1 ***ANLY 510***
Analytics II: Principles & Applications
Instructor: Ziyuan Huang, Ph.D.

2 ***Course Information***
- Everything you need to know about the course is in the syllabus – as things change during the course the syllabus will be updated on Canvas.
- Please check the syllabus frequently for updates mentioned in class.
- If you ever have any questions please contact me, I respond to email quickly – use Canvas email or HU email and be sure to include the course and section you are in as well as how your name shows up on Canvas.
- Make sure to read the syllabus and ask if you have any questions.

3 ***Grading***
- Weekly Labs 15% (previous weeks lab is due the day before the next class at midnight)
- 3 Exams 60% (20% each – cumulative by nature)
- 3 In-Class Assignments 15% (5% each – cumulative by nature)
- Attendance 10%
- The graduate school valid grades are A, B, C, and F. Please note that any grade below a C (anything below 70% as outlined above) is an F. Grading will not take place on a "curve" and late work cannot be made up. If you feel you are getting behind at the start you can either withdrawal from the class, or in special circumstances we can discuss an incomplete: Incompletes must be made up by the start of the following term.

4 ***Expectation***
- After this class, you should be able to know:
  1. How to recognize your data
  2. How to use different methods to analysis your data
  3. How to use proper ways to explain your findings
- This is a graduate level course and I expect graduate level work. As such, poorly formatted work (e.g., work missing your name, class, assignment, or work with copy and pasting besides code) will be given an automatic F.
- Cheating and plagiarism will be given an automatic F and will be reporting to the university.

5 ***What is Analytics***

6 ***Why Analytics?***
- Analytics refers to tools that help finding hidden patterns in DATA.
- Why it is so important?
- For the past few decades, businesses generate more DATA than they are able to use or know HOW TO USE.
- The Big Data revolution.
- The three Vs: Volume, Velocity and Variety.

7 ***Experimental Design***

8 ***Why Experiments?***
- We perform experiments to get *unbiased* answers to our questions.
- Planned experiments give us control over the data we collect; this allows us to ensure the

conditions are met for simpler analyses and stronger predictions.

- Observational experiment (data collected in the wild) offer almost no control but allows for the most natural experimental space and allows us to exam things we can't replicate in the lab (e.g., moral reasons).

9 ***The Experimental Procedure***

The most important part of any experiment is the design.

Why?

With a good design you can get clear answers to your questions, a poor design can lead to uninterpretable data or wrong conclusions.

What should we do?

How to ask the right question to identify what we don't know?

10 ***The Experimental Procedure***

1 - Identify the question: What are the issues/questions you want to know/solve/recognize?

For example, is price or quality more important in valuation of a product?

2 - Map out your question space: Make sure in the study, you are able to measure/detect the variable you actually want to.

For example, how will we manipulate price and quality? What other (confounding) factors might play a role in our experiment that we cannot fully control for?

3 - Design: How do we perform the research?

We need to pick the best method for answering our question and dealing with our confounds. How will we analyze our data and what are the assumptions of those analyses.

11 ***The Experimental Procedure***

4 - Conduct the experiment: Run the experiment as planned, ensuring all data points are generated without bias (e.g., different experimenter phrasing, differences due to translation, etc.)

5 - Understand the data: Once you have data you need to ensure it meets the assumptions for the analysis you wish to employ. You may need to correct the data in some way. Then conduct your analyses: My method is to employ the simplest test of my hypothesis, then I use more sophisticated analysis to get a better understanding the effect(s).

6 - Analysis and report: Using appropriate method(s) and be able to interpret your results in a proper way

12 **Data Collection**

13 ***Before collecting your data***

Two things should be known:

1. What to measure?
2. How to measure?

14 ***Variables***

What is a variable?

To test hypothesis we need variables.

- Independent variable: A variable we think is a cause, as know as predictor
  - For example, cola and weight
- Dependent variable: A variable we think is an effect, as know as outcome
  - For example, cola and weight

15 ***Levels of Measurement***

- The relationship between what is being measured and the number that represent what is being measured.
- Categorical variable: made up of category. For example, groups, conditions, etc.
  1. Binary variable: two directions
  2. Nominal: more than two groups
  3. Ordinal variable: when categories are ordered. For example, a race competition.

16 ***Levels of Measurement***

- Continuous variable: provides score for each entity.
  1. Interval: each entity has equal intervals.
  2. Ratio: adding a zero, and the scale should be meaningful. For example, age.
  3. Discrete: take only certain numbers. For example, likert scale.

17 ***Errors and Experiment***

- What is an error?
- Systematic error vs random error.
- Correlational Research methods vs Experimental Research Methods

18 ***Randomization***

- Why randomize?
  - To keep the error minimized.
- Why does randomization work?
- Relationship between randomization and central limit theorem?

19 ***Simple Comparative Analyses***

20 ***After data collection***

We talked about designing an experiment to compare whether Price on people's valuations of a product (lets call this a measure of preference).

When coming up with this experiment we would have had a hypothesis. Let's assume that people are generally influenced by price.

21 ***Hypothesis***

- What is a hypothesis?
- How and why do we need it?
- Null Hypothesis (H0): the hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
- Alternative Hypothesis (H1): a position that states something is happening.

22 ***After data collection***

- Back to our study, what are the hypothesis?
- Null Hypothesis?

- Alternative Hypothesis: Preferences for a product will be higher when the price is lower.
- Our dependent variable is what? Is it categorical, discrete, or continuous?
- Our predictor variable is what?

23 ***Central Tendency***

- Say we had 20 participants (10 per condition) give us their preferences for the product.
- We would like to get an idea of which condition leads to greater preference, what measure of central tendency should we use?
- Mean, median and Mode?
- Mean: 54 vs 68.4, Median: 57.5 vs 74.5

24

25 ***Distributions: Normal***

- A normal distribution has a given μ and σ, so we can say a given y is pulled from a normal distribution with a mean of μ and a standard deviation of σ, or *N*(μ, σ).
- The standard normal distribution has a μ = 0 and σ = 1.
- The Central Limit Theorem indicates that for samples from any distribution (even non-normal) with a large enough size the mean of those samples will be normally distributed.

26 ***Distributions: Chi Square (χ^2)***

- The χ^2 is estimated assuming that samples are drawn from a normal distribution.
- In short, the idea is if you have a normal distribution and you were to randomly draw and square samples from it the distribution of those values would be a χ^2 distribution.
- As such, the larger the sample the more normal the distribution becomes.
- Mostly use in inferential statistics, i.e. hypothesis testing.

27 ***Distributions: Students t or t***

- We use the *t* distribution when the variance is unknown and our sample size (*n*) is small.
- The *t* distribution is ~ normal with enough samples.
- When the number of samples is small the *t* distribution has a flatter center and thicker tails.

28 **Distributions: Snedecor's *F* or just *F***

- We use the F distribution when examining if two + conditions differ.
- Like other distributions the *F* is normal when sample sizes are sufficiently large.
- When the number of samples is small the *F* distribution will have increased positive skew).

29 ***T-tests***

30 ***Let's Do Some Data Science***

- We had a hypothesis : Preferences will be greater for a product of lower price than of higher price.
  - We can state this quantitatively as:
- H_0 = There will be no difference in preference (Null Hypothesis - $\bar{y}_1 = \bar{y}_2$)
- H_1 = Preferences for the products will not be equal - $\bar{y}_1 \neq \bar{y}_2$: as a one tail hypothesis - $\bar{y}_1 > \bar{y}_2$).

31 ***Let's Do Some Data Science***

- We need to decide our test statistic (which would work well here?) and a critical value (region) for the test (p < .05, .01, .001, etc.).
- There are two types of error we try to minimalize:
  1) Falsely rejecting the null when its true (Type I Error = α)
  2) Falsely rejecting the alternative hypothesis when it is true (Type II Error = β)
- To calculate Power to detect an effect we take 1 – β; many journals ask for .90 or higher.

32 ***The Null***

There are a few thing to bear in mind with a null result:

1) It could be that you have a great deal of error/noise (maybe people believe that lower price means lower quality.)
2) It could be you got unlucky and got a Type II error.
3) The effect does not exist.

33 ***T-Test: Assumptions***

1. Normality: sampling distribution is normally distributed
2. Type of variable: data are measured at least at the interval level.
3. Equal Variance: whether the variances are equal between the two groups
4. Independence: Scores in different conditions are independent.

34 ***Test It: Two Sample t-Test***

35 ***Test It: Two Sample t-Test unequal variances***

- The standard t-test assumed that the variances in each condition are ~ equal.
- Our values were fairly close: 22.96 and 28.66
- However, if we estimated them to be different (visually inspect and use Levine's) we need to change up our t-test a bit.

36 ***Test It: One Sample t-Test***

- In some instances we might want to compare some sample mean to a predicted mean (e.g., we are only interested if a treatment leads to an x amount increase).
- We could test for this using a fairly simple t score:

37 ***Test It: Paired Samples t***

- A clean design that helps limit external factors is a paired samples (within subject) design.

38 ***Confidence Levels***

Confidence intervals give us insight into how precise our estimates appear to be and the possible values we might expect an estimate or parameter to take.

$$\overline{y_1} - \overline{y_2} - (Critical\ t)S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < true\ mean\ difference\ falls\ between$$

$$> \overline{y_1} - \overline{y_2} + (Critical\ t)S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$54 - 68.4 - (2.1)11.61 < true\ mean\ difference\ falls\ between > 54 - 68.4 + (2.1)11.61$$

$$-38.79 < true\ mean\ difference\ falls\ between > 9.99$$

Or $M_{difference}$= 14.4, $CI_{95\%}$[-38.79, 9.99]

39 ***Applications***

40 ***One Sample t-test***

- Suppose we are interested in whether a new manufacturing technique takes more/less time than our current system in which the mean manufacturing time is 43.4 minutes.
- Our data would be a single column with observed times in the rows, like this:
- To perform a one sample t-test in R we would enter the command:

ttest(Time, mu = 43.4, alternative= "two.sided" "greater" or "less")

t = 1.4452, df = 9, p-value = 0.1823
t = 1.4452, df = 9, p-value = 0.09115
t = 1.4452, df = 9, p-value = 0.9088

41 ***Analysis***

- Assumption:

D'Agostino skewness test: Null -> the data does not have skewness
agostino.test
Shapiro-Wilk normality test: Null -> the data is normally distributed
shapiro.test

- T-test:

t.test(data, mu = mean)

42 ***In Class Practice***

- We know the mean is 35 and we would like to know if there is any different between the new machine.
-

43 ***Independent t-test***

- Suppose we have data on cholesterol levels for males and females and we wish to see if one sex has higher/lower LDL cholesterol (i.e., the bad kind) than the other.
- Our data would look something like this:

t.test(Males, Females, var.equal = TRUE/FALSE, *alternative*)

t = 0.75715, df = 18, p-value = 0.4588
t = 0.75715, df = 12.995, p-value = 0.4625

44 ***In Class Practice***

- Male: 169, 175, 172, 122, 135, 149, 177, 99, 103, 123, 256
- Female: 123, 105, 122, 136, 136, 156, 142, 109 ,123, 151, 107
- Is there any difference in terms of cholesterol level between male and female?

45 ***Paired Samples t-test***

- Suppose we are interested in whether a drug reduces the level of bad cholesterol significantly.
- We measure cholesterol before and after treatment.
- Our data would look something like this:

t.test(Before, After, paired = TRUE, alternative = 'two.sided')

46 ***Summary Write Up***

In the current study, the researcher examined the difference of cholesterol between gender. Performing an independent t-test (equal variances assumed) we find there is no significant difference between male (*M* = 145.5; *SD* = 34.96) and female (*M* = 136.2; *SD* = 16.92), *t*(18) = 0.76, *p* = . 45.

47 ***Categorical Variables***

48 ***Simple Tests For Categorical Dependent Variables***

In simple cases when dealing with a categorical dependent variable we use Chi Square (χ^2) analyses. There are two main types of Chi-Square test used in quantitative decision making.

1) Chi Square Goodness of Fit –when we wish to compare an observed frequency to an expected one.
2) Chi Square Test of Independence – when we wish to see if two groups differ in their observed frequencies across a categorical dependent variable.

49 ***Chi Square Goodness of Fit***

Suppose we know that the percentage of females in the population is 51% and we want to see if this percentage is also present in our class. In other words we wish to see if ~51% of our class are female as would be expected based on their proportion of the population.

We have XX females in our class and XX males so we can see that we are 51% female, but is this significant?

We first want to create an observed variable that has XX in the row for females and XX in the row for males. Then we create a variable with the expected percentages of males and females. Our data this would look something like:

In R chisq.test(Observed, p = Expected)

X-squared = XXX, df = 1, p-value = XX

50 ***Chi Test of Independence***

Imagine we have run a marketing study where we are interested in whether one of our two conditions increases the likelihood that a person buys a product.

Suppose this is the data we get:

We have to play with the data a bit, and since its fairly simple data we can just enter it:

AD1 <- c(25, 25) # Creating AD 1 Row

AD2 <- c(35, 15) # Creating AD 2 Row

Data <- as.data.frame(rbind(AD1, AD2)) # Creating Data Frame

names(Data) = c("Bought", "No") # Labeling Columns

chisq.test(Data) # Performing Test

X-squared = 3.375, df = 1, p-value = 0.06619

51 ***In Class Practice***

- Kidscalories – children either don't help (2) or help (1) making dinner and the amount of food

they consume is measured. Find which method gets them eating more.

- Cholestoraldata – cholesterol levels for the same individuals following a month of using two different brands of margarine. Determine if one lowers cholesterol more than another.
- PrioritiesData – students school priorities based on their location. Determine if there are any differences in priorities by location.
- VotingData – the order candidates from different parties are listed on the ballot and what candidate is selected. Determine if listing order impacts voting