STAT 350 Exam 2 FALL 2024

V1

Name: ___________________________________ PUID ____________________________

Instructor (circle one): Heekyung Ahn Evidence Matangi Timothy Reese Halin Shin

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**. If you bring any other papers into the exam, you will get a **zero** on the exam. 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. **For free response questions 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.

Question Number	Total Possible	Your points
Problem 1 (True/False) (2 points each)	12	
Problem 2 (Multiple Choice) (3 points each)	15	
Problem 3	26	
Problem 4	25	
Problem 5	27	
Total	105	

1. **(12 points, 2 points each) True/False Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect.

1.1. A researcher collects various values from a dataset, including the **sample mean** $\bar{x}$, the **sample variance** s^2**,** the **t-test statistic** T_{TS} and the p**-value.**

Ⓣ or Ⓕ Each of these values is an example of a statistic.

1.2. The p**-value** can be considered a continuous random variable as it is a function of the test statistic, which itself is a function of the data,

Ⓣ or Ⓕ and therefore it must follow a normal distribution, as all data-derived quantities do.

1.3. A researcher conducts a hypothesis test at a significance level $\alpha = 0.01$ and fails to reject the null hypothesis.

Ⓣ or Ⓕ This result indicates that the null hypothesis is true at a **99%** confidence level.

1.4. In a study let $X_{\mathrm{A}_1}, X_{\mathrm{A}_2}, \ldots, X_{\mathrm{A}_n}$ represent the first set of measurements and $X_{\mathrm{B}_1}, X_{\mathrm{B}_2}, \ldots, X_{\mathrm{B}_n}$ represent the second set of measurements, where each pair $\left(X_{\mathrm{A}_i}, X_{B_i}\right)$ for $i \in \{1, 2, \ldots, n\}$ is taken from the same subject and are dependent. Let the difference between measurements for each subject be $D_i = X_{\mathrm{A}_i} - X_{B_i}$ be normally distributed and let $\sigma_{\mathrm{A}}^2 = Var\left(X_{\mathrm{A_i}}\right)$, and $\sigma_{\mathrm{B}}^2 = Var\left(X_{\mathrm{B_i}}\right)$ for all $i \in \{1, 2, \ldots, n\}$.

Ⓣ or Ⓕ If the **covariance between pairs** is a **positive constant** for all $i \in \{1, 2, \ldots, n\}$ i.e., $\mathrm{Cov}(X_{\mathrm{Ai}}, X_{\mathrm{Bi}}) = \sigma_{\mathrm{AB}} > 0$, meaning that when one measurement is higher (or lower) than average, the other measurement is likely to be similarly higher (or lower) than average. Therefore,

$$\mathrm{Var}(\bar{D}) = \frac{\sigma_{\mathrm{D}}^2}{n} < \frac{\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{B}}^2}{n}.$$

1.5. A researcher is studying the relationship between physical activity and cholesterol levels. However, they also collect data on participants' diets, which are known to influence both physical activity and cholesterol levels.

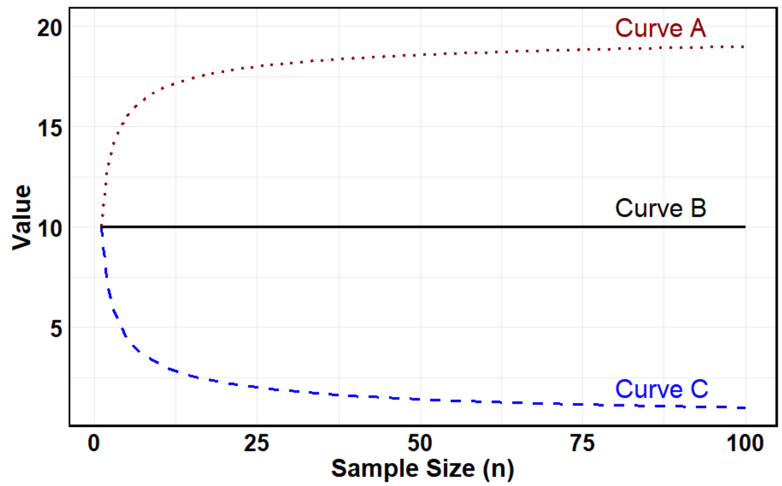
Ⓣ or Ⓕ Diet is considered a confounding variable in this study.

1.6. A researcher designed an experiment to test the effect of a new fertilizer on crop yield. They randomly assign the fertilizer treatment to half of the plots and leave the other half untreated. However, they notice that plots receiving fertilizer are closer to a water source.

Ⓣ or Ⓕ The random assignment is sufficient to ensure that the experiment is free from confounding variables.

2. **(15 points, 3 points each) Multiple Choice Questions.** Indicate the correct answer by completely filling in the appropriate circle. If you indicate your answer by any other way, you may be marked incorrect. **For each question, there is only one correct option letter choice.**

2.1. Assume $W_1, W_2, \cdots, W_n$ are **independent** samples drawn from some unknown distribution $f_W(w)$ with a **population mean** $\mu = 10$ and **population standard deviation** $\sigma = 10$. Which of the following statements is **FALSE** regarding the **distribution** of $\overline{W}$?



Ⓐ If the distribution $f_W(w)$ is heavily skewed, a larger sample is required to apply the central limit theorem.

Ⓑ **Curve A** represents the value of $sd(\overline{W})$ when the central limit theorem is not applicable.

Ⓒ **Curve B** represents the value of the $E[\overline{W}]$ for different sample sizes n.

Ⓓ **Curve C** indicates that the inference on $\mu_{\overline{W}}$ is more accurate as the sample size increases.

2.2. In the context of a one-sample procedure for constructing a **99% confidence interval** for the population mean μ, assuming all conditions for inference are met, which quantity is guaranteed to be within the interval?

Ⓐ 0

Ⓑ μ

Ⓒ σ

Ⓓ $\overline{x}$

Ⓔ **None of the above**

2.3. Consider an experiment in which a **sample of size 100** is drawn from a population with unknown mean **(μ)** and unknown standard deviation **(σ)**. The experiment is repeated using **ten different samples** of the **same size**, and a **99% confidence interval** is constructed for the **unknown mean** from **each sample**. Once all the intervals are computed, which of the following is always true?

Ⓐ The critical value used to calculate the confidence intervals is the same across the 10 replications of the experiment.

Ⓑ The numerical value at the center of the confidence interval is the same across the 10 replications of the experiment.

Ⓒ The margin of error is the same across the 10 replications of the experiment.

Ⓓ Each of the 10 computed confidence intervals contain the true mean **(μ)** with a **probability** of **0.99**.

Ⓔ Two or more of the above statements are correct.

2.4. Which of the following strategies can a researcher use to increase the power of a statistical hypothesis test?

Ⓐ Increase the sample size $\boldsymbol{n}$.

Ⓑ Increase the distance between the null value $\boldsymbol{\mu_0}$ and the alternative mean $\boldsymbol{\mu_A}$.

Ⓒ Reduce the population standard deviation $\boldsymbol{\sigma}$ by controlling extraneous variables.

Ⓓ Increase the significance level $\boldsymbol{\alpha}$ (Acceptable Type I error rate).

Ⓔ All of the above.

2.5. Suppose you are estimating a population parameter using two different estimators: Estimator A is unbiased but has high variance, while Estimator B is biased but has low variance. Which of the following statements is TRUE?

Ⓐ Estimator A is always preferred because it is unbiased.

Ⓑ Estimator B is always preferred because it has low variance.

Ⓒ Neither estimator is useful because both fail to provide accurate estimates of the true population parameter.

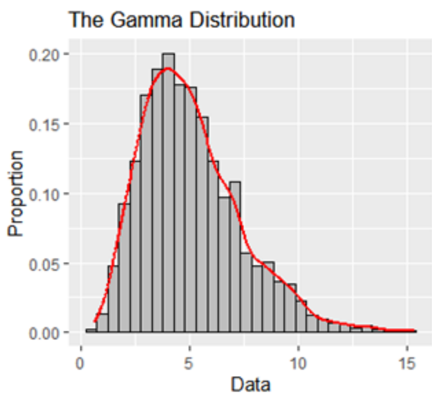
Ⓓ Depending on the context, Estimator B may be preferred if its bias is small, and variance is significantly lower than Estimator A's.

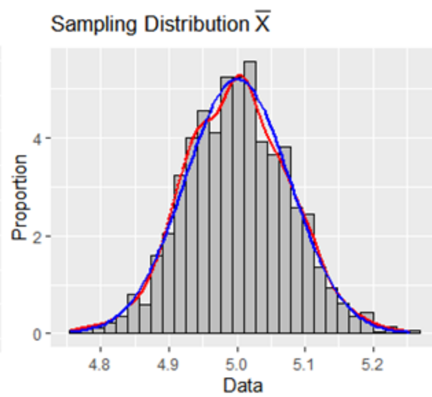
Ⓔ Both estimators are equally effective if the sample size is small enough.

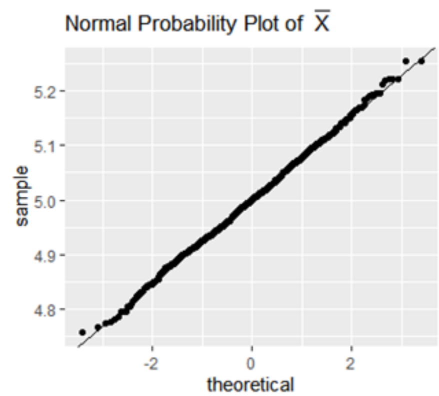
Free Response Questions 3-5. Show all work, clearly label your answers, and use **four decimal places**.

3. **(26 points)** An auto-insurance company plans to adjust policyholders' premiums based on historical data. According to the data, the claim amounts follow a gamma distribution with a **mean of 5k** and a **standard deviation of 2.25k**. The company expects **900 claims** to be filed in the upcoming month. These **900 claims** can be thought of as a **random sample** of **identically distributed** claims drawn from the **population distribution** of claim amounts. Assume the claims are **independent**.

   An enthusiastic statistician at the company conducted a simulation using **1,500 simple random samples**, each of size **900**, where each observation was randomly drawn from the **gamma distribution**. The following graphs are provided to support their analysis of $\bar{X}$, the average claim amount:







   a) **(10 points)** Describe the approximate distribution of $\bar{X}$, the average claim amount of **900** auto-insurance claims. Provide a detailed justification for why this approximation is valid, including the important theoretical principle that supports your result.

   b) **(3 points)** Find the mean and standard deviation of the sampling distribution of $\bar{X}$.

$\mu_{\bar{X}} =$ $\sigma_{\bar{X}} =$

c) **(3 points)** Select the correct code for determining the probability that the **average** of **900 claims** would be **greater** than **5.15k**?

Ⓐ pnorm(5.15, mean = 5, sd = 2.25, lower.tail = FALSE)

Ⓑ pnorm(5.15, mean = 5, sd = 2.25, lower.tail = TRUE)

Ⓒ pnorm(5.15, mean = 5, sd = 0.075, lower.tail = FALSE)

Ⓓ pnorm(5.15, mean = 5, sd = 0.075, lower.tail = TRUE)

Ⓔ pgamma(5.15, shape = 5, rate = 2.25, lower.tail = FALSE)

Ⓕ pgamma(5.15, shape = 5, rate = 2.25, lower.tail = TRUE)

d) **(10 points)** Suppose only **10 claims** are expected in the upcoming month. Can the same inference about the **average claim** be made in this case? Justify your answer based on relevant theoretical principles.

4. (25 points) GreatNotes is developing software that converts handwritten mathematical notations into typed text. To evaluate its performance, the company used 200 images of handwritten mathematical equations. Of these, 100 images were included in the training dataset, paired with their correct typed formats. The remaining 100 images were **withheld** from training to serve as a test set of new, unseen data.
After training, the company tested the algorithm on all 200 images, comparing each algorithm-generated output to its corresponding correct typed format to assess accuracy.

Each output was scored for accuracy, with scores ranging from 0 to 100. A numerical summary of the results is provided below:

	Training Data	Withheld Data	Training – Withheld
n	100	100	100
Sample Mean	96.4	95.2	1.2
Sample Standard Deviation	2.3	4.5	4.3

It is common for handwriting recognition algorithms to achieve higher accuracy on data used during training. However, for commercial success, GreatNotes must ensure that the algorithm performs comparably on new, **unseen data**. They will conclude that the algorithm fails to generalize if the **true mean accuracy** on **training data** is significantly higher than **the true mean accuracy** on **withheld data**.

Using a **91% confidence level**, perform a hypothesis test to determine whether the algorithm fails to generalize.

a) (2 points) Which two-sample method should be used?

Ⓐ Two-sample Independent Procedure

Ⓑ Two-sample Paired Procedure

b) (5 points) Perform the **first two steps** of the four-step hypothesis test.

c) (8 points) Compute the test statistic. Show work.

d) (3 points) Select the R code that would correctly compute the **p-value**:

Ⓐ `pnorm(test_statistic, lower.tail = TRUE)`

Ⓑ `pt(test_statistic,  df=147.42, lower.tail = TRUE)`

Ⓒ `pt(test_statistic, df=199, lower.tail = TRUE)`

Ⓓ `pnorm(test_statistic, lower.tail = FALSE)`

Ⓔ `pt(test_statistic, df=147.42, lower.tail = FALSE)`

Ⓕ `pt(test_statistic, df=199, lower.tail = FALSE)`

e) (7 points) The resulting ***p*-value** was approximately **0.009**. Provide the formal decision and interpret the conclusion in the context of the problem.

5. **(27 points)** Urbanization has been associated with an increase in coyote sightings in Georgia (Mowry et al., 2020). This has raised concerns about the role of coyotes in urban ecosystems, particularly regarding human-coyote conflicts and negative interactions with pets.

   Residents of Atlanta, GA, believe that coyotes are large on average, with a mean length of **at least 94 cm**. However, the Georgia Department of Natural Resources **(DNR)** suspects that the true mean length is **less than 94 cm**. To investigate, the DNR staff used **Geographic Information Systems (GIS)** to identify and randomly select sampling locations, supplemented by **satellite imagery** to capture **29 images of coyotes**. From these images, they measured the lengths of the coyotes.

   The **sample mean length** was **89.17 cm**. Based on historical data, the DNR has determined that coyote lengths follow a normal distribution with a **standard deviation of 9 cm**.

   a) **(8 points)** Calculate an appropriate **90% confidence interval** or **bound** to assess the belief of the true mean length of coyotes by the Georgian **DNR**. Clearly **specify** which **R output** from the last page of the exam you used.

b) **(5 points)** Interpret the results obtained from part (a) within the context of the problem.

c) **(14 points)** Carry out a hypothesis test on whether the data supports the claim made by the DNR staff. Use the information from above and on the last page of the exam to perform the **four-step hypothesis test**. Clearly specify which **R** output from the last page of the exam was used to obtain your conclusion. Test at $\boldsymbol{\alpha = 0.1}$.

Question 5 Code/Output:

Output 1

```
t.test(coyote_data, conf.level = 0.90, alternative = "greater", mu = 89.17)
t = 0.0012646, df = 28, p-value = 0.4995
```

Output 2

```
t.test(coyote_data, conf.level = 0.90, alternative = "less", mu = 94)
t = -2.5293, df = 29, p-value = 0.008671
```

Output 3

```
t.test(coyote_data, conf.level = 0.90, alternative = "two.sided", mu = 94)
t = -2.5293, df = 28, p-value = 0.01734
```

Output 4

```
z_TS <- (89.17-94)/9
cat("Test Statistic is: ",z_TS, "\n")
Test Statistic is:  -0.5366667
p_value <- pnorm(z_TS, lower.tail = TRUE)
cat("p-value is: ", p_value, "\n")
p-value is:  0.2957489
```

Output 5

```
z_TS <- (89.17-94)/(9/sqrt(29))
cat("Test Statistic is: ",z_TS, "\n")
Test Statistic is:  -2.890038
p_value <- pnorm(z_TS, lower.tail = TRUE)
cat("p-value is: ", p_value, "\n")
p-value is:  0.001925974
```

Output 6

```
z_TS <- (89.17-94)/9
cat("Test Statistic is: ",z_TS, "\n")
Test Statistic is:  -0.5366667
p_value <- 2*pnorm(z_TS, lower.tail = TRUE)
cat("p-value is: ", p_value, "\n")
p-value is:  0.5914979
```

Output 7

```
z_TS <- (89.17-94)/(9/sqrt(29))
cat("Test Statistic is: ",z_TS, "\n")
Test Statistic is:  -2.890038
p_value <- 2*pnorm(z_TS, lower.tail = TRUE)
cat("p-value is: ", p_value, "\n")
p-value is:  0.003851947
```

Output 8

> qnorm (p=0.05, lower.tail = TRUE) **-1.644854**	> qt (p=0.05, df = 28, lower.tail = FALSE) **1.701131**
> qnorm (p=0.1, lower.tail = FALSE) **1.281552**	> qt (p=0.1, df = 28, lower.tail = FALSE) **1.312527**
> qnorm (p=0.1, lower.tail = TRUE) **-1.281552**	> qt (p=0.1, df = 28, lower.tail = TRUE) **-1.312527**
> qnorm (p=0.05, lower.tail = FALSE) **1.644854**	> qt (p=0.05, df = 28, lower.tail = TRUE) **-1.701131**