STAT 350 Exam 2 SPRING 2025

V1

Name: _____ PUID _____

Instructor (circle one): Heekyung Ahn Evidence Matangi Timothy Reese Halin Shin
Class Start Time: ○ 10:30 AM ○ 11:30 AM ○ 12:30 PM ○ 1:30 PM ○ 2:30 PM ○ 3:30 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. 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.
4. The crib sheet can be a handwritten or type double-sided 8.5in x 11in sheet.
5. Keep your bag closed and cellphone stored away securely at all times during the exam.
6. If you share your calculator without permission or have a cell phone at your desk, you will get a **zero** on the exam. Do not take out your cell phone until you are next in line to submit your exam.
7. The exam is only 60 minutes long so there will be no 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.
8. **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.
9. All numeric answers should have **four decimal places** unless stated otherwise.
10. 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.
11. 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.
12. It is strictly prohibited to smuggle this exam outside. Your exam will be returned to you on Gradescope after it is graded.

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	24	
Problem 4	23	
Problem 5	31	
Total	105	

The rest of this page can be used for scratch work

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.

   a) Consider a sequence $X_1, X_2, \ldots, X_n$ of independent and identically distributed random variables drawn from a population $f_X(x)$ with finite mean μ, and finite variance $\sigma^2 > 0$. Define the sample sum as:

   $$S_n = \sum_{i=1}^{n} X_i,$$

   where the subscript n to explicitly indicates that the sum is over n random variables.

   Ⓣ or Ⓕ According to the **central limit theorem (CLT)**, the standardized sample sum

   $$\frac{(S_n - n\mu)}{\sqrt{n} \cdot \sigma}$$

   is exactly distributed as a standard Normal distribution regardless of the shape of $f_X(x)$, provided $n \geq 30$.

1.2. In a randomized block design (RBD), treatments are randomly assigned to experimental units within distinct blocks.

Ⓣ or Ⓕ This is done to balance rather than mitigate or remove the impact of extraneous variables on experimental results.

1.3. Denote $\tau_n = \frac{1}{\sqrt{n}} SD(X_1 + X_2 + \cdots + X_n)$, where the X_i's are independent and identically distributed with finite variance σ^2.

Ⓣ or Ⓕ Then it follows that $\tau_3 > \tau_4$.

1.4. A researcher is calculating the sample size needed for a confidence interval with a fixed margin of error. The population standard deviation is known.

Ⓣ or Ⓕ The required sample size increases as the confidence coefficient decreases.

1.5. The functions **pnorm()** and **pt()** are available functions in R (RStudio),

Ⓣ or Ⓕ they directly provide critical values for constructing confidence intervals and bounds.

1.6. When conducting a hypothesis test in which the alternative hypothesis is true,

Ⓣ or Ⓕ the probability of rejecting the null hypothesis increases by taking a larger sample size.

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. Which of the following techniques is **primarily used** to control or **reduce variability** arising from extraneous factors in experimental design?

Ⓐ Blocking

Ⓑ Replication

Ⓒ Randomization

Ⓓ Realism

Ⓔ All of the above

2.2. Time, resources, and practicality influence statistical studies. Which of the following is **essential** for drawing **generalizable conclusions** from a statistical investigation?

Ⓐ Selecting a representative sample

Ⓑ Having a sufficiently large sample size

Ⓒ Using appropriate well-established statistical methods

Ⓓ Clearly defining the population of interest

Ⓔ All of the above are essential for drawing generalizable conclusions

2.3. You used Glassgate, a platform where salaries are posted anonymously, to look up salary information for a small-sized consulting firm. You found salary data from 13 verified junior data scientists.
Which of the following statements is **NOT always true** regarding the sampling distribution of the average salary computed from these 13 junior data scientists?

Ⓐ If the population distribution has a strong positive skew, the Central Limit Theorem cannot be applied.

Ⓑ Without additional information about the population distribution, the exact sampling distribution cannot be determined.

Ⓒ The mean of the sampling distribution, $\mu_{\bar{X}}$, is equal to the population mean μ_X, regardless of sample size.

Ⓓ The standard deviation of the sampling distribution $\sigma_{\bar{X}}$, is always smaller than the population standard deviation σ_X when the sample size exceeds 2.

Ⓔ The central limit theorem ensures that the sampling distribution of $\bar{X}$ is approximately Normal.

2.4. A delivery company, CargoSwift Logistics, operates small vans that regularly transport packages between warehouses. Each trip includes a fixed set of loading equipment (metal securing racks, crates, and straps) weighing exactly **30 lbs**. The remaining cargo consists of **sixteen** individual packages, each with weights **independently** and **identically distributed** as follows:

$$X_i \sim \mathbf{Uniform}(a = 44, b = 56), \quad i = 1, 2, \ldots, 16.$$

Recall that $E[X_i] = \frac{a+b}{2}$ and $\mathbf{Var}(X_i) = \frac{(b-a)^2}{12}$.

The total weight T for a typical truckload of **16 packages** is given by:

$$T = 30 + \sum_{i=1}^{16} X_i.$$

CargoSwift's delivery vans have a maximum safe weight load capacity of **850 lbs**. Select the correct code to calculate the approximate probability that a randomly selected van containing **16 packages** would **exceed** the safe weight load.

Ⓐ
```
pt(1.4434, df = 15, lower.tail = FALSE)
```
Ⓑ
```
pnorm(1.4434, lower.tail = FALSE)
```
Ⓒ
```
pt(3.6084, df = 15, lower.tail = FALSE)
```
Ⓓ
```
pnorm(3.6084, lower.tail = FALSE)
```
Ⓔ
```
pt(5.7735, df = 15, lower.tail = FALSE)
```
Ⓕ
```
pnorm(5.7735, lower.tail = FALSE)
```
Ⓖ
```
punif(850, min = 704, max = 896, lower.tail = FALSE)
```

2.5. A nutritionist conducts a study to test whether a new dietary program significantly **reduces** cholesterol levels from a known baseline of **200 mg/dL**. They collect data from **25 participants** and observe a **sample standard deviation of 18 mg/dL**. They plan to perform a **lower-tailed hypothesis test** at significance level $\alpha = 0.05$.

Select the correct lines of R code to calculate the approximate **power** of this test for detecting a reduction in the mean cholesterol to **190 mg/dL**:

Ⓐ
```
cutoff <- 200 + qt(0.05, df=24, lower.tail=FALSE)*(18/sqrt(25))
pt((cutoff - 190)/(18/sqrt(25)), df=24, lower.tail=TRUE)
```
Ⓑ
```
cutoff <- 190 - qt(0.95, df=24, lower.tail=FALSE)*(18/sqrt(25))
 pt((cutoff - 200)/(18/sqrt(25)), df=24, lower.tail=TRUE)
```
Ⓒ
```
cutoff <- 200 - qnorm(0.95, lower.tail=FALSE)*(18/sqrt(25))
pnorm((cutoff - 190)/(18/sqrt(25)), lower.tail=FALSE)
```
Ⓓ
```
cutoff <- 200 + qnorm(0.05, lower.tail=FALSE)*(18/sqrt(25))
pnorm((cutoff - 190)/(18/sqrt(25)), lower.tail=FALSE)
```
Ⓔ
```
cutoff <- 200 - qt(0.05, df=24, lower.tail=FALSE)*(18/sqrt(25))
pt((cutoff - 190)/(18/sqrt(25)), df=24, lower.tail=TRUE)
```

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

3. **(24 points)** A car manufacturer advertises that its new compact SUV averages **40 miles per gallon** (mpg). A consumer advocacy group wants to **evaluate** this claim and believes the true average may be **lower**. They obtain **54** SUVs from the same model year and measure their fuel efficiency under combined city/highway driving, simulating what an average driver might encounter. After recording each vehicle's mileage under these controlled yet representative conditions, the group finds a **sample mean** of **37.8 mpg**. The population's standard deviation is unknown. They plan to conduct a hypothesis test at a **3% significance level** to assess the manufacturer's claim.

   a) **(3 points)** Identify the parameter of interest. Clearly state its symbolic notation and define it briefly in the context of this scenario.

Let μ_{MPG} denote the true average miles per gallon of the new compact SUV.

   b) **(4 points)** Write the null and alternative hypotheses clearly in symbolic notation.

$$H_0: \mu_{\text{MPG}} \geq 40$$

$$H_a: \mu_{\text{MPG}} < 40$$

   c) **(8 points)** The consumer advocacy group reports the corresponding **confidence bound** at **38.43**. Using this confidence bound and the provided critical values, deduce the value of the sample standard deviation. Provide your answer rounded to four decimal places.

`qnorm(0.03, lower.tail=FALSE) = 1.88`	`qt(0.03, df= 53, lower.tail=FALSE)= 1.92`
`qnorm(0.015, lower.tail=FALSE) = 2.17`	`qt(0.015, df= 53, lower.tail=FALSE)= 2.23`

```
qt(0.03, df= 53, lower.tail=FALSE)= 1.92
```

confidence upper bound → $38.43 = \bar{x} + t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$

$$38.43 = 37.8 + 1.92 \cdot \frac{s}{\sqrt{54}}$$

$$s = (38.43 - 37.8) \cdot \frac{\sqrt{54}}{1.92} = 2.4112$$

d) **(2 points)** Is the reported confidence bound of **38.43** in part **c)** an upper bound or a lower bound? Please mark the correct option.

Ⓐ Upper Bound

Ⓑ Lower Bound

e) **(7 points)** Using the results of parts **c)** and **d)**, obtain the result of the hypothesis test and write the formal contextual conclusion.

The **97%** confidence upper bound reported in c) of **38.43** is lower than the null value of $\boldsymbol{\mu_0 = 40.}$ Therefore, we would have evidence to reject the null hypothesis at a significance level $\boldsymbol{\alpha = 0.03}$.

The data **does** give support (**p-value** < 0.03) to the claim that the true average MPG of the new compact SUVs is less than 40 MPG.

The rest of this page can be used for scratch work

4. **(23 points)** 🕵️Special Agent Gibbs decided to pursue his career in 🎓academia specializing in 🛡️national security, post-traumatic stress, and investigation strategies. As part of his research, Gibbs requested access to a sensitive dataset containing information on veterans. Due to privacy and security considerations, the custodians of the dataset could not release it directly to Gibbs. Instead, they provided Gibbs with detailed descriptions of the available variables and asked him to submit clearly defined research questions. Their analyst team would then run analyses on the secure data and provide Gibbs with appropriate statistical summaries, test statistics, and supporting details. Assume none of the population standard deviations are known.

   a) **(3 points)** Five of Gibbs' research questions happen to be lower-tailed hypotheses, $H_a: \mu < \mu_0$. Which one of the following test statistics would be most likely to reject H_0? Assume the same degrees of freedom for all five test statistics.

   Ⓐ $t_{ts} = -2.25$

   Ⓑ $t_{ts} = -1.02$

   Ⓒ $t_{ts} = 0.02$

   Ⓓ $t_{ts} = 1.56$

   Ⓔ $t_{ts} = 3.02$

   b) **(3 points)** 🕵️Gibbs believes that veterans' financial status may vary by different factors, such as location, household composition, and health status, so he asked the analyst team if they can control these extraneous variables. Without direct access to the dataset himself, which of the following strategies is most appropriate for the analyst team to handle these extraneous variables based on Gibbs' request?

   Ⓐ Discard any cases which are deemed extreme by Gibbs to obtain a consistent dataset.

   Ⓑ Partition the data into blocks (or strata) aligned with these extraneous variables, thereby reducing their confounding influence during analysis.

   Ⓒ Randomly select a small number of veterans broadly representative of the entire population, simplifying the analysis.

   Ⓓ Ignore extraneous variables since a small bias is acceptable if it results in reduced variance.

   Ⓔ Control all extraneous variables from the beginning by requiring veterans to live according to randomly assigned conditions.

c) (3 points) Gibbs identified two variables: social isolation level (SIL, categorical) and mental stability score (MSS, numerical). SIL has four categories: *socially active*, *somewhat socially active, somewhat isolated,* and *completely isolated*. Specifically, Gibbs wants to see if the **mean difference** in **MSS** between ***somewhat isolated*** and ***completely isolated*** groups is **greater** than **20 points**.

The analysts, acting on Gibbs' request, paired **50 veterans** from the ***somewhat isolated*** group with **50 veterans** from the ***completely isolated*** group. Pairing was done based on demographics and veteran history to control possible extraneous factors. After matching, the difference in mental stability scores was computed for each pair as $\mathbf{D = Somewhat\ Isolated - Completely\ Isolated}$.

Which of the following hypothesis testing procedures is appropriate to answer Gibbs' question?

Ⓐ One-sample t-test

Ⓑ Two-sample Independent t-test

Ⓒ Two-sample Matched Pairs t-test

Ⓓ None of the above

d) (4 points) State clearly the null and alternative hypotheses using the appropriate mathematical notations.

$$\mathbf{D = Somewhat\ Isolated - Completely\ Isolated}$$

$$H_0: \mu_D \leq 20$$

$$H_a: \mu_D > 20$$

e) (3 points) Gibbs wants to find the p**-value** of a test statistic, $t_{ts} = 2.14$. Which of the following R code statements returns the correct p**-value**?

Ⓐ `pt(2.14, df = 50, lower.tail = FALSE)`

Ⓑ `pt(2.14, df = 50, lower.tail = TRUE)`

Ⓒ `pt(2.14, df = 49, lower.tail = FALSE)`

Ⓓ `pt(2.14, df = 49, lower.tail = TRUE)`

Ⓔ `pt(2.14, df = 99, lower.tail = FALSE)`

Ⓕ `pt(2.14, df = 99, lower.tail = TRUE)`

f) (7 points) The calculated p**-value** is **0.01868**. At a significance level of $\alpha = 0.02$, state your formal decision and conclusion in the context of the problem.

The p**-value** $= 0.01868 \leq 0.02 = \alpha$ therefore we have evidence to reject the null hypothesis H_0.

The data **does** give **some** support (p**-value** = **0.01868**) to the claim that the **true average mental stability score MSS** is **higher** for those veterans classified as **Somewhat Isolated** vs. **Completely Isolated**.

5. (31 points) The rapid growth of food delivery services has dramatically increased the number of two-wheeled couriers on city streets. While this surge has benefited local businesses and consumers, it has also led to increased traffic congestion and safety concerns, prompting some cities to consider stricter delivery regulations (MassLive, 2025).

California cities, in particular, are often at the forefront of adopting innovative transportation policies and stricter environmental and safety regulations. To better understand the potential impact of these regulations and inform future policy decisions, a food delivery analytics firm seeks to compare the total monthly mileage traveled by two-wheeled delivery couriers employed by a major third-party platform. Mileage is measured as the combined total distance traveled by all couriers within each city over a four-week period in late summer 2024. Using the summary statistics below, the firm aims to determine if there is a significant difference in the average total monthly mileage between California and non-California cities.

statistic	California cities	Non-California cities
n	11	11
$\bar{x}$	752,962	824,387.6
s	697,033.6	918,850.2

a) (4 points) State the assumptions necessary for the hypothesis test on the average total monthly mileage between California and non-California cities.

As this is a two-independent sample procedure we have the following assumptions:

- Independence:
  - The cities in each group (California and non-California) constitute independent, random samples.
  - Total monthly Mileage measurements within each city are independent of one another.
- Normality: The total monthly mileage in each group is approximately normally distributed. This is especially important given the small sample sizes.

b) (4 points) As part of your summer internship with the food delivery analytics firm, you have been asked to determine if there is a significant difference in average total monthly mileage between California and non-California cities. Using $\alpha = 0.03$, perform the **first two steps** of the four-step hypothesis test procedure.

Step 1: Identify and describe the parameter(s) of interest for the two populations.

$\mu_{\text{California}}$ and $\mu_{\text{non-California}}$ are the parameter of interest representing the population mean total monthly mileage of two-wheeled couriers across California and non-California cities.

Step 2: State the hypothesis.

$$H_0: \mu_{\text{California}} - \mu_{\text{non-California}} = 0$$
$$H_a: \mu_{\text{California}} - \mu_{\text{non-California}} \neq 0$$

c) (8 points) Calculate the appropriate test statistic for comparing the mean total mileage between California and non-California cities. Clearly show the steps and formula used.

$$t_{\text{TS}} = \frac{\bar{x}_{\text{California}} - \bar{x}_{\text{non-California}}}{\sqrt{\frac{s^2_{\text{California}}}{n_{\text{California}}} + \frac{s^2_{\text{non-California}}}{n_{\text{non-California}}}}}$$

$$t_{\text{TS}} = \frac{752,962 - 824,387.6}{\sqrt{\frac{697,033.6^2}{11} + \frac{918,850.2^2}{11}}} = \frac{-71425.6}{347738.3} = -0.2054$$

d) (3 points) Without additional assumptions, which of the following represents the most appropriate degrees of freedom for the test statistic found in part **c)**?

Ⓐ 10

Ⓑ 20

Ⓒ 18.646

Ⓓ None of the above

The analytics firm also wants to estimate the difference in the true mean total monthly mileage between California and non-California cities. They decided to construct a **97% confidence interval** for the difference.

e) (3 points) Select the correct R code and output that provides the correct critical value for constructing this **97% confidence interval**. Assume 'ν' represents the correct degrees of freedom if appropriate.

Ⓐ
```
qnorm(p=0.03/2, lower.tail = FALSE)
[1] 2.17009
```

Ⓑ
```
qt(p=0.03/2, df= ν, lower.tail = FALSE)
[1] 2.349237
```

Ⓒ
```
2*qnorm(p=0.03, lower.tail = FALSE)
[1] 3.761587
```

Ⓓ
```
2*qt(p=0.03, df= ν, lower.tail = FALSE)
[1] 4.004843
```

f) (6 points) Using the summary statistics provided and the critical value, calculate the **97% confidence interval** for the difference in mean total monthly mileage (California minus non-California). Clearly show all necessary formulas and steps.

$$\bar{x}_{\text{California}} - \bar{x}_{\text{non-California}} = 752,962 - 824,387.6 = -71425.6$$

$$\sqrt{\frac{s^2_{\text{California}}}{n_{\text{California}}} + \frac{s^2_{\text{non-California}}}{n_{\text{non-California}}}} = 347738.3$$

$$(\bar{x}_{\text{California}} - \bar{x}_{\text{non-California}}) \pm t_{0.03/2,\nu} \cdot \sqrt{\frac{s^2_{\text{California}}}{n_{\text{California}}} + \frac{s^2_{\text{non-California}}}{n_{\text{non-California}}}}$$

$$(-71425.6 - 2.349237 \cdot 347738.3, -71425.6 + 2.349237 \cdot 347738.3)$$

$$(-888,345.3, \quad 745,494.1)$$

g) (3 points) Provide an accurate interpretation of the confidence interval in context, explaining what it indicates about the difference in average total monthly mileage between California and non-California cities.

We are 97% confident that the true difference in average total monthly mileage is captured by the interval $(-888,345.3, \quad 745,494.1)$.
Since this interval includes both negative and positive values (including zero), it indicates that there is **no clear evidence** of a significant difference in the average monthly mileage between California and non-California cities based cities at the $\alpha = 0.03$ significance level.