STAT 350 Exam 2 FALL 2025

V1

Name: ________________________________	PUID ________________________

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

Select Class Meeting Days/Time

○ **T/Th** 9:00AM-10:15AM ○ **MW** 1:30PM-2:45PM
○ **MWF** 11:30AM-12:20PM ○ **MWF** 12:30 PM-1:20PM ○ **MWF** 1:30 PM-2:20PM
○ **MWF** 2:30 PM-3:20PM ○ **MWF** 3:30-4:20PM ○ **MWF** 3:30PM-4:20PM ○ **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. 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 typed double-sided 8.5in x 11in sheet.
5. 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.
6. 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.
7. **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 must 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.
8. All numeric answers should have **four decimal places** unless stated otherwise.
9. 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**.
10. 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.
11. 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	27	
Problem 4	28	
Problem 5	23	
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.

1.1. Assume that $X_1, X_2, \cdots, X_n$ is a random sample from a **Cauchy distribution**, which has an **undefined expected value** and **variance**. (i.e., $E[X_i]$ and $Var(X_i)$ are not finite.)

Ⓣ or Ⓕ Performing a one-sample hypothesis test for the population mean (μ) is valid if the sample size (n) is sufficiently large.

1.2. The lengths of pregnancies are normally distributed with a mean of 268 days and a standard deviation of 15 days.

Ⓣ or Ⓕ Then the probability that a randomly selected pregnant woman's pregnancy length is less than 265 is **larger** than the probability that the mean pregnancy length of a random sample of 40 pregnant women is less than 265.

1.3. In experimental design, researchers often encounter extraneous variables that may influence the response variable alongside the factor of interest.

Ⓣ or Ⓕ In a Randomized Block Design (RBD), blocks are used to control the factor of interest, while randomization controls extraneous variables.

1.4. A researcher conducts a hypothesis test and obtains a p**-value** of 0.03.

Ⓣ or Ⓕ This means there is a 3% probability that the null hypothesis is true.

1.5. A 95% confidence interval for the population mean μ is constructed from sample data

Ⓣ or Ⓕ A sample mean $\bar{x}$ falls within the 95% confidence interval for all possible samples from the population.

1.6. A researcher wants to estimate the average income in a city with diverse neighborhoods.

Ⓣ or Ⓕ In a stratified random sample, the city is divided into neighborhoods (strata), and then a few complete neighborhoods are randomly selected and all residents within those neighborhoods are surveyed.

2. **(15 points, 3 pts 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 unless specified.**

2.1. A student proposes an estimator for an unknown parameter θ the estimator $\hat{\theta}$ has expected value $\boldsymbol{E}[\hat{\boldsymbol{\theta}}] = \left(\frac{\boldsymbol{n}}{\boldsymbol{n}-1}\right)\boldsymbol{\theta} + \mathbf{5}$, where $\boldsymbol{n}$ is the sample size. What is the exact bias (for finite $\boldsymbol{n}$) and the asymptotic bias as $\boldsymbol{n} \to \infty$ of this estimator?

Ⓐ Exact bias = 5; Asymptotic bias = 0

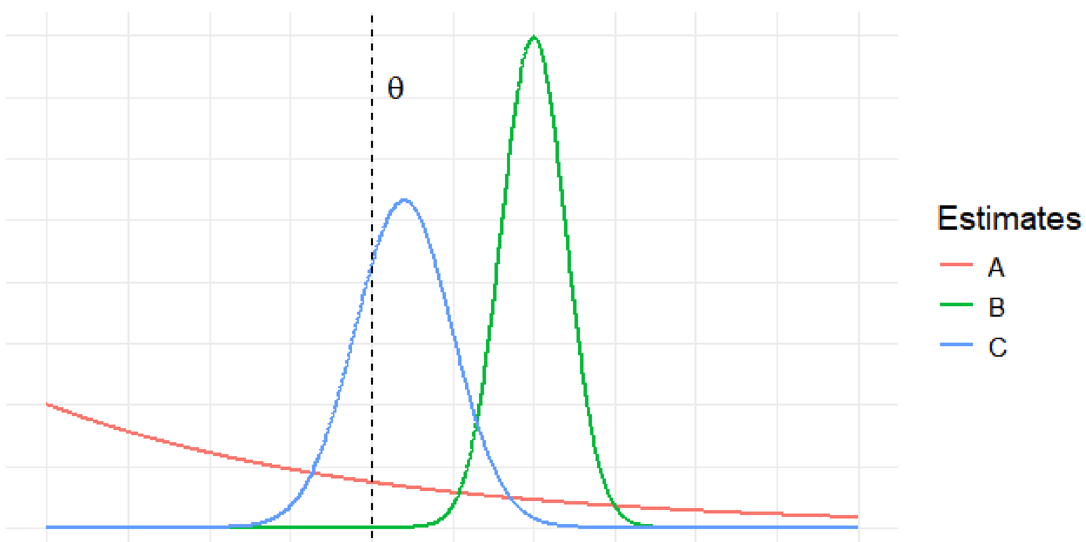
Ⓑ Exact bias = $\left(\frac{1}{n-1}\right) \cdot \boldsymbol{\theta}$; Asymptotic bias = 5

Ⓒ Exact bias = $\left(\frac{1}{n-1}\right) \cdot \boldsymbol{\theta}$ + 5; Asymptotic bias = 5

Ⓓ Exact bias = $\left(\frac{1}{n-1}\right) \cdot \boldsymbol{\theta}$ + 5; Asymptotic bias = 0

Ⓔ Exact bias = $\left(\frac{n}{n-1}\right) \cdot \boldsymbol{\theta}$ + 5; Asymptotic bias = θ

2.2. Three estimators, $\hat{\boldsymbol{\theta}}_A$, $\hat{\boldsymbol{\theta}}_B$, $\hat{\boldsymbol{\theta}}_C$, are constructed for an unknown target parameter $\boldsymbol{\theta}$, and their sampling distributions are visualized in the graph below. Which of the following statements is **TRUE** about the estimators?



Ⓐ $\hat{\boldsymbol{\theta}}_B$ is preferred over $\hat{\boldsymbol{\theta}}_A$ because $\hat{\boldsymbol{\theta}}_B$ has a smaller variance.

Ⓑ $\hat{\boldsymbol{\theta}}_A$ is preferred over $\hat{\boldsymbol{\theta}}_C$ if $\hat{\boldsymbol{\theta}}_A$ has a smaller bias.

Ⓒ $\hat{\boldsymbol{\theta}}_C$ is preferred over $\hat{\boldsymbol{\theta}}_A$ even if $\hat{\boldsymbol{\theta}}_A$ has a smaller bias.

Ⓓ On repeated samples, $\hat{\boldsymbol{\theta}}_A$ values hardly vary around the true parameter.

Ⓔ The best estimate can be determined only after obtaining realized values.

2.3. Two fertilizers are tested on different plots to compare their effects on crop yield. Summary statistics: $\boldsymbol{n_1 = 22, s_1 = 19.5}$ kg/hectare and $\boldsymbol{n_2 = 20, s_2 = 4.7}$ kg/hectare. Which statistical inference procedure is most appropriate for comparing mean yields?

Ⓐ One-sample $\boldsymbol{t}$-procedures

Ⓑ Two-sample paired $\boldsymbol{t}$-procedures

Ⓒ Pooled two-sample independent $\boldsymbol{t}$-procedures

Ⓓ Welch two-sample independent $\boldsymbol{t}$-procedures

2.4. A researcher wants to test the effect of a new type of feed on the weight gain of chickens. They have 100 chickens, but they are housed in 10 different coops (10 chickens per coop). The researcher knows that conditions (like temperature and lighting) vary slightly between coops, which might affect weight gain. To account for this, the researcher randomly assigns 5 chickens within each coop to the new feed and the other 5 to the standard feed.

Which experimental design technique is demonstrated by separating the chickens by coop before assigning the feed?

Ⓐ Completely Randomized Design

Ⓑ Randomized Block design

Ⓒ Simple Random Sample

Ⓓ Matched Pairs Design

Ⓔ Stratified Random Sampling

2.5. Chloride deposits are markers for early Mars' aqueous past with important implications for our understanding of Mars' climate and habitability. Purdue scientists are in the process of investigating high-resolution image surfaces of 33 chloride deposits from the southern highlands of Mars. Researchers from a different university have claimed that the mean diameter of chloride deposits is 1650m with the standard deviation of 779.42m. Studies of geological features suggest that the diameters of natural deposits tend to follow approximately symmetric distributions. Based on the Central Limit Theorem and assuming the other researchers' claim correctly describes the population, which of the following statements is incorrect?

Ⓐ The standard deviation of the sampling distribution of $\boldsymbol{\bar{X}}$ for Purdue investigations should be 779.42m.

Ⓑ The mean of the sampling distribution of $\boldsymbol{\bar{X}}$ for Purdue investigations is 1650m.

Ⓒ The sampling distribution of $\boldsymbol{\bar{X}}$ for Purdue investigations is approximately normal.

Ⓓ We cannot assume the population distribution of deposit diameters is exactly normal.

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

3. (27 points) A coffee company is testing a new, faster roasting machine. The old machine roasts beans to a target mean moisture level of 8.0%. The company suspects the new machine **(N)** produces a different mean moisture level than the old machine **(O)**.

They conduct an experiment. They take 16 batches of the same type of green coffee bean. For each batch, they split it in half, roasting one half with the new machine and the other half with the old machine. The moisture level for each roasted half is recorded.

The company calculates the difference for each batch: **D** = $\mathbf{Moisture_N - Moisture_O}$. The data for the 16 differences yields sample mean difference 0.25% and standard deviation of differences 0.60%. **The researchers have verified that the distribution of differences is approximately normal.**

a) (2 points) Which testing procedure is appropriate for this experiment?

Ⓐ Two-sample independent t-test

Ⓑ Two-sample paired t-test

b) (4 points) Explain what characteristic(s) in the experimental design motivated your choice of testing procedure in part **a)**.

Each batch is split into two halves with each half roasted using different machines. This means the two measurements within a batch are **not independent**; they share the same batch characteristics (bean type, initial moisture, etc.).

c) (2 points) Provide the first two steps of the four-step hypothesis testing procedure.

Step 1: Define the parameter(s) of interest.

$\mu_D: \mu_{\text{Moisture}_N} - \mu_{\text{Moisture}_O}$ representing the true population mean moisture difference associated with roasting from the two machines (New – Old).

Step 2: Define the hypothesis:

$H_0: \mu_D = 0$

$H_a: \mu_D \neq 0$

d) (10 points) Calculate the test statistic for this test. Show your work.

$$t_{TS} = \frac{\bar{D} - 0}{s_D/\sqrt{n}} = \frac{0.25}{0.6/\sqrt{16}} = 1.6667$$

e) (3 points) Select the appropriate code to compute the p-value below.

Ⓐ `pt(test_statistic,df = 15,lower.tail = TRUE)`

Ⓑ `2*pt(abs(test_statistic),df = 15,lower.tail = TRUE)`

Ⓒ `2*pt(abs(test_statistic),df = 15,lower.tail = FALSE)`

Ⓓ `pt(test_statistic,df = 15,lower.tail = FALSE)`

Ⓔ `pt(test_statistic,df = 25.8734,lower.tail = TRUE)`

Ⓕ `2*pt(abs(test_statistic),df = 25.8734,lower.tail = TRUE)`

Ⓖ `2*pt(abs(test_statistic),df = 25.8734,lower.tail = FALSE)`

Ⓗ `pt(test_statistic,df = 25.8734,lower.tail = FALSE)`

f) **(6 points)** The $\boldsymbol{p}$**-value** for the correct test was found to be 0.1805. Using a significance level of $\boldsymbol{\alpha = 0.1}$, state your formal decision and write a conclusion in the context of the problem.

Formal Decision: $\boldsymbol{p}$**-value** $= \mathbf{0.1805 > 0.1} = \boldsymbol{\alpha}$ **therefore we do not reject the null** hypothesis.

Conclusion: The data **does not** give support ($\boldsymbol{p}$**-value** = $\mathbf{0.1805}$) to the claim that the true mean difference in moisture levels produced by roasting from the two machines differs from 0.

4. **(28 points)** Jamie owns a small cranberry farm that primarily grows the Stevens variety, which is known for being sweeter and less tart than Early Black. Recently, he planted a small patch of Early Black cranberries for his daughter, who loves tart berries. After a few years, some regular customers have claimed that the Stevens cranberries have become more tart.
Jamie wants to test whether the presence of Early Black cranberries has caused an increase in tartness of his cranberries. Industry standards indicate that pH measurements from Stevens cranberry batches have an average pH of 2.6 with a standard deviation of 0.3. Research indicates that most people can detect a pH change of 0.2 (lower pH = more tart).

a) **(2 points)** Provide the first two steps of the four-step hypothesis testing procedure.

Step 1: Define the parameter(s) of interest.
$\boldsymbol{\mu}$ representing the true population mean tartness (pH)

Step 2: Define the hypothesis:
$\boldsymbol{H_0: \mu \geq 2.6}$
$\boldsymbol{H_a: \mu < 2.6}$

b) **(2 points)** Identify the mean and standard deviation of the sampling distribution of $\bar{X}$ under the null and alternative hypotheses to detect a pH change of 0.2. Since the sample size is currently unknown you may use n to represent it.

Under null: $\bar{X} \sim N\left(\mu_0 = 2.6 \quad , \sigma_{\bar{X}} = \frac{0.3}{\sqrt{n}}\right)$

Under alternative: $\bar{X} \sim N\left(\mu_a = 2.4 \quad , \sigma_{\bar{X}} = \frac{0.3}{\sqrt{n}}\right)$

c) **(10 points)** Calculate the minimum sample size required to achieve **90% statistical power** for detecting a pH difference of 0.2 in the direction that would indicate increased tartness at $\alpha = 0.04$. Show your work and clearly identify both: (i) the critical z-value for your significance level, and (ii) the critical z-value for your desired power.

`> qnorm(0.01, lower.tail = FALSE)` `[1] 2.326348`	`> qnorm(0.02, lower.tail = FALSE)` `[1] 2.053749`
`> qnorm(0.04, lower.tail = FALSE)` `[1] 1.750686`	`> qnorm(0.05, lower.tail = FALSE)` `[1] 1.644854`
`> qnorm(0.1, lower.tail = FALSE)` `[1] 1.281552`	`> qnorm(0.2, lower.tail = FALSE)` `[1] 0.8416212`

$z_\alpha = 1.750686$

$z_\beta = 1.281552$

$\bar{x}_{\text{cutoff}} = \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}} = 2.6 - 1.750686 \cdot \frac{0.3}{\sqrt{n}}$

Power: $P(\bar{X} < \bar{x}_{\text{cutoff}} | \mu = 2.4) = 0.9 = P(Z < z_\beta)$

This yields:

$\frac{\bar{x}_{\text{cutoff}} - \mu_a}{\sigma/\sqrt{n}} = z_\beta$

$\frac{\mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}} - \mu_a}{\sigma/\sqrt{n}} = z_\beta$

$$n = \left\lceil \left(\frac{z_\beta + z_\alpha}{(\mu_0 - \mu_a)/\sigma}\right)^2 \right\rceil = \left\lceil \left(\frac{1.281552 + 1.750686}{0.2/0.3}\right)^2 \right\rceil = 21$$

d) **(8 points)** Assume Jamie collects a random sample of **25** cranberry batches and measures the average pH of each batch. Determine the cutoff pH value ($\bar{x}_{\text{cutoff}}$) corresponding to a significance level $\alpha = 0.04$. Assume the population standard deviation remains $\sigma = 0.3$ (unchanged from the industry standard). Show your work.

`> qnorm(0.01, lower.tail = FALSE)` `[1] 2.326348`	`> qnorm(0.02, lower.tail = FALSE)` `[1] 2.053749`
`> qnorm(0.04, lower.tail = FALSE)` `[1] 1.750686`	`> qnorm(0.05, lower.tail = FALSE)` `[1] 1.644854`
`> qnorm(0.1, lower.tail = FALSE)` `[1] 1.281552`	`> qnorm(0.2, lower.tail = FALSE)` `[1] 0.8416212`

$$z_\alpha = 1.750686$$

$$\bar{x}_{\text{cutoff}} = \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}} = 2.6 - 1.750686 \cdot \frac{0.3}{\sqrt{25}} = 2.4950$$

e) **(3 points)** Which of the following R code correctly computes the statistical power of the test?

Ⓐ `pnorm((cutoff-2.6)/0.3, lower.tail = TRUE)`

Ⓑ `pnorm((cutoff-2.6)/0.06, lower.tail = TRUE)`

Ⓒ `pnorm((cutoff-2.6)/0.06, lower.tail = FALSE)`

Ⓓ `pnorm((cutoff-2.4)/0.3, lower.tail = FALSE)`

Ⓔ `pnorm((cutoff-2.4)/0.06, lower.tail = TRUE)`

Ⓕ `pnorm((cutoff-2.4)/0.06, lower.tail = FALSE)`

f) **(3 points)** Which of the following interventions will improve the statistical power of the test, assuming all other factors remain constant? **Select all that apply.**

Ⓐ Plant more Early Black cranberries on the farm.

Ⓑ Randomly sample more bags of cranberries from the farm.

Ⓒ Move Early Black bushes to a greenhouse with its own beehive.

Ⓓ Use a more precise pH measuring device.

5. **(23 points)** South Africa contributes significantly to the global dairy market. The most productive provinces (**States**) are Limpopo and Mpumalanga. Limpopo cows average 27.3 liters of milk a day, with a standard deviation of 2.4 liters. For Mpumalanga cows, the mean daily production is 25.0 liters, with a standard deviation of 3.2 liters. Assume that the milk production for these provinces follow normal distributions.

a) **(2 points)**

Ⓣ or Ⓕ We require sample sizes of at least 30 for the sampling distribution of the average daily milk production in both provinces to be approximately Normal.

b) **(8 points)** A random sample of 20 Limpopo cows was selected for a study and an average of 31 liters of milk per day was recorded. How many standard deviations is this average away from its population mean?

$$z = \frac{31 - 27.3}{2.4/\sqrt{20}} = 6.8945$$

c) **(10 points)** A Mpumalanga farmer has 20 cows. There is a 50% chance each day that **the total daily production from this herd** is at most how many liters? Justify your answer.

The sampling distribution of the total daily production from this heard of 20 cows from Mpumalanga would follow a normal distribution:

$$T = \sum_{i=1}^{20} X_i \sim N(\mu_T, \sigma_T^2)$$

where $\mu_T = n \cdot \mu = 20 \cdot 25 = 500$ and $\sigma_T^2 = n \cdot \sigma^2 = 20 \cdot 3.2^2 = 204.8$.

Since the 50th percentile of a normal distribution is the same as the population mean we find that there is a 50% chance that each day the total production from a herd of 20 cows from Mpumalanga would produce **500 liters** of milk.

d) **(3 points)** A Limpopo farmer has 20 cows. What is the probability that the **average milk production** for this herd **exceeds** 38 liters a day.

Ⓐ `pnorm((38-27.3)/2.4, lower.tail=FALSE)`

Ⓑ `1-pnorm((38-25)/3.2, lower.tail=TRUE)`

Ⓒ `1-pnorm((38-27.3)/0.5367, lower.tail=TRUE)`

Ⓓ `pt((38-27.3)/0.5367, df=19, lower.tail=FALSE)`