STAT 350 Exam 2 Fall 2023

V1

Name: _____	PUID _____

Instructor (circle one): **Anand Dixit** **Timothy Reese** **Halin Shin** **Khurshid Alam**

Class Start Time: ○ 11:30 AM ○ 12:30 PM ○ 1:30 PM ○ 2:30 PM ○ 3:30 PM ○ 4:30 PM ○ Online

As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue.

Instructions:

1. **IMPORTANT** Please write your **name** and **PUID clearly** on every **odd page.**
2. **Write your work in the box. Do not run over into the next question space.**
3. You are expected to uphold the honor code of Purdue University. It is your responsibility to keep your work covered at all times. Anyone caught cheating on the exam will automatically fail the course and will be reported to the Office of the Dean of Students.
4. It is strictly prohibited to smuggle this exam outside. Your exam will be returned to you on Gradescope after it is graded.
5. The only materials that you are allowed during the exam are your **scientific calculator**, **writing utensils, erasers, your crib sheet**, and **your picture ID**. Colored scratch paper will be provided if you need more room for your answers. Please write your name at the top of that paper also.
6. The crib sheet can be a handwritten or type double-sided 8.5in x 11in sheet.
7. Keep your bag closed and cellphone stored away securely at all times during the exam.
8. If you share your calculator or have a cell phone at your desk, you will get a **zero** on the exam.
9. The exam is only 60 minutes long so there will be no breaks (including bathroom breaks) during the exam. If you leave the exam room, you must turn in your exam, and you will not be allowed to come back.
10. You must show **ALL** your work to obtain full credit. An answer without showing any work may result in **zero** credit. If your work is not readable, it will be marked wrong. Remember that work has to be shown for all numbers that are not provided in the problem or no credit will be given for them. All explanations must be in complete English sentences to receive full credit.
11. All numeric answers should have **four decimal places** unless stated otherwise.
12. After you complete the exam, please turn in your exam as well as your table and any scrap paper that you used. Please be prepared to **show your Purdue picture ID**. You will need to **sign a sheet** indicating that you have turned in your exam.

Your exam is not valid without your signature below. This means that it won't be graded.

I attest here that I have read and followed the instructions above honestly while taking this exam and that the work submitted is my own, produced without assistance from books, other people (including other students in this class), notes other than my own crib sheet(s), or other aids. In addition, I agree that if I tell any other student in this class anything about the exam BEFORE they take it, I (and the student that I communicate the information to) will fail the course and be reported to the Office of the Dean of Students for Academic Dishonesty.

Signature of Student: _____

You may use this page as scratch paper.
The following is for your benefit only; we will not use this for grading:

Question Number	Total Possible	Your points
Problem 1 (True/False) (2 points each)	12	
Problem 2 (Multiple Choice) (3 points each)	9	
Problem 3	10	
Problem 4	22	
Problem 5	26	
Problem 6	21	
Extra Credit	5	
Total	105	

1. (12 points, 2 points each) True/False Questions. Please indicate the correct answer by filling in the circle. *<u>If you indicate the correct answer by any other way, you may receive 0 points for the question</u>*.

1.1. A one sample test statistic $\mathbf{T_{TS}} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ defines a procedure for assessing the consistency of the data with that of the null hypothesis and we evaluate this evidence using a ***p*-value**.

Ⓣ or Ⓕ The ***p*-value** associated with the test statistic $\mathbf{T_{TS}} = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ can also be considered a random variable.

1.2. In two distinct studies, two researchers obtained a sample of size $\boldsymbol{n = 11}$ from their respective populations both known to be normally distributed. In study 1, the researcher has knowledge of the population standard deviation, while in study 2, the researcher lacks knowledge of the population standard deviation, and it must be estimated. Both studies aim to construct a 98% confidence interval.

Ⓣ or Ⓕ The critical value t^* used to construct the confidence interval in study 2 will be larger than that of the critical value z^* used in study 1.

1.3. In a survey to ascertain the favored presidential candidate among eligible U.S. voters, the nation's electoral college system plays a pivotal role, reflecting the preferences of individual states. As part of this study, 500 random registered voters from each state are polled about their voting choices.

Ⓣ or Ⓕ The sampling design employed in this survey is **stratified random sampling**.

1.4. In a consumer study evaluating the **taste preferences** of different coffee blends, researchers are investigating the influence of various factors, including **coffee roast level (light, medium, dark), brewing method (drip, French press, espresso)**, and the **coffee beans' country of origin (Colombia, Ethiopia, Brazil)**. One factor that is beyond the researchers' control is the **participants' experience with coffee (occasional coffee drinker, daily coffee drinker, coffee enthusiasts)**. This diversity in coffee experience could introduce variability into the study. To address this potential source of variability, the researcher decides to conduct a randomized block design experiment.

Ⓣ or Ⓕ In this scenario, the blocks of the randomized block design would include all combinations of coffee roast level, brewing method, coffee beans' country of origin and participant's experience level with coffee.

1.5. The power associated with a statistical hypothesis test is stated to be **95%**.

Ⓣ or Ⓕ This indicates that the test has a **95%** sensitivity to detect the specific effect in the study when that effect is present.

1.6. A clinical trial was conducted which obtained statistical significance. The effect size was measured to be $\mathbf{0.19}$.

Ⓣ or Ⓕ Since the effect size was measured to be small, we would conclude the results to not be practically significant, regardless of the potential impacts of the study.

2. (9 points, 3 points each) Multiple Choice Questions. Please indicate the correct answer by filling in the circle. *If you indicate the correct answer by any other way, you may receive 0 points for the question*. For each question, there is only one correct option given.

2.1. Which of the following statements is accurate regarding hypothesis tests?

Ⓐ Type I Error can be considered as the rejection of the null hypothesis when the alternative hypothesis is true.

Ⓑ Power can be considered as the probability of rejecting the null hypothesis in favor of the alternative when the alternative is false.

Ⓒ When the probability of Type II Error decreases, power increases.

Ⓓ Type II Error can be considered as the rejection of the null hypothesis when the null hypothesis is true.

2.2. In a pharmaceutical study to develop a new pain relief medication, researchers investigate two **factors**: **dosage**, and **administration method**. Dosage has three levels: **low (50 mg)**, **medium (100 mg)**, and **high (150 mg) dosages** and the administration method also has three levels: **oral tablets**, **injectable solution**, and **transdermal patch**. The **response variable** is the **pain relief score (1 to 10)**. How many **treatment groups** result from the combinations of these factors?

Ⓐ **2**

Ⓑ **3**

Ⓒ **6**

Ⓓ **9**

Ⓔ **12**

2.3. In a statistical study, which of the following statements correctly describes the relationship between confidence interval, margin of error, sample size, and confidence level?

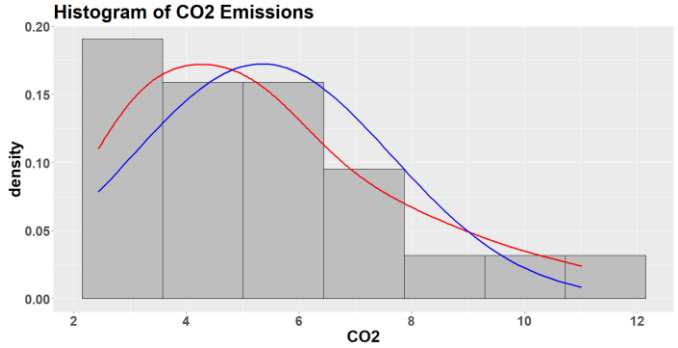
Ⓐ While keeping the confidence level constant, increasing the sample size decrease the margin of error, which narrows the confidence interval.

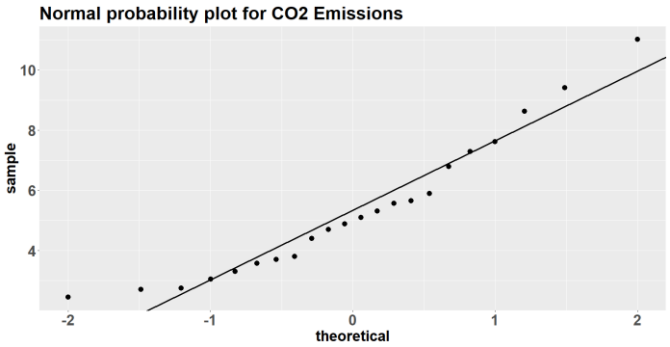
Ⓑ Increasing the sample size increases the confidence level, which narrows the confidence interval and reduces the margin of error.

Ⓒ While keeping the confidence level constant, increasing the sample size increases the margin of error, which widens the confidence interval.

Ⓓ Increasing the sample size decreases the confidence level, which widens the confidence interval and reduces the margin of error.

3. (10 points) The local environmental task force is conducting a study to estimate the average CO2 emissions in hectograms per mile (hectogram/mi) of vehicles in a city. A simple random sample of **22 vehicles** was taken, and the data is collected. The task force is interested in making inferences about the population mean CO2 emissions. The following graphs were generated from the above sample.





(a) (6 points) Using the above provided figures, what specific assumption(s) about the data can be tested to ensure the validity of performing statistical inference regarding the population mean CO2 emission of vehicles in a city? Please explain your answer and assess whether these assumptions are met based on the provided graphs. Be clear about what information is conveyed in each graph.

(b) (4 points) If the assumptions mentioned in **(a)** are not met, how could you adjust the data, to potentially satisfy these assumptions? If your answer to **(a)** is affirmative regarding the assumptions, specify any additional assumptions required for the analysis that cannot be determined from the graphs alone and explain how you would verify their validity.

4. (22 points) A company exclusively employs 22-inch cube-shaped boxes for its shipping requirements. To negotiate a favorable shipping rate, it is crucial for the company to ascertain the average weight of the filled boxes. There was a prevailing belief that the mean weight of these boxes was approximately 50 pounds. However, the executive responsible for negotiating the shipping rates held the opinion that the mean weight was greater than 50 pounds. To further investigate this matter, the executive selected a **sample** of **45 boxes**. The **sample** yielded a **mean weight** of **50.0381** pounds and a **sample standard deviation** of **0.1083** pounds. The results of various statistical tests are provided in the R output on the last page.

(a) (11 points) Does the data offer any evidence supporting the claim made by the executive? Using the above information and the information contained on the last page of the exam perform a four-step hypothesis test with a significance level $\boldsymbol{\alpha = 0.09}$. Clearly specify which output from the last page of the exam was used to obtain to obtain your conclusions.

i) Output number used for hypothesis test:

ii) Step 1:

iii) Step 2:

iv) Step 3:

v) Step 4:

(b) (7 points) Calculate an appropriate **confidence interval** or **bound** that corresponds to the hypothesis test conducted in part **(a)**. State if you are using a **confidence interval** or **bound** and be sure to mention the **confidence level** used for this **interval** or **bound**. Clearly specify which output from the last page of the exam was used to obtain to obtain your conclusions.

(c) (4 points) Interpret the result obtained in part **(b)** within the context of the problem.

5. (26 points) A group of dietitians conducted a study to investigate the impact of consuming pasta with and without a source of protein on the post-meal rise in blood sugar levels. They collected an SRS of 50 participants and randomly assigned each to either group 1 or group 2, until one group reached the size of 25. The remaining participants were assigned to the other group.

During the experiment, all the participants were on an identical, strictly controlled diet for one week. At the end of the week, Group 1 was given a meal consisting of pasta with an additional source of protein, and group 2 was given the same pasta dish without any additional source of protein. Blood sugar levels were measured for each participant both before and after the meal, and the differences ("after" minus "before" measured in milligrams per deciliter, mg/dL) were recorded. The summary statistics are shown in the table below.

	Group 1	**Group 2**	**Group1 – group2**
$\boldsymbol{n}$	**25**	**25**	**25**
$\boldsymbol{\bar{x}}$	**86.98**	**89.21**	**-2.23**
s	**3.88**	**3.39**	**4.74**

(a) (4 points) Considering the study's objectives and the method of data collection, which statistical methodology, between two-sample independent and two-sample paired, would be more suitable for analysis? Provide your reasoning.

(b) (3 points) The researchers would like to know whether eating pasta with a source of protein reduces the rise in blood sugar level compared to eating pasta alone. They have set a significance level of $\boldsymbol{\alpha = 0.05}$. State the null and the alternative hypothesis compatible with their question. Whether you have chosen two-sample independent or two sample-paired use the same order of subtraction as in the table **(Group1-Group2)**. Clearly define the parameters used in the hypothesis.

(c) (6 points) No explanation is required for this part. Assuming that all the assumptions are met and that the test statistic has been computed correctly and stored as **test_statistic**, and the degrees of freedom are correctly computed for both the two-independent and paired procedures.

1. Select the most appropriate code for computing the ***p*-value**.

Ⓐ `pt(test_statistic, df = 24, lower.tail = TRUE) = 0.01385`

Ⓑ `2*pt(test_statistic, df = 47.173, lower.tail = TRUE) = 0.036`

Ⓒ `2*pt(test_statistic, df = 24, lower.tail = TRUE) = 0.0277`

Ⓓ `pt(test_statistic, df = 47.173, lower.tail = TRUE) = 0.018`

2. Select the correct decision based on your choice of ***p*-value**.

Ⓐ Reject H_0

Ⓑ Accept H_a

Ⓒ Fail to reject H_0

Ⓓ Accept H_a

(d) (5 points) At **95% confidence**, what is the choice that is compatible with the hypotheses given in part (b): confidence interval, lower confidence bound, or upper confidence bound? **Explain your choice**.

(e) (8 points) How does zero relate to your selection in part (c)? In other words, if you opted for a confidence bound, is zero positioned above or below the bound? If you selected a confidence interval, is zero situated within or outside the interval? Clearly explain how you were able to determine this without explicitly computing the confidence bound or interval.

6. (21 points) In a semiconductor manufacturing company, the production process for microchips relies on a critical component that plays a vital role in the chips' performance. Even small deviations in these dimensions can lead to substantial losses. The company has a well-established process for producing these components and has been measuring the standard deviation of this critical dimension for years, obtaining a very precise estimate of $\boldsymbol{\sigma = 2}$ **nm (nano meter)**. It is also known that the critical dimension of these components follows a normal distribution.

The company is considering implementing an improved lithography technique in the manufacturing process that is expected to make the components even more precise, leading to a smaller critical dimension. However, implementing this change is extremely costly, and there is a lot at stake. They need to be sure that the improvement will result in a smaller critical dimension before investing in the new lithography technique.

The current critical dimension is measured to be on average **22 nm** and for the new technique to be worth the investment they need to have an average reduction of **2 nm** to justify the investment in the new lithography technique to their stake holders. A study is to be conducted in which the **Type I error rate** is to be controlled at **a significance level** $\boldsymbol{\alpha = 0.001}$. Additionally, the study requires a high degree of power to detect a reduction of **2nm** or more. The team working on the improved lithography technique is granted enough resources to manufacture **35 components** for testing this hypothesis. Is a sample size of $\boldsymbol{n = 35}$ enough to identify a **reduction** of **2nm** at a high degree of **power**?

(a) (3 points) Clearly state the parameter of interest and define the null and the alternative hypothesis of the study.

(b) (9 points) Determine $\boldsymbol{\bar{x}_{\text{cutoff}}}$, the point that forms the rejection region. Select an appropriate critical value for the calculation.

`qnorm(p = 0.001/2, lower.tail = FALSE)` `[1] 3.290527`	`qnorm(p = 1-0.995, lower.tail = FALSE)` `[1] 2.575829`
`qnorm(p = (1-0.995)/2, lower.tail = FALSE)` `[1] 2.807034`	`qnorm(p = 0.001, lower.tail = FALSE)` `[1] 3.090232`

(c) (9 points) Utilize the cutoff value calculated in part (b) to calculate the power associated with detecting a **decrease** in the critical dimension by **2nm**. **Clearly show the mathematical steps** required to obtain the power of the test and select the correct code and output for computing the power of this test from the table.

`> pnorm(1.642632, lower.tail = TRUE)` `[1] 0.9497705`	`> pnorm(2.625553, lower.tail = FALSE)` `[1] 0.004325417`
`> pnorm(2.386993, lower.tail = TRUE)` `[1] 0.9915066`	`> pnorm(2.386993, lower.tail = FALSE)` `[1] 0.008493407`
`> pnorm(2.825847, lower.tail = TRUE)` `[1] 0.9976422`	`> pnorm(2.825847, lower.tail = FALSE)` `[1] 0.002357789`

Extra Credit

(3 points) A two-sample independent analysis is conducted by a researcher who decides to use the pooled estimator of the variance to construct a **95% confidence interval** for the difference in means between **population A** and **population B**. Unbeknownst to the researcher, the variances of these populations follow the relationship $\boldsymbol{\sigma_A^2 = 25 \times \sigma_B^2}$. If the confidence interval is built using a sample from **population A** of size $\boldsymbol{n_A = 825}$ and from **population B** with size, $\boldsymbol{n_B = 46}$, what can be said about the width of the confidence interval and the true coverage probability of the confidence interval? The formula for the pooled estimate of the variance is given below as $\boldsymbol{s_p^2}$.

$$\boldsymbol{s_{\mathrm{p}}^2 = \left(\frac{n_{\mathrm{A}} - 1}{n_{\mathrm{A}} + n_{\mathrm{B}} - 2}\right) s_{\mathrm{A}}^2 + \left(\frac{n_{\mathrm{B}} - 1}{n_{\mathrm{A}} + n_{\mathrm{B}} - 2}\right) s_{\mathrm{B}}^2}$$

Ⓐ It is wide enough to capture the true mean difference $\boldsymbol{\mu_A - \mu_B}$ exactly **95%** of the time that this procedure is performed.

Ⓑ It is the same as the width of a **95%** confidence interval that is constructed without pooling.

Ⓒ It is narrower than expected on average, capturing the true mean difference $\boldsymbol{\mu_A - \mu_B}$ less than **95%** of the time that this procedure is performed.

Ⓓ It is wider than expected on average, capturing the true mean difference $\boldsymbol{\mu_A - \mu_B}$ more than **95%** of the time that this procedure is performed.

(2 points) For a two-independent sample procedure, provide a scenario where using the pooled estimator may be advantageous. Explain why this approach is beneficial in that particular situation.

This page is intentionally left blank for scratch paper.

Question 4 Code/Output:

In the following outputs on this page **you can assume** that **correct degrees of freedom (df)** and **confidence level (conf.level) was utilized** wherever appropriate.

Output 1

```
t.test(box_data, conf.level=??, alternative = "two.sided", mu = 0)
t = 3098.8, df = ??, p-value < 2.2e-16
```

Output 2

```
t.test(box_data, conf.level=??, alternative = "less", mu = 0)
t = 3098.8, df = ??, p-value = 1
```

Output 3

```
t.test(box_data, conf.level=??, alternative = "greater", mu = 0)
t = 3098.8, df = ??, p-value < 2.2e-16
```

Output 4

```
t.test(box_data, conf.level=??, alternative = "two.sided", mu = 50)
t = 2.3595, df = ??, p-value = 0.0228
```

Output 5

```
t.test(box_data, conf.level=??, alternative = "less", mu = 50)
t = 2.3595, df = ??, p-value = 0.9886
```

Output 6

```
t.test(box_data, conf.level=??, alternative = "greater", mu = 50)
t = 2.3595, df = ??, p-value = 0.0114
```

Output 7

`> qnorm(0.09, lower.tail = TRUE)` `[1] -1.340755`	`> qt(0.09, df, lower.tail = TRUE)` `[1] -1.362417`
`> qnorm(0.09, lower.tail = FALSE)` `[1] 1.340755`	`> qt(0.09, df, lower.tail = FALSE)` `[1] 1.362417`
`> qnorm(0.09/2, lower.tail = TRUE)` `[1] -1.695398`	`> qt(0.09/2, df, lower.tail = TRUE)` `[1] -1.733557`
`> qnorm(0.09/2, lower.tail = FALSE)` `[1] 1.695398`	`> qt(0.09/2, df, lower.tail = FALSE)` `[1] 1.733557`