Name:_______________________________________ **Date:**____________________

PSTAT 5A: Homework 04

Summer Session A 2023, with *Ethan P. Marzban*

1. In a particular iteration of PSTAT 5A, scores on the final exam had an average of 89 and a standard deviation of 40 The exact distribution of scores is, however, unknown. Suppose a representative sample of 100 students is taken, and the average final exam score of these 100 students is recorded.

   (a) Identify the population.

   Solution: The population is the set of all students in the aforementioned iteration of PSTAT 5A.

   Final Answer(s): No numerical answer.

   (b) Identify the sample.

   Solution: The sample is the 100 students that were selected.

   Final Answer(s): No numerical answer.

   (c) Define the parameter of interest. Use the notation discussed in Lecture 12.

   Solution: We use μ to denote population means; as such, let μ denote the true average final exam score of PSTAT 5A students.

   Final Answer(s): No numerical answer.

   (d) Define the random variable of interest. Use the notation discussed in Lecture 12.

   Solution: We use $\overline{X}$ to denote sample means; as such, let $\overline{X}$ denote the average final exam score of 100 randomly-selected students from PSTAT 5A.

   Final Answer(s): No numerical answer.

   (e) What is the sampling distribution of the random variable you defined in part (d) above? Be sure to check any conditions that might need to be checked!

Name:____________________________________ **Date:**__________________

Solution: The first question we ask ourselves is: is the population normally distributed? The answer is no. As such, we then ask ourselves: is the sample size greater than 30? The answer is yes. As such, we finally ask ourselves: is the population standard deviation known? The answer is yes. Hence, $\overline{X}$ will be normally distributed; specifically,

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \sim \mathcal{N}\left(89, \frac{40}{\sqrt{100}}\right) \sim \mathcal{N}(89, 4)$$

Final Answer(s): $\mathcal{N}(89,\ 4)$

(f) What is the approximate probability that the average score of these 100 students lies within 5 points of the true average score of 89?

Solution: We seek $\mathbb{P}(84 \leq \overline{X} \leq 94)$. As such, we compute

$$\begin{aligned}\mathbb{P}(84 \leq \overline{X} \leq 94) &= \mathbb{P}(\overline{X} \leq 94) - \mathbb{P}(\overline{X} \leq 84) \\ &= \mathbb{P}\left(\frac{\overline{X} - 89}{4} \leq \frac{84 - 89}{4}\right) - \mathbb{P}\left(\frac{\overline{X} - 89}{4} \leq \frac{94 - 89}{4}\right) \\ &= \mathbb{P}(Z \leq 1.25) - \mathbb{P}(X \leq -1.25) = 0.8944 - 0.1056 = 78.88\%\end{aligned}$$

Final Answer(s): 78.88%

2. Quinn is interested in performing inference on the average weight of Granny Smith apples in the Santa Barbara location of *Bristol Farms*. To that end, he takes a representative sample of 52 apples; the mean weight of his sample was 83g and the standard deviation of weights in his sample was 17g.

(a) Identify the population.

Solution: The population is the set of all apples at the Santa Barbara location of *Bristol Farms*.

Final Answer(s): No numerical answer.

(b) Identify the sample.

Solution: The sample is the set of 52 apples Quinn selected.

Final Answer(s): No numerical answer.

(c) Define the parameter of interest. Use the notation discussed in Lecture 12.

Solution: We let μ denote the true average weight of Granny Smith apples at the Santa Barbara location of *Bristol Farms*.

Final Answer(s): No numerical answer.

(d) Define the random variable of interest. Use the notation discussed in Lecture 12.

Solution: We let $\overline{X}$ denote the average weight of a sample of 52 Granny Smith apples, taken from the Santa Barbara location of *Bristol Farms*.

Final Answer(s): No numerical answer.

(e) What distribution do we use to construct confidence intervals for the true average weight of a Granny Smith apple at the Santa Barbara location of *Bristol Farms*?

Solution:

- Is the population normally distrbuted? No.
- Is the sample size greater than 30? Yes.
- Is the population standard deviation known? No, only the sample standard deviation.

As such, we use the t distribution with $n - 1 = 52 - 1 = 51$ degrees of freedom; i.e. the t_{51} distribution.

Final Answer(s): Use the t_{51} distribution.

(f) Construct a 95% confidence interval for the true average weight of a Granny Smith apple at the Santa Barbara location of *Bristol Farms*.

Solution: Our confidence interval will be of the form

$$\overline{x} \pm t_{51,\ \alpha} \cdot \frac{s}{\sqrt{51}}$$

Here, $\overline{x} = 83$ and $s = 17$. Now, the $t-$table does not actually have a row for 51 degrees of freedom; as such, we can either obtain an approximate value by simply using 50 degrees of freedom (which gives us a value of $t_{51,\alpha} \approx 2.01$), or we can simply use Python (which also gives us a value of around 2.01). As such, our confidence interval becomes

$$83 \pm (2.01) \cdot \frac{17}{\sqrt{52}} = \boxed{[78.26147\ ,\ 87.73853]}$$

Final Answer(s): [78.26147 , 87.73853]

3. Meta recently launched the social media app *Threads*. As the new resident Data Scientist for Meta's Santa Barbara division (congratulations!), you would like to determine the true proportion of Santa Barbara residents that have made a *Threads* account. Your supervisor believes that 47% of all Santa Barbara residents have made a *Threads* account; in a representative sample of 120 residents, however, you observe that only 48 of these sampled individuals have made a *Threads* account. You would like to use your data to test your supervisor's claims against a two-sided alternative, at a 5% level of significance.

   (a) Define the parameter of interest.

   Solution: Let p denote the true proportion of Santa Barbara residents that have made a *Threads* account.

   Final Answer(s): No numerical answers.

   (b) Define the random variable of interest.

   Solution:
   Let $\widehat{P}$ denote the proportion of Santa Barbara residents in a sample of 120 that have made a *Threads* account.

   Final Answer(s): No numerical answers.

   (c) State the null and alternative hypotheses in terms of the parameter of interest.

Solution: Our null hypothesis is that $p = 0.47$; we are told to adopt a two-sided alternative, meaning our hypotheses take the form

$$\left[\begin{array}{rl} H_0: & p = 0.47 \\ H_A: & p \neq 0.47 \end{array} \right.$$

Final Answer(s): No numerical answers.

(d) What is the observed value of the test statistic?

Solution: First note that $\widehat{p} = (48/120) \cdot 100 = 0.4$. Thus,

$$\mathrm{ts} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.4 - 0.47}{\sqrt{\frac{0.47 \cdot (1-0.47)}{120}}} \approx -1.54$$

Final Answer(s): -1.54

(e) What distribution does the test statistic follow, assuming the null is correct?

Solution: We know that TS will be normally distributed under the null, provided that:

1) $np_0 = (120)(0.47) = 56.4 \geq 10$ ✓

2) $n(1 - p_0) = (120) \cdot (1 - 0.47) = 63.6 \geq 10$ ✓

Since both conditions hold, we can conclude that

$$\mathrm{TS} \overset{H_0}{\sim} \mathcal{N}(0,\ 1)$$

Final Answer(s): $\mathcal{N}(0,\ 1)$

(f) What is the critical value of the test?

Solution: Recall that for a general α level of significance, the critical value is found to be :

- -1 times the $(\alpha/2) \times 100^{\text{th}}$ percentile of the standard normal distribution, which is equivalent to

- the $[1 - (\alpha/2)] \times 100^{\text{th}}$ percentile of the standard normal distribution

Since $\alpha = 0.05$, this leads us to a critical value of 1.96.

Final Answer(s): 1.96

(g) Conduct the test, and phrase your conclusions in the context of the problem.

Solution: We reject the null only when the absolute value of the observed value of the test statistic exceeds the critical value. Here, $|\text{ts}| = |-1.54| = 1.54 < 1.96$ meaning we fail to reject the null:

> At an $\alpha = 0.05$ level of significance, there was insufficient evidence to reject the null that 47% of Santa Barbara residents have a *Threads* account, in favor of the alternative that the true proportion is *not* 47%.

Final Answer(s): Fail to reject the null.

4. (**Deriving the Lower-Tailed Test of Proportions**). Consider testing the set of hypothesis

$$\left[\begin{array}{rl} H_0: & p = p_0 \\ H_A: & p < p_0 \end{array}\right.$$

at an arbitrary α level of significance. Define the test statistic TS to be

$$\text{TS} = \frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

(a) Show that $\text{TS} \stackrel{H_0}{\sim} \mathcal{N}(0,\ 1)$. If your answer depends on a set of conditions to be true, explicitly state those conditions.

Solution: So long as we are able to invoke the CLT for Proportions, we will be good. Hence, we need to first assure that both:

1) $np_0 \geq 10$
2) $n(1 - p_0) \geq 10$

Assume the above conditions are true. Then, under the null (i.e. assuming the true value of p is actually p_0), the CLT for proportions tells us

$$\widehat{P} \sim \mathcal{N}\left(p_0, \ \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

which means (by our familiar Standardization result)

$$\frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\sim} \mathcal{N}(0, \ 1)$$

and we are done.

Final Answer(s): No numerical answer.

(b) Argue, in words, that the test should be of the form

$$\texttt{decision}(\mathrm{TS}) = \begin{cases} \texttt{reject } H_0 & \text{if TS} < c \\ \texttt{fail to reject } H_0 & \text{otherwise} \end{cases}$$

for some constant c. As a hint, look up the logic we used in Lecture 13 to derive the two-tailed test, and think in terms of statements like "$\widehat{p}$ is *far away* from p_0". **You do not have to find the value of c in this part**.

Solution: If the null hypothesis states that the true value of p is p_0, and if we observe an instance of $\widehat{p}$ that is much less than p_0, we are more inclined to believe the alternative (i.e. that $p < p_0$) is true. In other words, we would reject the null for *small* values of TS; namely, our rejection region takes the form $(-\infty, c)$.

The key assertion, however, is that we would only really reject the null in favor of the alternative that $p < p_0$ if TS were small in *raw value*, **NOT** in absolute value. Said differently, observing a very large value of TS would **NOT** necessarily lead credence to the claim that $p < p_0$, and hence we would **NOT** reject the null in favor for the alternative if TS were large in the positive direction.

Final Answer(s): No numerical answer.

(c) Now, argue that c must be the $\alpha \times 100^{\text{th}}$ percentile of the standard normal distribution (**NOT** scaled by negative 1), thereby showing that the full test takes the form

$$\texttt{decision}(\mathrm{TS}) = \begin{cases} \texttt{reject } H_0 & \text{if TS} < z_\alpha \\ \texttt{fail to reject } H_0 & \text{otherwise} \end{cases}$$

©Ethan P. Marzban

where z_α denotes the $(\alpha) \times 100^{\text{th}}$ percentile of the standard normal distribution.

Solution: Recall that the level of significance α is precisely the probability of committing a Type I error; i.e. the probability of rejecting the null when the null were true:

$$\alpha = \mathbb{P}_{H_0}(\text{TS} < c)$$

Since, under the null, $\text{TS} \sim \mathcal{N}(0,\ 1)$ (as was shown in part (a) above), this means that c must satisfy

$$\mathbb{P}(Z < c) = \alpha$$

where $Z \sim \mathcal{N}(0,\ 1)$; i.e. c is the α^{th} percentile of the standard normal distribution.

Final Answer(s): No numerical answer.

PLEASE NOTE: You may be expected to use this result on future homework/quizzes/exams.

©Ethan P. Marzban