Introductory Statistics - Midterm # 2
Answer Key

1. The Lumina Foundation and Gallup would like to investigate how important are the laws in the state where the college is located on access to reproductive health services is in your decision to stay enrolled until you have graduated. Results were based on web survey responses, conducted Oct. 26 - Nov. 17, 2022, from 6,008 U.S. adults currently enrolled in an associate degree, bachelor's degree, certificate or certification program. It is reported that 72% of currently enrolled college students report that the reproductive health laws in the state where their college is located are at least somewhat important to their decision to stay enrolled.
   a. What is the point estimate for the population proportion?

      The point estimate is given as 72%.

   b. Calculate the mean and standard error for the sampling distribution of sample proportions in this scenario.
      Mean = 0.72
      SE = √[0.72 * (1 - 0.72) / 6,008] = √[0.2016 / 6,008] ≈ √0.0000335552 ≈ 0.005792685 ≈ 0.0058

   c. Verify that the conditions for creating a confidence interval for population proportion are satisfied in the context of this problem.
      - Random sampling: The survey was designed to be random or had a sufficiently large and diverse participant pool, this condition is met.
      - Independence: The sample size is less than 10% of the population, therefore, this condition is met.
      - Large sample: $n\hat{p}$ = 6,008 * 0.72 ≈ 4,324 and $n(1 - \hat{p})$ ≈ 6,008 * 0.28 ≈ 1,684. Both of these values are greater than 10, so the condition is met.

   d. Assume the level of confidence is 95%. What is the z critical value, z*, that correspond to the confidence level?
      z = 1.96

   e. Calculate the margin of error for the sample proportion.
      ME = z*SE ≈ 1.96*0.005792685 ≈ 0.01135 ≈ 0.0114

   f. Using the statistical tool, find the 95% confidence interval and write the interpretation of the confidence interval in context of the problem.

95% confidence interval: (0.7086, 0.7314)

Interpretation: We are 95% confident that the true proportion of currently enrolled college students who consider reproductive health laws in their state important to their decision to stay enrolled lies between approximately 70.86% and 73.14%.

g. Supposed that a news article claimed that three-quarter of enrolled college students report that the reproductive health laws in the state where their college is located are at least somewhat important to their decision to stay enrolled. Using the confidence interval, does the data support the news article claim? Explain your answer.

The confidence interval does not include the possibility that the true proportion is 75%, in fact the values are lower than that. Therefore, based on this confidence interval, we cannot conclude that the data strongly supports the news article's claim. It suggests that the true proportion might be somewhat lower than the claim made in the article.

h. What sample sizes would be needed for a 95% confidence level and a margin of error of 1%?

9,604

i. What would happen to the 95% confidence interval if we took another sample of only 3000 U.S. adults? Explain.

If we were to take a new sample of 3000 U.S. adults, we should expect to have a wider 95% confidence interval, because the smaller sample size introduces more variability and uncertainty into the estimate of the population proportion.

2. The Lumina Foundation and Gallup would like to investigate how important are the laws in the state where the college is located on access to reproductive health services is in your decision to stay enrolled until you have graduated. Results were based on web survey responses, conducted Oct. 26 - Nov. 17, 2022, from 6,008 U.S. adults currently enrolled in an associate degree, bachelor's degree, certificate or certification program. It is reported that 72% of currently enrolled college students report that the reproductive health laws in the state where their college is located are at least somewhat important to their decision to stay enrolled.[1]

   At the 5% significance level, does the data indicate that less than three-quarter of enrolled college students report that the reproductive health laws in the state where their college is located are at least somewhat important to their decision to stay enrolled?

   a. Write the null and alternative hypotheses for this scenario.

[1] https://news.gallup.com/poll/474365/reproductive-health-laws-factor-college-decisions.aspx

Null Hypothesis: The proportion of enrolled college students who report that reproductive health laws in their state are at least somewhat important is equal to 0.75 (75%):
$H_0: p = 0.75$

Alternative Hypothesis: The proportion of enrolled college students who report that reproductive health laws in their state are at least somewhat important is less than 0.75 (75%):
$H_A: p < 0.75$

b. Verify that the conditions for the one-sample z-test for proportions have been met.
   - Random sampling: The survey was designed to be random or had a sufficiently large and diverse participant pool, this condition is met.
   - Independence: The sample size is less than 10% of the population, therefore, this condition is met.
   - Large sample: $n\hat{p} = 6{,}008 * 0.75 \approx 4{,}506$ and $n(1 - \hat{p}) \approx 6{,}008 * 0.25 \approx 1{,}502$. Both of these values are greater than 10, so the condition is met.

c. Use the statistical tool and calculate the test statistic and P-value.
   Test statistic = -5.3701
   P-value = 0

d. Will the null hypothesis be rejected? Explain.
   Yes, we reject the null hypothesis because P-value = 0 < 0.05.

e. Write the conclusion of this hypothesis test in context of the problem.
   The data provides strong evidence to suggest that the true proportion of college students who find these laws at least somewhat important is less than 75%.

f. In this study, did you show "statistical significance?" “practical significance?” Explain.
   The study likely demonstrated statistical significance because the reported p-value was 0, indicating that the results are highly unlikely to be due to random chance alone. While statistical significance suggests that there is a difference in responses, the practical significance would require further analysis to determine whether this difference has meaningful implications for policy or decision-making regarding reproductive health laws.

g. In the context of the problem, what might happen if a type I error occurs?
   If a Type I error occurs, the researchers incorrectly reject the null hypothesis and conclude that less than 75% find the laws important when this may not be the case. This could lead to unwarranted policy changes or actions.

h. In the context of the problem, what might happen if a type II error occurs?
If a Type II error occurs, it means that the researchers failed to detect statistical evidence that less than 75% find the laws important when this is the case. This could result in no policy changes or actions when there is a genuine issue related to the importance of these laws.

3. In 2020, researchers worked to develop a safe and effective vaccine against SARS-CoV-2, the coronavirus that causes COVID-19. A clinical trial was conducted with more than 30,000 adult volunteers nationwide for the Moderna COVID-19 vaccine. Participants were 18 years of age or older with no known previous SARS-CoV-2 infection. Volunteers were randomly assigned to receive either two doses of the investigational vaccine (100 micrograms each) or two shots of a saline placebo.

   The investigators recorded 196 cases of symptomatic COVID-19 among participants who received the vaccine at least 14 days after they received their second shot. Only 11 of these cases were in the group that received the vaccine, with none severe. In contrast, 185 of the cases occurred in the placebo group, 30 of which were severe.[2]

   Their results are organized in the table.

	Non-severe COVID	Severe COVID
Received the vaccine	11	0
Received the placebo	155	30

   a. Calculate the associated sample statistic for the difference in proportions of severe cases for this scenario.
   p_vaccine - p_placebo = 0/11 - 30/185 = 0 - 0.16216 = -0.16216

   b. Using the statistical tool, find the 95% confidence interval. Identify the standard error, margin of error, confidence level, and lower and upper bounds of the confidence interval for estimating the true difference in proportion of severe COVID cases between the vaccinated and placebo groups.
   SE = 0.0271
   ME = 0.05311
   Confidence Level = 95%
   95% confidence interval: (-0.2153, -0.109)

[2] https://www.nih.gov/news-events/nih-research-matters/experimental-coronavirus-vaccine-highly-effective

c. Write the interpretation of the confidence interval in context of the problem.
We are 95% confident that the interval (-0.2153,-0.109) captures the true difference in proportions of severe cases between the vaccine and placebo groups. Because the confidence interval contains all negative values, we can conclude that we are 95% confident that the vaccine is likely to be effective in reducing the risk of severe cases of symptomatic COVID-19 when compared to a placebo.

At the 5% significance level, does the data indicate that the vaccine is effective at reducing the likelihood of an individual catching a severe case of COVID?

d. Write the null and alternative hypotheses for this scenario.
Null Hypothesis: The proportion of severe symptomatic COVID-19 cases among those who received the vaccine (p_vaccine) is equal to the proportion of symptomatic cases among those who received the placebo (p_placebo).
H_0: p_vaccine = p_placebo

Alternative Hypothesis: The proportion of severe symptomatic COVID-19 cases among those who received the vaccine (p_vaccine) is less than the proportion of severe symptomatic cases among those who received the placebo (p_placebo).

H_A: p_vaccine < p_placebo or H_A: p_vaccine - p_placebo < 0

e. Verify that the conditions testing a two-sample z-test for difference in proportions are satisfied.
- Random sampling: Participants were recruited from a nationwide pool of adult volunteers, which suggests that random sampling was used to select participants from the population of interest, this condition is met.
- Independence: It's assumed that participants in the trial were independent of each other, this condition is met.
- Large Sample: p_c = (11+30)/(11+185) ≈ 0.2092
  n_1 * p_c = 11*0.2092 ≈ 2.3
  n_2 * p_c = 185*0.2092 ≈ 38.7
  n_1 * (1-p_c) = 11*(1-0.2092) ≈ 8.7
  n_2 * (1-p_c) = 185*(1-0.2092) ≈ 146.3
  The number of COVID cases is not large enough to meet the normal condition.

f. Use the statistical tool and calculate the test statistic and P-value.
Test statistic = -1.45
P-value = 0.0734

g. Write the conclusion of this hypothesis test in context of the problem.

Because P-value = 0.073 > 0.05 = α, we fail to reject the null hypothesis. This indicates that there is not strong enough statistical evidence to suggest that the Moderna COVID-19 vaccine is effective in reducing the likelihood of catching a severe case of COVID-19.

4. The average score on the math section of 1,737,678 test takers of the 2022 SAT is 521 out of 800.[3] For a simple random sample of 100 students, we found the same average with the standard deviation is 120.
   a. Verify the conditions for creating a confidence interval for population mean.
      - Random sampling: it's mentioned that the sample is a simple random sample of 100 students, which satisfies this condition.
      - Independence: This condition is often assumed to be met if the sample is drawn using random sampling methods and the sample size is less than 10% of the population, therefore, this condition is met.
      - Large sample: The sample size is 100, which is larger than 30, so the Central Limit Theorem suggests that the sampling distribution of the sample mean will be approximately normal.

   b. Using the statistical tool, find the 95% confidence interval and write the interpretation of the confidence interval in context of the problem.
      95% confidence interval: (497.2, 544.8)
      Interpretation: Based on the sample, we are 95% confident that the true mean math score of all students who took the SAT in 2022 falls within the interval from 497.2 and 544.8.

   c. Suppose that we want to test the hypothesis with a significance level of 0.05 that the average score on the math section of the SAT scores for a sample of 100 Lumen high school students who took the test is higher than the national average. Suppose that the average score on the math section of the SAT scores of 100 students at Lumen high school is 540 with a standard deviation of 120. What can we conclude?
      H_0: mu = 521
      H_A: mu > 521
      SE = 12
      Test statistic = 1.5833
      P-value = 0.0583
      Because P-value = 0.0583 > 0.05, we fail to reject the null hypothesis.
      This means that we do not have enough evidence that the average score on the math section of the SAT scores for a sample of 100 Lumen high school students who took the test is higher than the national average.

[3] https://reports.collegeboard.org/media/pdf/2022-total-group-sat-suite-of-assessments-annual-report.pdf

5. The average scores on the math section of the 2022 SAT for Black students (452) are considerably lower than those of Asian students (633), Hispanic/Latino students (473), Native Hawaiian/Other Pacific Islander students (464), White students (543), American Indian/Alaska Native students (463), and Two or More Races students (543).[4]

   Let's use a two-sample t confidence interval and two-sample t-test for population means to investigate if there are significant differences on the average scores on the math section of the 2022 SAT for Black students and Hispanic/Latino students.

   Note: The number of test takers who identify as Black students on the 2020 SAT is 201,645. The number of test takers who identify as Hispanic/Latino students on the 2020 SAT is 396,422. We will use the standard deviation of the math sections of all of the 2022 SAT for both races/ethnicity, which is 120.

   a. Verify that the conditions for the population means have been met.
      - Independent: The samples of Black students and Hispanic/Latino students must be independent of each other, this condition is met.
      - Random sapling/assignment: These students were assigned to their respective groups (Black or Hispanic/Latino), this condition is satisfied.
      - Sample size: These are very large sample sizes, this condition is met.

   b. Using the statistical tool, find the 95% confidence interval. Identify the point-estimate, standard error, margin of error, confidence level, and lower and upper bounds of the confidence interval for estimating the differences on the average scores.
      Point estimate: 452-473 = -21
      Standard error: 0.328
      Margin of error: 0.643
      Confidence level: 955
      Confidence interval: (-21.643, -20.357)

   c. Write the interpretation of the confidence interval in context of the problem.
      We are 95% confident that the true difference in mean math scores between Black students and Hispanic/Latino students is between -21.643 and -20.357.

   d. Write the null and alternative hypotheses for this scenario.
      Null Hypothesis: There is no significant difference in the average math scores between Black students and Hispanic/Latino students on the 2022 SAT.

[4] https://reports.collegeboard.org/media/pdf/2022-total-group-sat-suite-of-assessments-annual-report.pdf

Alternative Hypothesis: There is a significant difference in the average math scores between Black students and Hispanic/Latino students on the 2022 SAT.

e. Use the statistical tool and calculate the test statistic and P-value.
Test statistic: t = -63.979
P-value < 0.0001

f. Will the null hypothesis be rejected? Explain.
Because the P-value is less than 0.001, which is less than the significance level of 0.05, we reject the null hypothesis.

g. Write the conclusion of this hypothesis test in context of the problem.
Based on the data set, there is strong evidence indicating that there exist a significant difference in average math scores between Black students and Hispanic/Latino students.

Additionally, the interval does not contain zero, it suggests that there is a statistically significant difference between the average math scores of Black students and Hispanic/Latino students. The direction of the difference (negative) indicates the Black students have higher average math scores than the Hispanic/Latino students.

h. An article has been published with the title “SAT math scores mirror and maintain racial inequity”. Based on the information provided and your analysis between two races/ethnicity, do you agree or disagree with the title of the article? Justify your answer using your statistical analyses results above.
Based on the information provided and the statistical analysis conducted, we have rejected the null hypothesis, which means that there is a significant difference in average math scores between Black students and Hispanic/Latino students on the 2022 SAT.

Given this result, it is reasonable to agree with the title of the article, "SAT math scores mirror and maintain racial inequity." The analysis suggests that there is a disparity in average math scores between these two racial/ethnic groups, which aligns with the concept of racial inequity in standardized testing outcomes.

However, it's essential to note that statistical significance does not imply causation or provide a complete understanding of the underlying factors contributing to these differences. The article may explore further factors and contexts contributing to these score disparities and their implications for educational equity and fairness.

6. Does 10K running time decrease when the runner listens to music?
Nine runners were timed as they ran a 10K with and without listening to music. The running times in minutes are shown below.

With music	53	50	38	53	54	38	63	53	42
Without music	56	54	42	46	54	41	68	58	41

Assume a normal distribution, what can be concluded at 0.01 level of significance?

a. For this study, explain why t-test for the difference between two dependent population means should be used instead of the test for two independent population means.
Because the same group of runners was tested under two different conditions: with and without listening to music. These two sets of running times are paired or dependent because each runner's performance was measured twice, once with music and once without music. Using a paired or dependent samples t-test allows us to account for the individual differences in running abilities among the same group of runners. It assesses whether there is a statistically significant difference in the mean running times when the runners listened to music compared to when they didn't.

b. Write the null and alternative hypotheses for this scenario.
Null Hypothesis: There is no significant difference in running times between the two conditions (with and without music).
H_0: $\mu_d = 0$
Alternative Hypothesis: There is a decrease in running times when runners listen to music.
H_A: $\mu_d < 0$

c. Use the statistical tool and calculate the test statistic and P-value.
Test statistic: $t = -1.368$
P-value = 0.1042

d. Interpret the P-value in context of the study.
The p-value of 0.1042 is considered relatively large in comparison to 0.01 level of significance. It indicates that the data is reasonably consistent with what you expect if the null hypothesis (no significance difference in running times between the two conditions) were true. In another word, based on the P-value, we do not have enough evidence to reject the null hypothesis.

e. Write the conclusion of this hypothesis test in context of the problem.
There is not enough statistical evidence to conclude that listening to music significantly decreases running times for the group of runners.