unit for numeric variables
    * Go from one group to another for categorical data
    * Often described as β in the regression equation
  - A p-value is a statement about detecting statsitical differences
  - Effect size and p-values are not directly related
    * A small p-value can be associated with a meaningless difference between the groups (in real life)
    * Or a large p-value could be associated with a very large $\hat{\beta}$
  - p-values do NOT talk about importance or influence, those are effect sizes
  - Be prepared to comment on whether a large p-value with a large effect is more meaningful than a small p-value with a small effect
    * There is no "correct" answer but I want to make you stretch your thinking
    * And to rationally discuss a position
- Multiple Comparisons and p-hacking
  - What is the multiple comparisons problem?
    * Running multiple tests is more likely to find a small p-value by accident/H_0 is true
    * Raises doubts about how reliable your results are
    * The ASA statement on p-values mentioned the number of hypothesis checked because it ties into multiple comparisons
    * Several ways to deal with this (you don't need to know them)
  - What is p-hacking? Why is it bad?

* Forcing your model to get small p-values
* Usually violating assumptions while you do it
* Or you ignore the multiple comparisons problem
* Fundamentally violates the point and ideas of science which is to move the football of human knowledge forward and instead is usually seen as a self-serving career move
  - Or it's because they are ignorant of the subject matter which is still worrisome
  - SIDE NOTE: Hanlon's Razor (wiki) states that "never attribute to malice that which is adequately explained by stupidity" and is a personal favorite
* Originates out of "publish or perish" mentality
* Strength of evidence in hypothesis testing is (hopefully) going to alleviate this some

6 Math You'll Need to be Comfortable With

- Calculate a test statsitic for a t-test (see formula sheet)
- Using an estimated linear model regression equation to make a prediction
- Using an estimated regression equation with a categorical predictor (ie with indicator variables) to make a prediction
- Given an actual observation and a prediction find a residual
- Going from r to R^2 and back again
  - Be careful on if we need a positive or negative square root
  - Decision is based on the direction of the graph

7 Table

Test Name	σ^2 known?	Assumptions	Relevant Distribution	Test Statistic	Conf. Interval
z-test	yes	(Random), (IID), (Large n or the pop is normal)	Normal	$\frac{\bar{x}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$	$\bar{x} \pm z_{1-\alpha/2}\sqrt{\sigma^2/n}$
t-test	no	(Random), (IID), (Large n or the pop is normal)	t-distribution w/ df = n-1	$\frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{n}}}$	$\bar{x} \pm t_{df=n-1}\sqrt{s^2/n}$