

Name (Last, First):

Lopez Sepulveda,
Kevin

Student ID:

U57902827

Assignment 11

METCS544A3A4_F2024

Instructions:

1. For answering programming questions, please use Adobe Acrobat to edit the pdf file in two steps **[See Appendix: Example Question and Answer]**:
 - a. Copy and paste your R code as text in the box provided (so that your teaching team can run your code);
 - b. Screenshot your R console outputs, save them as a .PNG image file, and paste/insert them in the box provided.
 - c. Show all work—credit will not be given for code without showing it in action, including a screenshot of R console outputs.
2. To answer non-programming questions, please type or handwrite your final answers clearly in the boxes. Show all work - credit will not be given for numerical solutions that appear without explanation in the space above the boxes. **You're encouraged to use R to graph/plot the data and produce numerical summaries; please append your code and screenshot of the outputs at the end of your PDF submission.**

[Total 99 pts = 84 pts + 15 Extra Credit pts]

Grading Rubric

Each question is worth 3 points and will be graded as follows:

3 points: Correct answer with work shown

2 points: Incorrect answer but attempt shows some understanding (work shown)

1 point: Incorrect answer but an attempt was made (work shown), or **correct answer without explanation (work not shown)**

0 points: Left blank or made little to no effort/work not shown

Reflective Journal [3 pts]

(Copy and paste the link to your live Google doc in the box below)

https://drive.google.com/drive/folders/1_8qcBjQVMfZggF42UYJuHQzMoBcyAy0Q?usp=drive_link

Part I. The Normal Distribution and Combining Normal Random Variables (4 x 3 = 12 pts)

1) Consider a set of 9000 scores on a national test that is known to be approximately normally distributed with a mean of 500 and a standard deviation of 90.

(a) What is the probability that a randomly selected student has a score greater than 600?

[Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

Answer:

13.4%

$$z = (x - \mu) / \sigma$$

$$z = 1.111$$

$$P(Z > 1.111) = 1 - P(Z \leq 1.111) = 1 - 0.866 = 0.134$$

$$13.4\%$$

(b) How many scores are there between 450 and 600? [Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

Answer:

5193scores

$$z = (x - u) / \sigma$$

$$\text{for } x = 450: \approx -0.556$$

$$\text{for } x = 600: \approx 1.111$$

$$P(450 \leq X \leq 600) = P(Z \leq 1.111) - P(Z \leq -0.556)$$

$$P(450 \leq X \leq 600) = 0.866 - 0.289 = 0.577$$

$$57.7\% * 900 = 5193$$

(c) Megan needs to be in the top 1% of the scores on this test to qualify for a scholarship. What is the minimum score Megan needs? [Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

Answer:

710

$$z = 2.33$$

$$x = u + z * \sigma$$

$$X = 500 + (2.33 * 90) = 500 + 209.7 = 709.7$$

$$710$$

2) Gus and Cal go bowling every week. Gus's scores are normally distributed with a mean of 175 pins and a standard deviation of 30 pins. Cal's scores are normally distributed with a mean of 150 pins and a standard deviation of 40 pins. Assume that their scores in any given game are independent. Let G be Gus's score in a random game, C be Cal's score in a random game, and D be the difference between Gus's and Cal's scores where $D = G - C$. What is the probability that Gus will knock down more pins than Cal? [Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

Answer:

69.2%

$$D = G - C$$

$$\mu_D = \mu_G - \mu_C$$

$$\mu_D = 175 - 150 = 25$$

$$\sigma(D) = 50$$

$$P(D > 0) = 0.6915$$

$$P(Z > -0.5) = 1 - P(Z \leq -0.5) = 1 - 0.3085 = 0.6915$$

$$69.2\%$$

Part II. Sampling Distribution of Sample Proportions (8 x 3 = 24 pts)

1) (12 pts) Suppose a large candy machine has 15% orange candies. Imagine taking an SRS of 25 candies from the machine and observing the sample proportion \hat{p} of orange candies. [Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

(a) What is the mean of the sampling distribution of \hat{p} ? Why?

Answer:

up=0.15
0.15

(b) Find the standard deviation of the sampling distribution of \hat{p} . Check to see if the 10% condition is met.

Answer:

sigma(p) \approx 0.0714
10% Condition: The condition is met if the population size
 \geq
250
 $N \geq 250$.

(c) Is the sampling distribution of \hat{p} approximately Normal? Check to see if the Large Counts condition is met.

Answer:

$n \cdot p = 3.75$, which is less than 10, so the condition is not satisfied.
 $n \cdot (1-p) = 21.25$, which is greater than 10.
the Large Counts condition is not met, and the sampling distribution of \hat{p} is not

(d) If the sample size were 225 rather than 25, how would this change the sampling distribution of \hat{p} ?

Answer:

If the sample size were $n=225$:
The mean remains the same (0.15).
The standard deviation decreases (≈ 0.0238).
The shape of the sampling distribution becomes approximately Normal because the Large Counts condition is met.

2) (12 pts) The Harvard College Alcohol Study finds that 67% of college students support efforts to "crack down on underage drinking". Does this result hold at a large local college? To find out, college administrators surveyed an SRS of 100 students and found that 62 support a crackdown on underage drinking. What is the probability that the proportion in an SRS of 100 students is 0.62 or less? Does this refute the claim made by Harvard? **Explain using the 4 step process. Answer:**

Step1: State

We aim to find the probability of observing a sample proportion $\hat{p} \leq 0.62$ if $p = 0.67$

Step2: Plan

$n \cdot p \geq 10$ and $n \cdot (1-p) \geq 10$

$n \cdot p = 100 \cdot 0.67 = 67$

$n \cdot (1-p) = 100 \cdot 0.33 = 33 (\geq 10)$

$p = 0.67$

≈ 0.047

Step3: Do

We calculate the z-score for $\hat{p} = 0.62$

≈ -1.06

$P(Z \leq -1.06) \approx 0.145$

$P(\hat{p} \leq 0.62) \approx 0.145$

Step 4: Conclude

The probability of obtaining a sample proportion of 0.62 or less under the assumption

=

0.67

$p = 0.67$ is approximately 0.145 (14.5%).

This probability is not very small (e.g., it is greater than the common significance level of 0.05), so we do not have strong evidence to refute Harvard's claim that

=

0.67

$p = 0.67$.

The observed sample proportion of 0.62 could reasonably occur due to random sampling variability.

Part III. Sampling Distribution of a Difference in Sample Proportions (12 pts)

In a single town, there are two high schools: North and South (each high school has more than 2000 students in it). At North High School, the principal says that they have 18% of students arriving tardy each day. At South High School, the principal claims that they have 22% of students arriving tardy each day. The North principal plans to take an SRS of 250 students to see how many are tardy, and the South principal plans to take an SRS of 200 students to see how many are tardy. Let p_N represent the proportion of students arriving late at North and p_S represent the proportion of students arriving late at South.

(a) Describe the shape, center, and spread of the sampling distribution of $\hat{p}_N - \hat{p}_S$.

Answer:

Shape: Approximately Normal.

Center: The mean is
-0.04.

Spread: The standard deviation is approximately
0.0381.

(b) The principals say that if the proportion of students tardy in South's sample is 10% or higher than the proportion of students tardy in North's sample, they will implement a new tardy policy. What is the probability of this happening? Use the sample distribution you created in part (a).

Answer:

The probability of this happening is approximately 0.058 (or 5.8%)

Part IV. Sampling Distribution of Sample Means (2 x 6 = 12 pts)

1) (6 pts) The Wechsler Adult Intelligence Scale (WAIS) is a common “IQ test” for adults. The distribution of WAIS scores for people over 16 years of age is approximately normal with mean 100 and standard deviation 15. What is the probability that the average WAIS score of an SRS of 60 people is 105 or higher? [Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

Answer:

To determine the probability that the average WAIS score of a simple random sample (SRS) of 60 people is 105 or higher, we calculate the sampling distribution of the sample mean. The sampling distribution has the same mean as the population ($\mu = 100$) and a standard error of $\sigma = \sigma/\sqrt{n} = 15/\sqrt{60} \approx 1.937$. Standardizing the sample mean of 105 yields a Z-score of $Z = (x - \mu)/\sigma = (105 - 100)/1.937 \approx 2.58$. Using the standard normal distribution, the probability of $Z \geq 2.58$ is $1 - P(Z \leq 2.58)$. From the Z-table, $P(Z \leq 2.58) \approx 0.9951$, so $P(Z \geq 2.58) = 1 - 0.9951 = 0.0049$. Therefore, the probability that the average WAIS score for the sample is 105 or higher is approximately 0.0049.

2) (6 pts) A company that owns and services a fleet of cars for its sales force has found that the service lifetime of disc brake pads varies from car to car according to a Normal distribution with mean $\mu = 55,000$ miles and standard deviation $\sigma = 4500$ miles. The company installs a new brand of brake pads on 8 cars. The average life on the pads on these 8 cars turns out to be 51,800 miles. What is the probability that the sample mean lifetime is 51,800 miles or less if the lifetime distribution is unchanged? (The company takes this probability as evidence that the average lifetime of the new brand of pads is less than 55,000 miles.) [Write your numerical answer in the box, and your work of deriving and calculating in the empty space]

Answer:

To calculate the probability that the sample mean lifetime of the new brake pads is 51,800 miles or less, we assume the population distribution remains unchanged. The lifetime of brake pads follows a normal distribution with a mean (μ) of 55,000 miles and a standard deviation (σ) of 4,500 miles. For a sample of size $n = 8$, the sampling distribution of the sample mean (X) has a mean $\mu_X = 55,000$ and a standard error $\sigma_X = \sigma/\sqrt{n} = 4500/\sqrt{8} \approx 1592.5$. To standardize the observed sample mean $X = 51,800$, we calculate the Z-score as $Z = (X - \mu_X)/\sigma_X = (51,800 - 55,000)/1592.5 = -3,200/1592.5 \approx -2.01$. Using the standard normal distribution, the probability $P(Z \leq -2.01)$ is approximately 0.0217. Therefore, the probability that the sample mean lifetime is 51,800 miles or less is approximately 0.0217.

Part V. Sampling Distribution of a Difference in Sample Means (24 pts)

1) **(18 pts)** For the following situations, match the distribution type with the corresponding situation. Each distribution will be used only once.

- A. Normal Probability Distribution
- B. Combining Normal Random Variables Distribution
- C. Sampling Distribution for Sample Proportions
- D. Sampling Distribution for a Difference in Sample Proportions
- E. Sampling Distribution for Sample Means
- F. Sampling Distribution for a Difference in Sample Means

 F 1) A researcher is comparing the effectiveness of two different medications in reducing blood pressure. They randomly assign 50 patients to receive Medication A and 50 patients to receive Medication B. What is the probability that the difference in the mean reduction in blood pressure between the two groups is less than 5 mmHg?

 B 2) In a factory, the weights of products produced follow a normal distribution with a mean of 500 grams and a standard deviation of 20 grams. If a package contains 10 of these products, what is the probability that the total weight of the package is less than 4950 grams?

 E 3) In a study on the heights of sunflowers, it's found that the heights follow a normal distribution with a mean of 150 centimeters and a standard deviation of 20 centimeters. A researcher randomly selects a sample of 25 sunflowers from a field. What is the probability that the average height of the sample is greater than 155 centimeters

 A 4) A company manufactures light bulbs with a mean lifespan of 800 hours and a standard deviation of 50 hours. What is the probability that a randomly selected light bulb will last more than 850 hours.

 D 5) In a study comparing the effectiveness of two different advertising strategies, Strategy A and Strategy B, a marketing team wants to determine if there's a significant difference in the proportion of customers who make a purchase after seeing the ad. In a sample of 200 customers exposed to Strategy A, 40 customers made a purchase. In a sample of 250 customers exposed to Strategy B, 65 customers made a purchase. What is the probability that the difference in the proportion of customers who made a purchase between the two strategies is greater than 0.05?

 C 6) A survey was conducted to see the proportion of smartphone users who prefer iPhone. A research team randomly selects 800 smartphone users. Among them, 520 users prefer iPhone. What is the probability that in another random sample of 800 smartphone users, the proportion preferring iPhone is less than 0.6?

2) (6 pts) In a clinical trial investigating the effectiveness of a new medication for allergies in dogs, 60 dogs were randomly assigned to two groups: 30 dogs went to Group M, where they received the new medication, while 30 dogs went to Group P, where they received a placebo. After a month of treatment, the severity of allergy symptoms in each dog was assessed using a standardized scale. The mean score for Group M was 3.5 units with a standard deviation of 1.2 units, while the mean score for Group P was 4.2 units with a standard deviation of 1.5 units. A higher score means that more allergy symptoms were observed. What is the probability that the difference in the mean scores between the two groups (Group M – Group P) is greater than 0?

Answer:

To calculate the probability that the difference in mean scores between Group M and Group P ($\mu_M - \mu_P$) is greater than 0, we use the sampling distribution for the difference in sample means. The sample sizes for both groups are $n_M = 30$ and $n_P = 30$. The means and standard deviations are $\bar{X}_M = 3.5$, $\bar{X}_P = 4.2$, $s_M = 1.2$, and $s_P = 1.5$. The standard error for the difference in sample means is calculated as $SE = \sqrt{(s_M^2/n_M + s_P^2/n_P)} = \sqrt{(1.2^2/30 + 1.5^2/30)} = \sqrt{(1.44/30 + 2.25/30)} = \sqrt{(0.048 + 0.075)} = \sqrt{0.123} \approx 0.35$. The observed difference in sample means is $\bar{X}_M - \bar{X}_P = 3.5 - 4.2 = -0.7$. To standardize, we calculate the Z-score: $Z = ((\bar{X}_M - \bar{X}_P) - 0)/SE = (-0.7 - 0)/0.35 = -2.0$. Using the standard normal distribution table, the probability $P(Z \leq -2.0)$ is approximately 0.0228. Since we are looking for the probability that the difference is greater than 0, we use $P(Z > 0) = 1 - P(Z \leq -2.0) = 1 - 0.0228 = 0.9772$. Therefore, the probability that the difference in mean scores between Group M and Group P is greater than 0 is approximately 0.9772.

Part VI. Extra Credit: Statistical Programming (12 pts)

Scores and Grades

Initialize the scores of 100 students as shown below:

```
scores <- read.csv("scores.csv")
```

- a) Show the default histogram of the student scores. Save the result of the histogram into a variable. Using only the **counts** and **breaks** property of this variable, write the R code to produce the following output. The code for the following output should not refer to the individual scores.

```
3 students in range (35,40]
4 students in range (40,45]
10 students in range (45,50]
13 students in range (50,55]
17 students in range (55,60]
27 students in range (60,65]
13 students in range (65,70]
8 students in range (70,75]
3 students in range (75,80]
2 students in range (80,85]
```

- b) Using the **breaks** option of the histogram, show the histogram and the custom output as shown below so that students in the range (70,90] get an A grade, (50,70] get a B grade, and (30-50] get a C grade. The code for the following output should not refer to the individual scores.

```
17 students in C grade range (30,50]
70 students in B grade range (50,70]
13 students in A grade range (70,90]
```

Answer: Copy and paste your R code in the box below (not an image but the text).

```
library(readr)
scores <- read_csv("CS544-BU/scores.csv")
View(scores)

histogram_result <- hist(scores$Score, plot = FALSE)

histogram_result

for (i in 1:length(histogram_result$counts)) {
  cat(histogram_result$counts[i], "students in range",
      "(", histogram_result$breaks[i], ",", histogram_result$breaks[i + 1], "]\n")
}

custom_breaks <- c(30, 50, 70, 90, max(scores$Score))

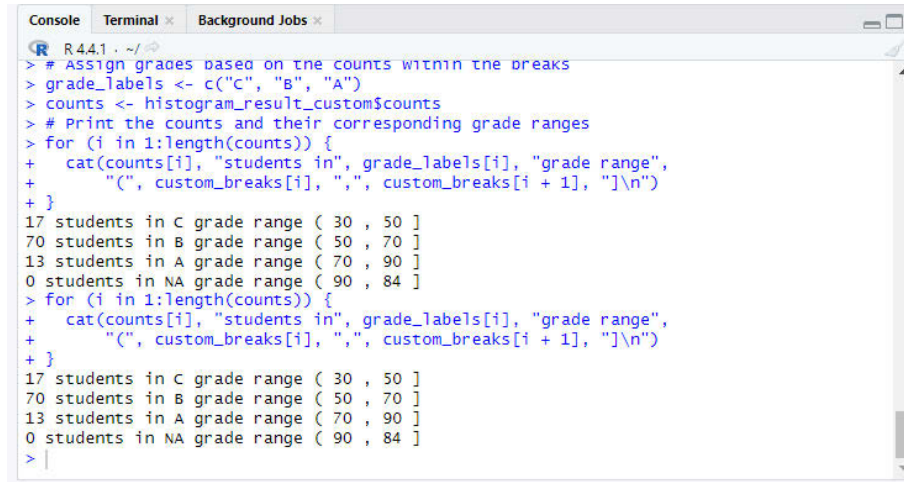
histogram_result_custom <- hist(scores$Score, breaks = custom_breaks, plot = FALSE)

histogram_result_custom

grade_labels <- c("C", "B", "A")
counts <- histogram_result_custom$counts

for (i in 1:length(counts)) {
  cat(counts[i], "students in", grade_labels[i], "grade range",
      "(", custom_breaks[i], ",", custom_breaks[i + 1], "]\n")
}
```

Screenshot of your R console outputs and paste the image in the box below



```
R 4.4.1. ~/
> # Assign grades based on the counts within the breaks
> grade_labels <- c("C", "B", "A")
> counts <- histogram_result_custom$counts
> # Print the counts and their corresponding grade ranges
> for (i in 1:length(counts)) {
+   cat(counts[i], "students in", grade_labels[i], "grade range",
+       "(", custom_breaks[i], ",", custom_breaks[i + 1], "]\n")
+ }
17 students in C grade range ( 30 , 50 ]
70 students in B grade range ( 50 , 70 ]
13 students in A grade range ( 70 , 90 ]
0 students in NA grade range ( 90 , 84 ]
> for (i in 1:length(counts)) {
+   cat(counts[i], "students in", grade_labels[i], "grade range",
+       "(", custom_breaks[i], ",", custom_breaks[i + 1], "]\n")
+ }
17 students in C grade range ( 30 , 50 ]
70 students in B grade range ( 50 , 70 ]
13 students in A grade range ( 70 , 90 ]
0 students in NA grade range ( 90 , 84 ]
> |
```

Appendix: Example Question and Answer for R programming questions:

Calculate the sum $\sum_{j=0}^n r^j$, where r has been assigned the value 1.08, and compare with $(1 - r^{n+1})/(1 - r)$, for $n = 10, 20, 30, 40$.

Answer: Copy and paste your R code in the box below (not an image but the text).

```
r <- 1.08
n <- c(10, 20, 30, 40)
sum1 <- c()
for(i in n){
  x <- 0:i
  sum1 <- c(sum1, sum(r^x))
}
sum1 # This gives the calculated sums for n = 10, 20, 30, 40.

sum2 <- (1 - r^(n + 1)) / (1 - r)
sum2

sum2 - sum1 # The formula works.
```

Screenshot of your R console outputs and paste the image in the box below

```
> r <- 1.08
> n <- c(10, 20, 30, 40)
> sum1 <- c()
> for(i in n){
+   x <- 0:i
+   sum1 <- c(sum1, sum(r^x))
+ }
> sum1 # This gives the calculated sums for n = 10, 20, 30, 40.
[1] 16.64549 50.42292 123.34587 280.78104
> sum2 <- (1 - r^(n + 1)) / (1 - r)
> sum2
[1] 16.64549 50.42292 123.34587 280.78104
> sum2 - sum1 # The formula works.
[1] 0 0 0 0
```

THE END