Stats II
Initial Assessment
January 13, 2026
Barboza-Salerno

Instructions:

This material covers analyses typically taught in a Statistics I course. To establish a baseline understanding before we begin Statistics II, I would like you to complete a brief assessment.

Details:

- **Date:** Tuesday, January 13
- **Submission:** Upload your completed assessment by the end of the class period on the same day. I am giving an hour grace period just in case there are any technical issues.
- **Purpose:** This is not a graded assignment. It is intended to gauge your starting point. I am not concerned about whether you complete every question or provide exact answers.
- **Guidelines:**
    - Spend the allotted class time on the assessment. No extra time will be provided, as the goal is to observe your approach within a time constraint.
    - If you encounter a question you do not understand, simply move on to the next one. If you have a question you are free to use any resource you feel can help you answer the question, or you can text me, but please do not consult each other. I would

Focus:

The purpose of this exercise is to evaluate your problem-solving process rather than your ability to derive precise answers. The key skills we will develop in this class are:

1. Identifying the appropriate statistical method to address a research question.
2. Interpreting and extending your findings to inform policy analysis.

Please approach the assessment with an open mind and a focus on process rather than perfection. In case you have difficulty answering any of these questions, you can rest assured that we will review everything throughout the semester as we cover more advanced topics.

I am happy to review the questions with you and will provide solutions after the assessment is completed.

PART I: THEORETICAL

1. (*Introduction to data*). The table below describes an intervention to reduce PTSD among violence-exposed youth. It is important to understand the difference between row and column percentages, and when to use each. Using the table below:
    a. Among all youth in a treatment group, what proportion had PTSD by the end of the first year?
    b. What proportion of youth had PTSD in the control group by the end of the first year?

	0-30 days		**0-365 days**	
	PTSD	No PTSD	PTSD	No PTSD
Treatment	33	191	45	179
Control	13	214	28	199

2. (*Relationships between variables*): The independence assumption is key to all statistical analyses you covered in Stats I. This is because it is a key assumption of the statistics used, such as t-tests, etc. **True or False**: A pair of variables are either related or not (independent). Two variables can't be both associated and independent.

3. (*IQR*). Oftentimes, I use the interquartile range to describe meaningful changes in my regression models. I do so because a change in the IQR is more meaningful and can provide a means of standardizing variables. Therefore, knowing what an IQR is and how it's calculated is important. Answer these questions relevant to the IQR:
    a. Assume that Q1 describes the 25^{th} percentile. Define Q1 in words.
    b. Assume Q3 is the 75^{th} percentile. Q3 = 300; Q1 = 100. Calculate the IQR.
    c. Describe why extreme observations affect the standard deviation more than the interquartile range (IQR).

4. Although not obvious, contingency tables rely heavily on probability theory. This is one reason why probability theory is covered in Stats I. Use the table below to answer the following questions:
    a. What does the .458 represent in the table?
    b. What does the .139 at the intersection of **No** and **Big** represent?

	None	**Small**	**Big**	**Total**
Yes	149/367=.406	168/367=.458	50/367=.136	1
No	400/3554=.113	2657/3554=.748	495/3554=.139	1

5. (*Probability Distributions*): The table below suggests three salary distributions for social workers in Columbus. Only one is correct. Which one must it be? What is wrong with the other two?

Income range ($1000s)	**0-25**	**25-50**	**50-100**	**100+**
A	.18	.39	.33	.16
B	.38	-.27	.52	.37
C	.28	.27	.29	.16

6. (*Independent Random Variables*): About 9% of people who are exposed to a natural disaster develop symptoms of PTSD. Suppose 2 people are selected at random from the US population. Assume that these 2 people are independent.
    a. What is the probability they both developed symptoms?
    b. What is the probability that neither did?
    c. Explain what 'independence' means and why your calculation depends on this assumption.

7. Use the data from above. Now, suppose 5 people are selected at random.
    a. What is the probability they all develop PTSD symptoms?
    b. What is the probability none did?

8. Find the probability a randomly selected person who was not given the COVID-19 vaccine died from COVID using the following table showing joint probabilities.

	Vaccinated	
	Yes	No
Lived	.0382	.8252
Died	.0010	.1356

9. A tree diagram is a helpful way of visualizing data and can be a valuable tool to organize outcomes and probabilities around various data structures. The COVID-19 vaccination data above fits this description. The population can first be split by 'vaccination status', meaning whether or not they were vaccinated.
    a. Use this as a starting place and create a tree diagram for the data in the table.
    b. Use the tree diagram to calculate the probability that a random person was vaccinated and lived.
    c. Use the tree diagram to calculate the probability that the person died (regardless of whether they were vaccinated)

10. It is important to be comfortable using notation to some extent and to digest data presented in multiple ways. Compute the variable's mean, variance, and standard deviation below.

xi	P(X = xi)	xi · P(X = xi)
0	0.20	0.00
137	0.55	75.35
170	0.25	42.50

Hint: first, compute the mean of the above distribution. Then, use the mean you calculated to compute the variance. From there, the standard deviation is easy.

11. (*Distributions of Random Variables*). Let *X* represent a random variable from $N(\mu = 3, \sigma = 2)$. In other words, the notation means the distribution is normally distributed (N), with a mean of 3 and a standard deviation of 2. Suppose we observe that x = 5.19.
    a. Find the Z-score of x
    b. Use the Z-score to determine how many standard deviations above or below the mean, x, falls.

12. The social work licensing exam follows a nearly normal distribution with mean 92.6 and standard deviation 3.6.
    a. Compute the Z-score for an exam score of 95.4.
    b. Compute the Z-score for an exam of 85.8.
    c. Which of the two observations is more unusual?
    d. Both students come to you to calculate and interpret their percentile on the exam.

e. What proportion of test takers scored higher than these two?

13. The SAT test is another standardized test that is approximately normally distributed. The mean is 1500 and the standard deviation is 300, i.e., $(\mu = 1500, \sigma = 300)$.
    a. What is the probability of scoring at least 1630 on the SAT?
    b. One year something unusual happened and more than 75% of test takers scored above 1600. How does that affect the calculation.
    c. From a social justice standpoint, describe the problem with interpreting aptitude using a statistical analysis.

14. (Binomial theorem): It is essential for you to understand key distributions used in social work research. The normal distribution is the most often used distribution. There are others, such as the binomial distribution defined by 'success' or 'failure.' Use the binomial theorem to answer the following question. On 70% of days, a therapist's office admits at least one new client. On 30% of the days, no clients are admitted.
    a. What is the probability that the therapist will admit a new client on exactly three days this week?
15. (*Hypothesis testing*): As you know, I am not only a statistician, I am also an attorney. I have noticed many parallels between the two seemingly unrelated disciplines, law and statistics. One is the critical use of evidence. Another is the basic hypothesis testing framework used by both. Consider the following: A US court is weighing two possible claims about a defendant: she is not guilty or guilty.
    a. How can a hypothesis testing framework be used to evaluate these claims?
    b. What is the problem with falsely using the term 'innocent' rather than 'not guilty' in this framework?
16. (*Confidence Intervals*): In a sample of 100 students from the 2022 National Child Health Survey, the average number of days per year a child was reported ill was 2.78, with a standard deviation of 2.56 days.
    a. Compute a 95% confidence interval for the average for all children from the NCHS. Assume the conditions for normality are met.
    b. Interpret the confidence interval

17. UCLA estimates the cost of in-state tuition is 23,000$/year. I estimate it at 33,000$/year. What are the null and alternative hypotheses to test whether this claim is accurate?

18. What is the substantive difference between a p-value and the significance level of a test?

19. If the null hypothesis is true, how often should the p-value be less than .10?

20. Describe the difference between practical and statistical significance. Provide an example.

21. (*t-test*): Set up and implement a hypothesis test to determine whether, on average, there is a difference in outcomes for the treatment and control group using a dependent sample t-test. The summary statistics for the paired differences are as follows: N=73; mean difference = 12.76, standard deviation of the difference = 14.26.

22. (*independent samples t-test*): The summary statistics below show newborn weights for mothers who smoked and did not smoke during pregnancy.

	Smoked	**Did not smoke**
Mean	6.78	7.18
St. Dev.	1.43	1.6
Sample Size (N)	50	100

    a. What is the estimate of the population difference?
    b. Compute the standard error of the estimate from (a)

23. (**ANOVA**): College departments offer many lectures of the same course. Consider that CSW offers three lectures of an intro stats course to PhD students. We want to determine if there is a statistically significant difference in first exam scores across three classes (A, B, and C).
    a. Describe the appropriate hypotheses to determine whether there are any differences between the classes.
    b. An ANOVA (Analysis of Variance) was conducted for the data, and the summary results are in the table below.

	Df	**Sum Sq**	**Mean Sq**	**F-Value**	**p-value**
Class	2	1290.11	645.06	3.48	.033
Residuals	161	29810.13	185.16		
				Spooled=13.61 on df = 161	

Is the difference significant at the .05 level. Why or why not?

PART II: Applied

This part of the assessment examines your proficiency with downloading, cleaning, recoding, and performing statistical analysis for applied work. We are keeping this very simple, as these procedures are typically extremely time-consuming. I request you to download the Census Bureau's National Child Health Survey (NSCH) for 2023. You can download the data here: NSCH Datasets. Click on the link for topical data and input files (use whichever format you are more comfortable with- either SAS or STATA – it does not matter. Feel free to use any software package you like to answer the following questions. One nice thing about this survey is the online codebook located here: NSCH Codebook. This allows you to browse for variables of interest easily.

One of the papers I want to write about regards traumatic brain injury in children. Since we are focused on social justice, I am curious about differences across race and ethnicity. I also have a research focus on childhood bullying. Let's begin to examine these important issues by asking the following questions:

a. What is the prevalence of concussion or brain injury in children? Use an appropriate test statistic and draw a conclusion based on your assessment. *Note*: in 2023, there are two variables; do not use the confirmed measure.
b. Are there racial differences in concussion or brain injury? *Note:* use the race variable. There is no need to recode race/ethnicity, but if you do not understand the difference between race and ethnicity, please ask.
c. Are children with a concussion or brain injury more likely to bully others? *Note*: Bullying is measured using 7 categories, which is in accordance with the definition of bullying in the literature. For this question, just focus on comparing 'not a bully' to any bullying, i.e., compare category 1 with 2-7.
d. Choose one question above and provide a compelling visualization OR interpretation of the data, along with 1-2 sentences describing the implications for your findings.