# PSTAT 5A: Homework 4

Please note, you are responsible for completing your homework and submitting it to Gradescope. Though the Syllabus allows for the option to not submit homework, it is highly recommended.

1. The U.S. Department of Housing and Urban Development defines a person or household to be "rentburdened" if 30% or more of the individual/household's income is spent on housing. A recent survey revealed that 42% of households in a representative sample of 150 households were rent-burdened.
   a. Define the parameter of interest.

**Solution:** Let $p$ denote the proportion of households that are rent-burdened.

   b. Define the random variable of interest.

**Solution:** Let $\widehat{P}$ denote the proportion of housholds in a representative sample of 150 that are rent-burdened.

c. Construct a 95% confidence interval for the true proportion of rent-burdened households, and interpret your interval in the context of the problem.

**Solution:** Our first task is to identify the sampling distribution of $\widehat{P}$, which entails checking the success-failure conditions. Since we don't know the value of $p$, we use the substitution approximation:

1) $n\widehat{p} = (150) \cdot (0.42) = 63 \geq 10$ ✓

2) $n(1-\widehat{p}) = (150) \cdot (1-0.42) = 87 \geq 10$ ✓

Since both conditions are met, we can invoke the Central Limit Theorem for Proportions to conclude that $\widehat{P}$ will be approximately normally distributed. Hence, our Confidence Interval will take the form

$$\widehat{p} \pm z \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where $z$ is the appropriately-selected quantile of the normal distribution. Since we want a 95% confidence interval, we take $z$ to be negative one times the $(1-0.95)/2 \times 100\% = 2.5^{\text{th}}$ percentile of the standard normal distribution, which we know is around 1.96. Hence, our confidence interval is

$$(0.42) \pm (1.96) \cdot \sqrt{\frac{(0.42) \cdot (1-0.42)}{150}} = (0.42) \pm (1.96) \cdot (0.0402)$$
$$\approx [0.341208\ ,\ 0.498792]$$

The interpretation of this interval is:

> We are 95% certain that the true proportion of rent-burdened households is between 34.12% and 49.88%.

d. Would you expect an 80% confidence interval for the true proportion of rent-burdened households to be wider or narrower than the 95% confidence interval you constructed in part (c)? Explain briefly.

**Solution:** We know that higher confidence levels lead to wider confidence intervals, meaning an 80% CI should be narrower than a 95% one.

e. Construct an 80% confidence interval for the true proportion of rent-burdened households, and interpret your interval in the context of the problem.

**Solution:** We can re-use a lot of our work from part (c) above: all we really need to change is our value of $z$. Now, we use negative one times the the $(1 - 0.8)/2 \times 100\% = 10^{\text{th}}$ percentile of the standard normal distribution, which from our normal table gives us a value of around 1.28. (Recall that this is equivalent to finding the $90^{\text{th}}$ percentile, due to the symmetry of the normal distribution.) Hence, our confidence interval becomes

$$(0.42) \pm (1.28) \cdot (0.0402) \approx [0.368544\,,\ 0.471456]$$

and the interpretation of this interval is:

> We are 80% certain that the true proportion of rent-burdened households is between 36.9% and 47.1%.

2. 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.

b. Identify the sample.

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

c. Define the parameter of interest. Use the correct notations for the distribution.

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

d. Define the random variable of interest. Use the correct notations for the distribution.

**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.

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!

**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,

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

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

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

3. 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*.

b. Identify the sample

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

c. Define the parameter of interest. Use the correct notations for the distribution.

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

d. Define the random variable of interest. Use the correct notations for the distribution.

**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*.

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.

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

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

4. 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.

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.

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

$$\begin{cases} H_0: & p = 0.47 \\ H_A: & p \neq 0.47 \end{cases}$$

d. What is the observed value of the test statistic?

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

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

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

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

f. What is the critical value of the test?

**Solution:** Recall that for a general $\alpha$ 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 $[1-(\alpha/2)] \times 100^{th}$ percentile of the standard normal distribution

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

g. Conduct the test, and phrase your conclusion 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, $|ts| = |-1.54| = 1.54 < 1.96$ 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%.

5. **[EXTRA CREDIT 6 points] (Deriving the Lower-Tailed Test of Proportions)**. Consider testing the set of hypothesis

$H_0: p = p_0$

$H_A: p < p_0$

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

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

a. Show that TS ~ N(0, 1) under the null hypothesis. 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

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

which means (by our familiar Standardization result)

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

and we are done.

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

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

for some constant c. As a hint, look up the logic we used in lecture to derive the two-tailed test, and think in terms of statements like "$\hat{p}$ 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.

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

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

where $z_\alpha$ denotes the $(\alpha)\times 100$th percentile of the standard normal distribution.

**Solution:** Recall that the level of significance $\alpha$ 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 $\alpha^{\text{th}}$ percentile of the standard normal distribution.