CSCI 3022 – *Introduction to Data Science with Probability and Statistics* – **FINAL SOLUTIONS**

Final Exam – Summer 2024 (Theoretical Portion – 75 points)

08/08/2024

1. If you're interested in starting your own candy store and have a good credit rating, you could likely secure a bank loan for franchises like Candy Vibe, The Fudge Lodge, Corn Karmel, and Rocky Mountain Chocolate Express. Below are the startup costs (in thousands of dollars) for a selection of candy stores. These costs are assumed to follow a roughly normal distribution.

95, 173, 129, 95, 75, 94, 116, 100, 85

(a) Find a 90% confidence interval for the population average startup costs μ for candy store franchises. (10 points)

Entering the above data into list L_1, the `TInterval : Data` option can be used with `List` $= L_1$, `Freq` $= 1$, `C-Level` $= .90$. The result is $(88.639, 125.14)$.

(b) What does this confidence interval mean in the context of the problem? (5 points)

We are 90% confident that the interval \$88,639 to \$125,140 is one that contains the average startup cost for all candy store franchises.

2. From a random sample of n = 40 current major league baseball players, a 90% confidence interval for the population mean μ of home run percentages for all current major league baseball players was determined to be 1.93 to 2.65.

(a) What does this imply that the sample mean $\bar{x}$ of home run percentages was? What is ***E (standard error)*** in this case? (5 points). (Hint: If you do not know the population standard deviation the t-interval is used not z interval)

The sample mean $\bar{x}$ is the center of the confidence interval. So, that implies $\bar{x} = 2.29$.
The maximal margin of error E is the distance from $\bar{x}$ to either endpoint of the interval, or equivalently, half the length of the interval. So, that implies $E = 0.36$.

(b) Determine a 99% confidence interval for the population mean μ of home run percentages.
(HINT: First, use E to find the value of s, then use n and sample mean $\bar{x}$ along with the s.) (10 points)

We know the value of E is 0.36 from above. We also know by formula that $E = t_{0.90} \cdot \frac{s}{\sqrt{n}}$. Note that $t_{0.90} = \texttt{InvT(0.95, 39)} = 1.684875$. So, $0.36 = 1.684875 \cdot \frac{s}{\sqrt{40}}$ which implies $1.3513405 = s$. Then `TInterval: Stats` with $\bar{x} = 2.29$, $S_x = 1.3513405$, $n = 40$, and `C-Level` $= .99$, yields $(1.7114, 2.8686)$.

The 99% confidence is longer than the 90% confidence, as we would expect. To make a more confident statement with the same data, we need to stretch that interval out to allow for a bit more error in our sample.

3. A *CNN* poll asked the question, "What is earth's biggest current issue?" Nineteen percent of people asked replied, "burglaries and crime." Assuming that the sampling error had a margin of plus or minus 3 percentage points. Following the convention that the margin of error is based on a 95% confidence interval, find a 95% confidence interval for the percentage of the entire population that would respond "Crime and violence" to the question asked by the pollsters. (5 points)

Since we are following the convention that the margin of error creates a 95% confidence interval, that interval would simply be $19 - 3 = 16$ to $19 + 3 = 22$. So, we are 95% confident that the interval 16% to 22% is one that contains the true population percentage.

4. If the *P*-value in a statistical test is less that or equal to the level of significance for the test, do we reject or fail to reject H_0? Does this imply that there IS or IS NOT enough evidence in the data (and the test being used) to justify the rejection of H_0? (5 points)

When the P-value is less than or equal to the level of significance, we REJECT the null hypothesis H_0. This means that there IS enough evidence in the data to justify the rejection of H_0 and choose the alternate hypothesis H_1...although it is NOT proof that H_1 is true beyond all doubt.

5.

Weatherwise magazine is published in association with the American Meteorological Society. Volume 46, Number 6 has a rating system to classify Nor'easter storms that frequently hit New England states and can cause much damage near the ocean coast. A *severe* storm has an average peak wave height of 16.4 feet for waves hitting the shore. Suppose that a Nor'easter is in progress at the severe storm class rating.

(a) Let us say that we want to set up a statistical test to see if the wave action (i.e., height) is dying down or getting worse. What would be the null hypothesis regarding average wave height?

$H_0 : \mu = 16.4$ feet.

(b) If you wanted to test the hypothesis that the storm is getting worse, what would you use for the alternate hypothesis?

$H_1 : \mu > 16.4$ feet.

(c) If you wanted to test the hypothesis that the waves are dying down, what would you use for the alternate hypothesis?

$H_1 : \mu < 16.4$ feet.

(d) Suppose you do not know whether the storm is getting worse or dying out. You just want to test the hypothesis that the average wave height is *different* (either higher or lower) from the severe storm class rating. What would you use for the alternate hypothesis?

$H_1 : \mu \neq 16.4$ feet.

(e) For each of the tests in parts (b), (c), and (d), would the area corresponding to the P-value be on the left, on the right, or on both sides of the mean? Explain your answer in each case.

(b) Right; (c) Left; (d) Both Sides. That is, for (b), we use a right-tailed test; for (c), we use a left-tailed test; and for (d), we use a two-tailed test.

6. Gentle Ben is a Morgan horse at a Colorado dude ranch. The mean glucose level for horses should be μ = 85 mg/100 ml (Reference: *Merck Veterinary Manual*). Over the past 8 weeks, a veterinarian took weekly glucose readings from this horse (in mg/100 ml) and found the sample mean $\bar{x}$ = 93.8 mg/100 ml. Do the data indicate that Gentle Ben has an overall average glucose level higher than 85 mg/100 ml?

(a) State the appropriate null and alternate hypothesis for this test. Is this a left-tailed, right-tailed, or two-tailed test? (5 points)

$H_0 : \mu = 85\text{mg}/100\text{ml}$ $H_1 : \mu > 85\text{mg}/100/\text{ml}$ This is a right-tailed test.

(b) If we assume that x has a normal distribution and that we know from past experience that σ = 12.5, then the corresponding P-value is about 0.0232. Verify this is correct using the appropriate statistical test. At the α = 0.05 level, do these data indicate that Gentle Ben has an overall average glucose level higher than 85 mg/100 ml? Explain. (5 points)

Using `Z-Test`, one verifies that the P-value is about 0.0232. Since $P \leq \alpha$, at the 0.05 significance level we choose to REJECT the null hypothesis. In other words, the data suggests (but do NOT prove) that Gentle Ben has an overall average glucose level higher than 85mg/100ml.

7. You draw a random sample of size n = 64 from a population with mean μ = 50 and standard deviation σ = 16. From this, you compute the sample mean, $\bar{x}$. **Hint: Use the Central Limit Theorem**

(a) What are the expectation and standard deviation of $\bar{x}$? (5 points)

$$\mathrm{E}[\bar{X}] = \mu = 50,$$
$$\mathrm{sd}[\bar{X}] = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{64}} = 2.$$

(b) Approximately what is the probability that the sample mean is above 54? (5 points)

The sample mean has expectation 50 and standard deviation 2. By the central limit theorem, the sample mean is approximately normally distributed. Thus, by the empirical rule, there is roughly a 2.5% chance of being above 54 (2 standard deviations above the mean).

b

(c) Do you need any additional assumptions for this question to be true? Hint look at sample size (5 points)

No. Since the sample size is large ($n \geq 30$), the central limit theorem applies.