STAT 350 Final Exam 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 sheets (2)**, 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.
6. The crib sheet can be a handwritten or type double-sided 8.5in x 11in sheet.
7. Keep your bag closed and cellphone always stored away securely 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 120 minutes long. If you need a bathroom break, raise your hand to request approval. Breaks will only be permitted when an escort is available, and you must confirm that no cell phones or electronic devices are in your possession before leaving the room.
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 (digital)**.

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)	18	
Problem 2 (Multiple Choice) (3 points each)	15	
Problem 3	16	
Problem 4	30	
Problem 5	42	
Problem 6	44	
Total	**150+15** (Extra Credit) = **165**	

1. (18 points, 2 points each) True/False Questions.

1.1. Given a dataset that contains multiple real outliers,

Ⓣ or Ⓕ the best measures of spread for this data would be the interquartile range (IQR).

1.2. Let $\boldsymbol{X}$ follows a binomial distribution with fixed p and sufficiently large number of trials $\boldsymbol{n}$. The estimator $\widehat{\boldsymbol{p}} = \frac{\boldsymbol{X}}{\boldsymbol{n}}$, representing the sample proportion of successes, is derived from the random variable $\boldsymbol{X}$, which is the sum of $\boldsymbol{n}$ independent and identically distributed Bernoulli trials,

Ⓣ or Ⓕ the sampling distribution of $\widehat{\boldsymbol{p}} = \frac{\boldsymbol{X}}{\boldsymbol{n}}$ is approximately normal.

1.3. The time it takes for a customer to complete a transaction at a store follows an exponential distribution with a rate of $\boldsymbol{\lambda = 1.4}$ per minute.

Ⓣ or Ⓕ The probability that a transaction lasts less than 1 minute is greater than the probability that a transaction lasts between 1 and 3 minutes.

1.4. The time it takes for a customer to complete a transaction at a store is uniformly distributed between 2 and 10 minutes.

Ⓣ or Ⓕ The probability that a transaction lasts between 1 and 6 minutes is greater than the probability that a transaction lasts between 6 and 10 minutes.

1.5. A research team investigates whether consuming a spoonful of apple cider vinegar before meals prevents blood sugar spikes. They selected 30 pairs of identical twins, randomly assigning each twin in a pair to one of two groups.

Ⓣ or Ⓕ In this scenario, a two-sample independent procedure is appropriate to compare the groups.

1.6. In a one-way ANOVA analysis for a factor with nine levels, the F-test resulted in rejection of the null hypothesis. If all possible pairs of levels are to be compared,

Ⓣ or Ⓕ the Multiple Comparisons step would involve 72 paired comparisons.

1.7. The **98%** lower confidence bound on the true slope of a simple linear regression line, $\boldsymbol{\beta_1}$, gives the value of **-0.43**. Then

Ⓣ or Ⓕ there is **0.98** probability that $\boldsymbol{\beta_1}$ is greater than **-0.43**.

1.8. In simple linear regression,

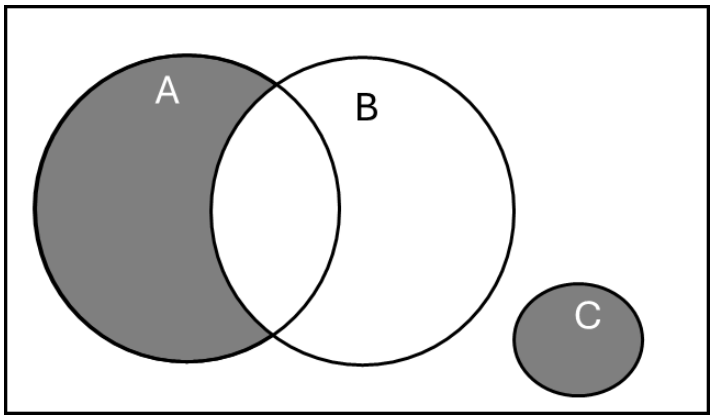
Ⓣ or Ⓕ all influential points must be outliers.

1.9. In simple linear regression,

Ⓣ or Ⓕ prediction intervals are wider than confidence intervals for the mean response at the same value of the predictor variable.

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. Select the expression that does **NOT** correctly represent the probability of the colored area in the Venn diagram shown below.



Ⓐ $\boldsymbol{P(A \cup B) - P(B) + P(C)}$

Ⓑ $\boldsymbol{P(A) - P(A \cap B) + P(C)}$

Ⓒ $\boldsymbol{P(A) - P(B) + P(C)}$

Ⓓ $\boldsymbol{P(A \cup C) - P(A \cap B)}$

Ⓔ $\boldsymbol{P(A \cup B \cup C) - P(B)}$

2.2. Fréchet distribution is a heavily skewed, right-tailed continuous distribution that is used for modeling extreme events such as earthquake magnitudes, daily rainfall totals, and large insurance claims. Which of the following statements is TRUE about Fréchet distribution?

Ⓐ The mean is the largest among the measures of central tendency, followed by the mode and the median.

Ⓑ A small sample size is adequate to apply the central limit theorem to the distribution of the sample mean.

Ⓒ For samples from this distribution, the median and variance are recommended measures of central tendency and spread, respectively.

Ⓓ The interquartile range (IQR) is preferred for describing the spread of a population with a Fréchet distribution because it is less sensitive to outliers.

Ⓔ None of the above statements are TRUE for the Fréchet distribution.

2.3. Suppose $\boldsymbol{X}$ is a random variable with $\boldsymbol{E[2^X] = 16, Var(X) = 32}$, and $\boldsymbol{E[3X + 2] = 8}$. Let a new random variable $\boldsymbol{Y}$ be defined as $\boldsymbol{Y = 2^X - \frac{1}{4}X^2}$**.** What is $\boldsymbol{E[Y]}$**?**

Ⓐ **0**

Ⓑ **7**

Ⓒ **8**

Ⓓ **36**

Ⓔ None of the above

2.4. An ANOVA F-test was performed on a dataset with two treatment levels ($\boldsymbol{k = 2}$), resulting in a test statistic of $\boldsymbol{f_{ts} = 3.28}$. If the same dataset is used for a hypothesis test on the difference of means of the two levels, then $\boldsymbol{t_{ts}^2 = 3.28}$ only if

Ⓐ The null value, $\boldsymbol{\Delta_0}$ is 0.

Ⓑ The hypothesis test is two-tailed $\boldsymbol{t}$-test.

Ⓒ The observations within and across the two levels are assumed to be independent.

Ⓓ The population variances are assumed to be equal, and the pooled variance estimate is used to construct the test statistic.

Ⓔ All of the above must hold simultaneously.

Ⓕ We cannot be certain because the $\mathbf{F_{TS}}$ and $\mathbf{T_{TS}}$ are test statistics for two different hypothesis tests, each with distinct assumptions and interpretations.

2.5. Which of the following statements is true for simple linear regression?

Ⓐ All diagnostics plots rely on the residuals/errors.

Ⓑ The assumption that the response is a simple random sample (SRS) for each fixed value of the explanatory variable is easy to verify.

Ⓒ A scatter plot can be used to assess both linearity and normality.

Ⓓ A scatter plot can be used to assess both homogeneity of variance and linearity.

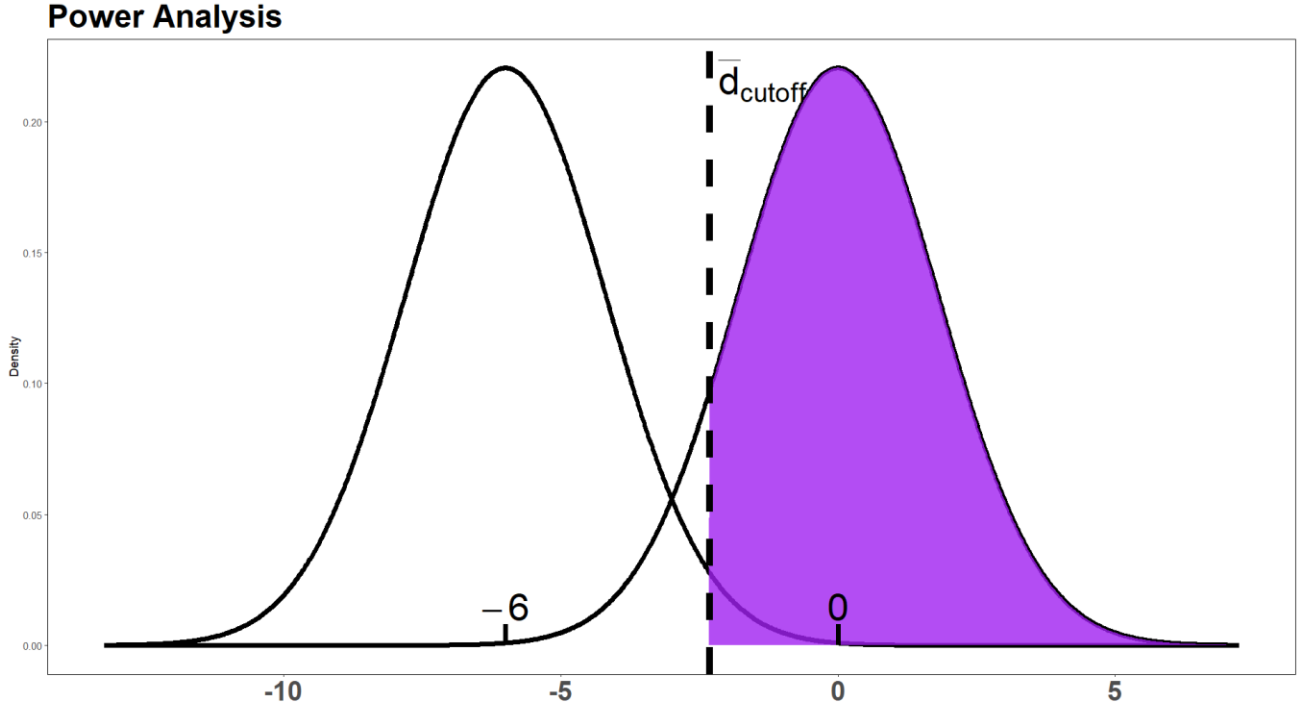
Ⓔ All of the above are true for simple linear regression.

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

3. **(16 points)** A dietitian is studying the effectiveness of a new dietary supplement on weight loss. To evaluate its impact, the dietitian measures the weight of **44 individuals** before and after a **4-week regimen** with the **supplement**.
   The difference in weight **(d = weight after- weight before)** is calculated for each participant. The dietitian wants to establish whether the supplement results in a **decrease in weight**. The dietitian would like the test to have **at least 90% power** to detect an **average weight decrease of 6 lbs**, which is deemed an important reduction.
   The test is conducted at a **significance level** of $\boldsymbol{\alpha = 0.05}$. This leads to the following hypotheses:

   $$\boldsymbol{H_0: \mu_d \geq 0}$$
   $$\boldsymbol{H_a: \mu_d < 0}$$

   For this paired test, the standard deviation of the differences, $\boldsymbol{\sigma_d}$, is assumed to be known with a value of **12 lbs**. To assist in this analysis, the dietitian asked an intern to generate the power graph for the test. The intern provided the graph below and shaded a portion of the graph to indicate what they believe to be the power of the test.



   a) (3 points) Select the correct option for what the purple shaded region in the graph actually represents.

   Ⓐ The intern is correct; it is the **power** of the test in detecting an alternative $\boldsymbol{\mu_{d_a} = -6.}$

   Ⓑ The intern is wrong; it is in fact the probability of **Type I error**.

   Ⓒ The intern is wrong; it is in fact the probability of **Type II error**.

   Ⓓ The intern is wrong, and it is none of the above options.

b) (13 points) Using the R output below, calculate the power of the test to detect an **average weight decrease of 6 lbs**. Determine whether the sample size is sufficient to meet the dietitian's requirement for **at least 90% power**.
Provide a detailed explanation of your calculations, including all steps and reasoning, to receive full credit. This includes computing $\bar{d}_{\text{cutoff}}$ and writing out full probability statements. Submitting only a formula and answer is not sufficient for full credit.

> qnorm (p=0.05, lower.tail = FALSE) **1.644854**	> qt (p=0.05, df = 43, lower.tail = FALSE) **1.681071**
> qnorm (p=0.1, lower.tail = FALSE) **1.281552**	> qt (p=0.1, df = 43, lower.tail = FALSE) **1.301552**
> pnorm(1.671771, lower.tail = TRUE) **0.9527153** > pnorm(1.671771, lower.tail = FALSE) **0.04728474**	> pnorm(-2.975653, lower.tail = TRUE) **0.001461827** > pnorm(-2.975653, lower.tail = FALSE) **0.9985382**
> pt(1.671771, df = 43, lower.tail = TRUE) **0.9490839** > pt(1.671771, df = 43, lower.tail = FALSE) **0.05091613**	> pt(-2.975653, df = 43, lower.tail = TRUE) **0.002391402** > pt(-2.975653, df = 43, lower.tail = FALSE) **0.9976086**

4. **(30 points)** Halin is a new student in STAT 350 and has never coded in R before. To improve her skills, she spends at most $\mathbf{40}$ **minutes** each weekday doing R self-study. Her daily workflow is as follows:
   - The time it takes until she runs into an error, denoted by **T**, follows Exponential distribution with **an average time of** $\mathbf{25}$ **minutes**.
   - When she sees an error, she tries **debugging,** which **succeeds with probability** $\mathbf{0.7}$.
     - If $\boldsymbol{T > 40}$, she does not run into an error during her study session that day.
     - If $\boldsymbol{T \leq 40}$, she encounters an error and attempts to debug it:
       - Debugging succeeds with probability $\mathbf{0.7}$, after which she feels happy and ends her study session for the day.
       - Debugging fails with probability $\mathbf{0.3}$, and she immediately stops her study session and goes to office hours for help.
   - Each day's workflow is independent of other days.

   a) (8 points) What is the probability that Halin will complete her study session on a given weekday without encountering an error?

   b) (12 points) On a given weekday, what is the probability that Halin does not need to go to office hours?

c) **(10 points)** Suppose Halin has continued her **independent** study for **20 weekdays**. Use the information from the previous question to determine the probability that Halin **will visit** office hours exactly **5 times** over the **20 weekdays**.

5. **(42 points)** A clothing retail company aims to boost profit during the upcoming holiday season, and its marketing team has decided to use four advertising strategies: an **Email Ad Campaign** (electronic email sent to customers), a **Direct Mail Ad Campaign** (personalized flyers and brochures sent to customers), one **Social Media Ad Campaign**, and one **AI-Powered Ad Campaign**. The last strategy, newly introduced by the marketing team, is based on the belief that customers are more likely to make purchases when they experience human-like interactions online. To test this hypothesis, **180 loyal customers** were randomly divided into four groups of 45 customers each. Each group was exposed to one advertising strategy, and their purchase amounts for the year were recorded. The summary information is provided below.

	Email Ad Campaign	Direct Mail Ad Campaign	Social Media Ad Campaign	AI-Powered Ad Campaign
n_i	45	45	45	45
$\bar{x}_i$	449.45	450.39	453.42	455.72
s_i^2	68.16	104.6542	112.3643	147.7942

a) **(3 points)** Which of the following assumptions is **NOT** required to perform one-way ANOVA? Assume a factor has k levels.

Ⓐ Population variances are equal across the k groups.

Ⓑ An independent sample is randomly drawn from each of the k groups.

Ⓒ Observations within each group are independent of observations in other groups.

Ⓓ The sample sizes from each of the k groups are the same.

Ⓔ The sample means are normally distributed for each of the k groups.

Ⓕ All of the above assumptions are required.

b) (2 points) Determine whether the homogeneity of variance assumption is valid or invalid. Mathematically support your answer.

c) (4 points) Clearly identify the factor of interest, specify how many levels this factor has, and describe what the quantitative response variable measures. Using this information, define the parameters of interest and state the null hypothesis and alternative hypothesis.

d) (12 points) Complete the ANOVA table.
Clearly, show your work in the box provided below.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F statistic
Factor	**3**			**3.424137**
Error			**108.2432**	
Total				

e) (7 points) The p**-value** was found to be $\mathbf{0.0185}$. Test your hypotheses at a significance level of $\alpha = \mathbf{0.05}$. Provide the formal decision and interpret the conclusion in the context of the problem. You may assume all assumptions are valid.

f) (3 points) Based on your conclusion, determine whether you should proceed to conduct a Tukey HSD test. You may assume all assumptions have been met.

Ⓐ Conduct the Tukey HSD test because it can identify specific pairs of means that are significantly different when the ANOVA results show a significant difference.

Ⓑ Do not conduct the Tukey HSD test because the ANOVA results indicate that the population means are not significantly different.

Ⓒ There is insufficient information provided to decide whether a Tukey HSD test should be conducted.

g) (3 points) Regardless of your conclusion for part f) The researchers had decided to conduct a Tukey's HSD wherein the overall level of significance was fixed at 5%. Let $\mathbf{df_E}$ denote the correct degrees of freedom for error. Choose the correct Tukey's parameter from the following.

Ⓐ qtukey(0.95, nmeans = 3, df = dfe, lower.tail=TRUE) = 3.342793

Ⓑ qtukey(0.95/2, nmeans = 4, df = dfe, lower.tail=TRUE) = 1.925889

Ⓒ qtukey(0.95/2, nmeans = 3, df = dfe, lower.tail=TRUE) = 1.533567

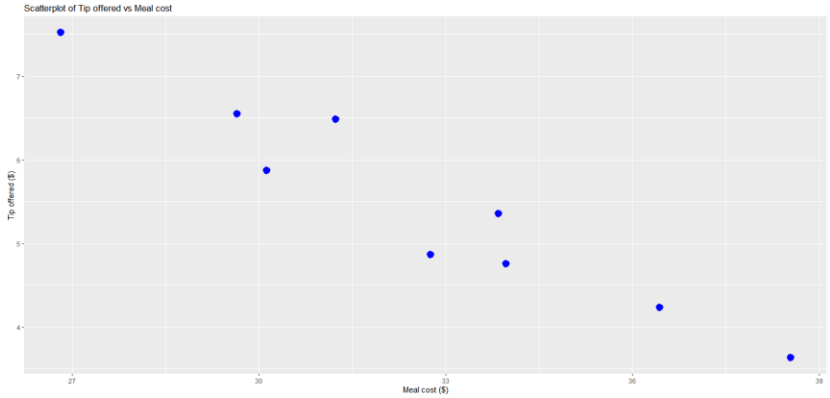
Ⓓ qtukey(0.95, nmeans = 4, df = dfe, lower.tail=TRUE) = 3.66811

h) (8 points) Using the summary information and the Tukey parameter above, construct a 95% confidence interval for the true difference in the average yearly amount spent by customers exposed to the **AI-powered Ad Campaign** and those exposed to the **Direct Mail Ad Campaign**.
Based on the confidence interval, determine whether there is statistically significant evidence of a difference in the average yearly amount spent by customers between these two groups. Justify your answer.

6. **(44 points)** A tip is often considered an expression of gratitude to waitstaff for the service provided. A STAT350 student sought to explore the relationship between the cost of a meal for a single diner (x) and the tip amount offered (Y). The student selected nine specific meal costs and, for each cost, recorded a randomly selected tip amount from diners at a restaurant in Greater Lafayette.

Meal cost ($)	33.85	31.24	26.82	38.54	33.97	36.44	30.13	29.65	32.76
Tip ($)	5.35	6.48	7.52	3.63	4.75	4.23	5.87	6.55	4.86

a) **(5 points)** Describe the relationship between meal cost and the tip amount based on the scatter plot below.



b) (15 points) The following summary statistics were realized from the data above;

$\boldsymbol{n = 9}$ $\quad \boldsymbol{\bar{x} = 32.600}$ $\quad \boldsymbol{\bar{y} = 5.4711}$

$\boldsymbol{\sum_{i=1}^{9} x_i y_i = 1570.9020}$ $\quad \boldsymbol{\sum_{i=1}^{9} x_i^2 = 9668.3960}$ $\quad \boldsymbol{\sum_{i=1}^{9} y_i^2 = 281.7746}$

(i) Determine the slope $\boldsymbol{(b_1)}$ of the least-squares regression line.

(ii) Determine the intercept $\boldsymbol{(b_0)}$ of the least squares regression line.

(iii) Write out the equation of the regression line.

c) (6 points) List the assumptions of simple linear regression that can be evaluated using diagnostic plots. For each assumption, specify all the diagnostic plots that can be used to assess it. To receive full credit, you must include all relevant plots for each assumption.

d) (18 points) The following output was obtained using RStudio for the tip-meal cost data. You may assume in what follows that all assumptions have been met for simple linear regression.

```
Residual standard error: 0.3783 on 7 degrees of freedom
Multiple R-squared: 0.9191, Adjusted R-squared: 0.9075
F-statistic: 79.49 on 1 and 7 DF, p-value: 4.534e-05
```

(i) Interpret the coefficient of determination, $\boldsymbol{R^2}$, given in the output above.

(ii) Use the output above and your results in **(b)** to compute the Pearson correlation coefficient ($\boldsymbol{r}$).

(iii) Inference for the parameters, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ is crucial to understanding the baseline response and the effect of the explanatory variable on the mean response.

For the test on the slope of the mean response line $\boldsymbol{\beta_1}$:

- Use the output above to calculate the $\boldsymbol{t}$-test statistic.
- Specify the associated degrees of freedom.

(iv) Based on the R output above, perform a hypothesis test following the **four general steps** of hypothesis testing at $\boldsymbol{\alpha = 0.01}$ using the **F-test** procedure to determine whether there is a significant linear association between meal cost and tip offered. Provide a **formal decision** and **conclusion** in the context of the problem.