Name (Last, First):	NEU ID:
Kunjangada Arun Sushma	002473132

Assignment 10

NEU_COE_INFO6105_Fall2024

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 or Python code as text in the box provided (so that your teaching team can run your code);
    b. Screenshot your R or Python 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 or Python 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 or Python 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.**
3. **[Total 78 pts = 75 pts + 3 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://docs.google.com/document/d/1ptEhnYHniNtT1yxDPcvXK7LpJaPGzi80BCSGZZhom7Y/edit?usp=sharing

I. Inference for a Difference in Proportions (6 x 3 = 18 pts)

Use the following situation to answer the questions that follow.

A driving school owner believes that Instructor A is more effective than Instructor B at preparing students to pass the state's driver's license exam. An incoming class of 100 students in randomly assigned to two groups, each of size 50. One group is taught by Instructor A; the other is taught by Instructor B. At the end of the course, 30 of Instructor A's students and 22 of Instructor B's students pass the state exam. Let p_A be the proportion of Instructor A's students who pass the exam, and p_B be the proportion of Instructor B's students who pass the exam.

Because we are not randomly selecting from the population, we should not use the pooled population proportion. We are performing an experiment with randomly assigned groups, so we can use the pooled sample proportion.

| C |

1) Which of the following statements is FALSE?

(A) We use the pooled sample proportion because we are assuming the null hypothesis is true.
(B) We do not have to check that the sample size is less than 10% of the population size because we are performing an experiment.
(C) Because we are not randomly selecting from the population, we should use the pooled population proportion.
(D) The null hypothesis for this test would be H_0: $p_A - p_B = 0$
(E) The alternative hypothesis for this test would be H_a: $p_A - p_B > 0$

| E |

2) What is the z test statistic?

$$Z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

(A) $\dfrac{0.6-0.44}{\sqrt{\frac{.52(.48)}{100} + \frac{.52(.48)}{100}}}$

$\hat{p}_A = \frac{30}{50} = 0.6$ $\quad \hat{p}_B = \frac{22}{50} = 0.44$ $\quad \hat{p} = \frac{30+22}{100} = 0.52$

(B) $\dfrac{0.6-0.44}{\sqrt{\frac{.6(.4)}{100} + \frac{.44(.56)}{100}}}$

Substituting:

(C) $\dfrac{0.6-0.44}{\sqrt{\frac{.6(.4)}{50} + \frac{.44(.56)}{50}}}$

$$Z = \frac{0.6 - 0.44}{\sqrt{0.52(1-0.52)\left(\frac{1}{50} + \frac{1}{50}\right)}} = \frac{0.6 - 0.44}{\sqrt{(0.52)(0.48)\left(\frac{1}{50} + \frac{1}{50}\right)}}$$

(D) $\dfrac{0.6-0.44}{\sqrt{\frac{(.52)(.48)}{100}}}$

$$= \frac{0.16}{\sqrt{(0.52)(0.48)(0.04)}} = 1.601281538$$

(E) $\dfrac{0.6-0.44}{\sqrt{(.52)(.48)\left(\frac{1}{50} + \frac{1}{50}\right)}}$

$\simeq 1.601$

$$z = \frac{\hat{P}_A - \hat{P}_B}{\sqrt{\hat{P}(1-\hat{P})}\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$$

$\hat{P}_A = \frac{30}{50} = 0.6$, $\hat{P}_B = \frac{22}{50} = 0.44$, $\hat{P} = \frac{30+22}{100} = 0.52$

B. $\boxed{0.0547}$ 3) What is the p-value?

(A) 1.6012
(B) 0.0547
(C) 0.6
(D) 0.1094
(E) 0.52

$$z = \frac{0.6-0.44}{\sqrt{(0.52)(1-0.52)\left(\frac{1}{50}+\frac{1}{50}\right)}} = \frac{0.6-0.44}{\sqrt{(0.52)(0.48)\left(\frac{1}{50}+\frac{1}{50}\right)}}$$

$$= \frac{0.16}{\sqrt{(0.52)(0.48)(0.04)}} = 1.601281538 \approx 1.601$$

Null Hypothesis (H_0): $P_A - P_B = 0$ (No difference in proportions)
Alternative Hypothesis (H_A): $P_A - P_B > 0$ (Instructor A's proportion is greater)

$P(z > 1.601) = 1 - P(z \le 1.601) \approx 1 - 0.9453 = 0.0547$

B. 4) Do these results give convincing evidence at the 5% significance level that Instructor A is more effective?

$P = 0.0547$ $\alpha = 0.05$. If $P \le \alpha$, reject H_0 If $P > \alpha$, fail to reject H_0

(A) Because our p-value is lower than 5%, we reject H_0. We have convincing evidence that Instructor A is more effective.

$0.0547 > 0.05$ fail to reject H_0

(B) Because our p-value is higher than 5%, we fail to reject H_0. We do not have convincing evidence that Instructor A is more effective.
(C) Because our p-value is lower than 5%, we reject H_0. We have convincing evidence that Instructor B's students do not pass the state exam.
(D) Because our p-value is higher than 5%, we fail to reject H_0. We do not have convincing evidence that there is a difference between Instructor A and Instructor B.
(E) Because our p-value is lower than 5%, we reject H_0. We have convincing evidence that there is a difference between Instructor A and Instructor B.

D. 5) Which of the following errors could you have made from your conclusion above?

(A) Type I; saying that Instructor A was more effective when they were really not.
(B) Type II; saying that Instructor A was more effective when they were really not.
(C) Type I; saying that we do not have convincing evidence that Instructor A was more effective when they really were.
(D) Type II; saying that we do not have convincing evidence that Instructor A was more effective when they really were.
(E) None of the above errors are correct.

Type II Error:
- We fail to reject H_0 when H_0 is actually false.
- Failing to detect that Instructor A is truly more effective when that is indeed the case.

6) Create a 95% confidence interval for the difference in the proportion of students who pass the exam between Instructor A and Instructor B. Do we come to the same conclusion as we did in (4)? Why or why not?

Answer:

$$(\hat{p}_A - \hat{p}_B) \pm z^* \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}$$

z^* for 95% confidence ≈ 1.96

$$\hat{p}_A - \hat{p}_B = 0.6 - 0.44 = 0.16$$

Margin of error: $1.96 \times \sqrt{\frac{0.6(1-0.6)}{50} + \frac{0.44(1-0.44)}{50}}$

$$= 1.96 \times \sqrt{\frac{0.6 \times 0.4}{50} + \frac{0.44 \times 0.56}{50}} = 1.96 \times \sqrt{0.0048 + 0.004928}$$

$$= 0.193316023$$

$$\approx 0.193$$

Confidence Interval:

$$0.16 \pm 0.193 \Rightarrow (-0.033, 0.353)$$

Since the interval includes 0, there is no significant evidence to conclude that Instructor A is more effective at the 95% confidence level.

∴ The confidence interval aligns with the hypothesis test result in (4). There is no convincing evidence that Instructor A is more effective.

II. Confidence Interval for Means (4 x 3 = 12 pts)

A nutritionist wants to estimate the average amount of sugar in a certain brand of yogurt. She takes an SRS of 20 containers and measures the sugar content, in grams. She creates a 90% confidence interval for the true amount of sugar in this brand of yogurt: (11.42, 12.58).

(a) What is the point estimate from this sample?

Answer:

$$\text{Point estimate} = \frac{\text{Lower Bound} + \text{Upper Bound}}{2}$$

$$= \frac{11.42 + 12.58}{2} = \frac{24.00}{2} = 12.00$$

Point estimate = 12 grams.

(b) What is the margin of error?

Answer:

Margin of Error = Upper Bound − Point Estimate

= 12.58 − 12.00

= 0.58

Margin Error = 0.58 grams

(c) Interpret the 90% confidence interval in the context of the problem.

Answer:

We are 90% confident that the true average amount of sugar in this brand of yogurt is between 11.42 grams to 12.58 grams.

(d) Interpret the confidence level of 90% in the context of the problem.

Answer:

If we were to repeat this process many times (taking samples of 20 containers and constructing confidence intervals) about 90% of these intervals would contain the true mean sugar content of this brand of yogurt.

III. Significant Test for Means (4 x 3 = 12 pts)

Functional Fitness claims that its new exercise program helps their members lose 5 pounds in the first month. To test this claim, you select a random sample of 13 members who completed the program and record their weight loss (in pounds) as follows:

5.5	5.2	6	5.5	6.8	4	4.2	5.5	6.2	5.1	4.8	6.6	3.9

Is there evidence, at the 10% significance level, that this new exercise program helps members lose more than 5 pounds in their first month?

Answer:

Null Hypothesis $(H_0) \Rightarrow \mu = 5$ (Average weight loss is 5 pounds)

Alternative Hypothesis $(H_a) \Rightarrow \mu > 5$ (The average weight-loss is greater than 5)

Sample size $(n) = 13$

Sample mean $(\bar{x}) = \dfrac{\text{Sum of all values}}{n}$

$$\bar{x} = \frac{5.5+5.2+6+5.5+6.8+4+4.2+5.5+6.2+5.1+4.8+6.6+3.9}{13}$$

$$= \frac{69.3}{13} = 5.330769231 \approx \underline{5.33}$$

Standard deviation $(s) = \sqrt{\dfrac{\sum(x_i-\bar{x})^2}{n-1}} =$

$\sum(x_i-\bar{x})^2 = (5.5-5.33)^2+(5.2-5.33)^2+(6-5.33)^2+(5.5-5.33)^2+(6.8-5.33)^2+(4-5.33)^2+(4.2-5.33)^2+(5.5-5.33)^2+(6.2-5.33)^2+(5.1-5.33)^2+(4.8-5.33)^2+(6.6-5.33)^2+(3.9-5.33)^2$

$= 10.5077$

$$s = \sqrt{\frac{10.5077}{13-1}} = 0.9357572691 \approx \underline{0.936}$$

Test Statistics for one sample: $t = \dfrac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \dfrac{5.33-5}{\frac{0.936}{\sqrt{13}}} \approx \underline{1.27}$

For $t = 1.27$, p-value $= 1 - 0.8980 = 0.102$

pvalue $(0.102) > \alpha\,(0.10)$ we fail to reject the null Hypothesis

There is not enough evidence to support that the new excercise programs helps members lose more than 5 pounds in their first month.

IV. Margine of Error and Matched Pairs (5 x 3 = 15 pts)

1) Your local gym wants to estimate the mean amount of time (in hours) that their members spend each week at their center. They would like a 95% confidence interval with a margin of error of 30 minutes. Based on previous data, they know that the population standard deviation is 2.2 hours. What is the smallest sample size that will allow this confidence interval to be built?

Answer:

population standard deviation (σ) = 2.2 hours, confidence level = 95% ($z^* \simeq 1.96$)

$E = z^* \times \frac{\sigma}{\sqrt{n}}$

Margin of error ~~for~~ of 30 min $\Rightarrow \frac{30}{60}$ hours = 0.5 hours

$$n = \left(\frac{z^* \cdot \sigma}{E}\right)^2 \Rightarrow n = \left(\frac{1.96 \times 2.2}{0.5}\right)^2$$

$$= \left(\frac{4.312}{0.5}\right)^2 = (8.624)^2 = 74.373376 \simeq \underline{\underline{74.37}}$$

∴ We need atleast 75 people in our sample.

2) The number of hours college students spend studying per week follows a normal distribution. If you go three standard deviations out on either side of the mean, you get an interval of values from 10 hours to 34 hours. We want to estimate the mean study hours of a group of college students with 90% confidence and a margin of error of 2 hours. What is the sample size needed to get this confidence interval?

Answer:

Margin of Error (E) = 2 hours, Confidence level = 90% ($z^* \simeq 1.645$)

From the interval [10, 34], the range is 34 − 10 = 24

$$\sigma = \frac{\text{Range}}{6} = \frac{24}{6} = 4$$

$$n = \left(\frac{z^* \cdot \sigma}{E}\right)^2$$

$$n = \left(\frac{1.645 \times 4}{2}\right)^2 = \left(\frac{6.58}{2}\right)^2 = (3.29)^2 = 10.8241 \simeq \underline{\underline{10.82}}$$

∴ We need atleast 11 students

Use the following scenario for questions 3 – 5 ***[Mark your choice in the boxes provided].***

Does listening to positive affirmations tapes (for example, "I am in charge of my life") while sleeping decrease depression? Dr. Goodyear wants to test this theory! Dr. Goodyear randomly selected 7 of his patients with mild depression. He gives them a "test" to score their depression on a 100 point scale, where the higher the score, the more clinically depressed you are. He then assigns the 7 patients to listen to an affirmation tape for one month while sleeping. After one month, the patients are given the test again.

	Test Scores						
Person	A	B	C	D	E	F	G
Before	54	57	68	43	42	49	66
After	52	50	67	44	30	38	61

D.

3) Which of the following conditions must be met in order to use a *t*-procedure on these paired data? The t-test assumes normality of the differences, not the individual data sets.

(A) The distribution of both before times and after times must be approximately Normal.
(B) The distribution of before times and the distribution of differences (before – after) must be approximately Normal.
(C) Only the distribution of before times must be approximately Normal.
(D) Only the distribution of differences (before – after) must be approximately Normal.
(E) All three distributions—before, after, and the difference—must be approximately Normal.

E.

4) If μ_d represents the mean different in before – after, which of the following would be the appropriate hypothesis to test if positive affirmations make a student complete a puzzle faster?

Null Hypothesis (H_0): $\mu_d = 0$

(A) H_0: $\mu_d > 0$ and H_a: $\mu_d = 0$
(B) H_0: $\mu_d \neq 0$ and H_a: $\mu_d = 0$
(C) H_0: $\mu_d = 0$ and H_a: $\mu_d < 0$
(D) H_0: $\mu_d = 0$ and H_a: $\mu_d \neq 0$
(E) H_0: $\mu_d = 0$ and H_a: $\mu_d > 0$

The null hypothesis states there is no difference in puzzle completion times, meaning $\mu_d = 0$
Alternative Hypothesis (H_a): $\mu_d > 0$
If affirmations makes students faster, the After times are smaller than the Before times, so $\mu_d > 0$ (positive difference).

D.

5) You run a one sample t test on the mean difference in before time – after time and obtain a p-value of 0.0155. If you significance level is 0.05, which of the following is correct? $p = 0.0155$ $\alpha = 0.05$

Confidence level ≠ significance level $p < \alpha$

I. We are 95% confident that we can reject the null hypothesis.
II. We have convincing evidence there is a positive difference in test scores. ($p < 0.05$)
III. The probability of making a Type I error is 0.0155 (Type I error rate $= \alpha = 0.05$)

(A) I and III only (B) I and II only (C) II and III only (D) II only (E) I, II, and III

V. Difference in Means (6 x 3 = 18 pts)

Do high school girls and boys differ in how many hours they get per night? At a large high school, 35 girls were randomly selected and they got an average of 6.7 hours of sleep the night before, with a standard deviation of 0.65 hours. 40 boys were randomly selected and they got an average of 6.4 hours of sleep the night before, with a standard deviation of 0.45 hours. Construct a 90% confidence interval and conduct a two-sided significance test at the 10% significance level. Do these two results support each other? Explain.

Answer:

Girls: Sample size(n_1) = 35, $\bar{x}_1 = 6.7$ hours, $s_1 = 0.65$ hours

Boys: $n_2 = 40$, $s_2 = 0.45$ hours, $\bar{x}_2 = 6.4$ hours

Confidence level: 90%

Significance level for hypothesis test: $\alpha = 0.10$

Since 2 tailed interval, $\alpha/2 = 0.10/2 = 0.05$

$$CI = \boxed{(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot SE}$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(0.65)^2}{35} + \frac{(0.45)^2}{40}} = \sqrt{\frac{0.4225}{35} + \frac{0.2025}{40}} = 0.1308966332 \simeq 0.1309$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1-1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} = \frac{\left(\frac{0.65^2}{35} + \frac{0.45^2}{40}\right)^2}{\frac{\left(\frac{0.65^2}{35}\right)^2}{34} + \frac{\left(\frac{0.45^2}{40}\right)^2}{39}} = 59.3911735 \simeq 59$$

$df = 59$ and 90% confidence level, critical t-value $t^* \simeq 1.671$

$$CI = (6.7 - 6.4) \pm 1.671 \times 0.1309 = 0.3 \pm 0.2187$$

$$CI = (0.0813, 0.5187)$$

Null Hypothesis (H_0): $\mu_1 - \mu_2 = 0$ (No ~~hours~~ difference in the sleep hours between girls and boys)

Alternative Hypothesis (H_a): $\mu_1 - \mu_2 \neq 0$ (There is difference in sleep hours between girls & boys).

$$t = \frac{(\text{Mean}_1 - \text{Mean}_2)}{SE} = \left(\frac{6.7 - 6.4}{0.1309}\right) = \frac{0.3}{0.1309} \simeq 2.291$$

- Since the calculated test statistic $t (= 2.291) > t^* (= 1.671)$, we reject the hypothesis

* Confidence Interval: The 90% confidence interval for the difference in sleep hours is (0.082, 0.518), which does not include 0. This suggests a significant difference between the sleep hours of girls and boys at the 90% confidence level.

* We reject the null Hypothesis because the calculated test statistic $t = 2.291 > t^* = 1.671$.

THE END