MATH 254: Introduction to Statistics **Mintaek Lee**

Lesson 3: Large Sample Inference for a Population Proportion

Corresponding Workbook Module: 4

Example 3.1 *Mintaek spent his entire paycheck to buy 40 scratch lottery tickets that claim 35% of their tickets are winning tickets. After scratching all of them, he found 11 of them were winning tickets, each with $1 prize. Devastated and desperate, Mintaek wants to file a lawsuit against the state lottery commission that they are falsely advertising their winning odds. Use the appropriate statistical methods and set up the hypotheses that best matches Mintaek's interests.*

Step 1: Note that the state lottery commission's claim is H_0 since Mintaek has not found any evidence against them at this point. The alternative hypothesis is Mintaek's claim (what he wants to find out) that the odds of winning is less than 35%. He is not interested in checking whether the odds of winning is higher than 35%. Hence, the H_A is $p < 0.35$.

- H_0: $p = 0.35$, the population proportion of winning lottery tickets is equal to 35%.
- H_A: $p < 0.35$, the population proportion of winning lottery tickets is **less than** 35%.

Step 2: We first need to check if the problem is indeed a binomial setting, using 'BINS' (See workbook page 124 for the description of BINS).

- **B**inary: there are only two possible outcomes: winning ticket or not a winning ticket.
- **I**ndependence: presumably, each lottery ticket is issued randomly.
- **N**umber: number of tickets Mintaek purchased was set as 40 in advance.
- **S**uccess: since there would certainly be more than 400 (which is 10 times the sample size, $n = 40$) lottery tickets issued statewide, population size is at least 10 times the sample size. Thus it is reasonable to assume that the probability of winning (success probability) is stable (i.e. same for every ticket scratched).

Now we need to check if we can apply the one-sample z-test for proportions here. The only requirement is that we have a binomial setting (which we checked above) and have a large enough sample (See workbook page 124 for description of the conditions).

We check that the number of success (number of winning tickets) $= 11 > 10$ and the number of failure (number of non-winning tickets) $40 - 11 = 29 > 10$.

Since all conditions are met, we can use the one-sample z-test for proportions to make an inference on p, the population proportion (proportion of ALL winning tickets statewide). For the significance level, you may use $\alpha = 0.05$ which is an industry standard, unless otherwise specified.

Step 3: We can now find the z-test statistic value and corresponding P-values.

We first find the following information from the question: (1) the sample proportion of winning odds, $\hat{p} = \dfrac{11}{40} = 0.275$, (2) the presumed population proportion from the null hypothesis (or, just the null hypothesis value), $p_0 = 0.35$, and (3) the sample size, $n = 40$. Then we have

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{0.275 - 0.35}{\sqrt{\frac{0.35(1-0.35)}{40}}} \\ &= -0.99 \end{aligned}$$

where p_0 is the hypothesized population proportion from the hypotheses.

From the standard normal distribution (z) table, we can see that our z-test statistic $z = -0.99$ corresponds to 0.1611 as the **lower** tail probability (area to the left from $z = -0.99$ is 0.1611 OR $P(Z \leq -0.99) = 0.1611$). This lower tail area calculated from $z = -0.99$ can be interpreted as the probability of having 27.5% (because $\hat{p} = 0.275$) or LESS of tickets purchased as winning tickets is 0.1611 if H_0 was true. Since our H_A is $p < 0.35$ (one-sided test), we are only interested in the lower tail probability, so our P-value would be 0.1611.

Step 4: In this case, we would fail to reject the null hypothesis that the population proportion of winning lottery tickets is 35% under the significance level of $\alpha = 0.05$. It is because our P-value was larger than α. **In other words, the probability of obtaining a sample with 27.5% or less of tickets being winning tickets is about 16.11% if the true population proportion of winning lottery tickets were indeed 35%.** It is not a particularly surprising result if the actual odds of winning was indeed 35%. Therefore, we do not have sufficient evidence to suggest that population proportion of winning lottery tickets is **less than** 35%

IF THIS WAS A TWO-SIDED TEST WHERE H_A***:*** $p \neq 0.35$

Step 1: H_0 is the same, just H_A needs to be changed.

- H_0: $p = 0.35$, the population proportion of winning lottery tickets is equal to 35%.
- H_A: $p \neq 0.35$, the population proportion of winning lottery tickets is **different from** 35%.

Step 2: Same.

Step 3: For a two-sided test, we are interested in both lower and upper tails. It is because we are interested in finding out p (the population proportion of winning lottery tickets) is significantly less than 0.35 OR higher than 0.35.

The upper tail area calculated from $z = 0.99$ (Notice it is a positive number. Why?) can be interpreted as the probability of obtaining a sample with $z = 0.99$ or MORE is 0.1611 if H_0 were true. Notice it is very similar to the one for the lower tail, just the opposite interpretation. Since the standard normal distribution is symmetric, this upper tail area would be the same as the lower tail area. Therefore, we can obtain area of both tails by doubling one tail area. Since the lower tail area was 0.1611, our P-value would just be $2 \times 0.1611 = 0.3222$.

Step 4: We would still fail to reject the null hypothesis that the population proportion of winning lottery tickets is 35% under the significance level of $\alpha = 0.05$ since our P-value was larger than α. Therefore, we do not have sufficient evidence to suggest that population proportion of winning lottery tickets is **not equal to** 35%

Example 3.2 *Since Mintaek realized that he can't sue the state lottery commission with the information he has, he now wants to estimate the population proportion of winning lottery tickets. So, he decides to construct a 96% confidence interval. It will provide him with the estimation of the population proportion of winning lottery tickets with 96% confidence/certainty.*

Step 1: Technically, you still would need to check all the conditions for the one-sample z-confidence interval for proportions. However, you may notice that they are identical to ones for the one-sample z-test for proportions. So, you could omit this section if you already answered Question 1.

Same as **Step 1** from **Question 1**.

Step 2: We first need to calculate the margin of error in order to construct the appropriate z-confidence interval. We find from the t-distribution table, the z-critical value (z^*) corresponding to 96% confidence level is 2.054. We can then find the margin of error as

$$\begin{aligned} m &= z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= 2.054 \sqrt{\frac{0.275(1-0.275)}{40}} \\ &= 2.054 \times 0.0706 \\ &= 0.145 \end{aligned}$$

Notice that we use the **standard error** of sample proportion, $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, instead of the **standard deviation** of sample proportion, $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$ as we did in Question 1. This is because in Question 1, we are calculating the likelihood of our sample assuming that H_0 is true, so we use $\sigma_{\hat{p}}$. However, we do not have such assumptions here. We are making our best estimation of p using the sample we have. Therefore, we use the best estimate we have for $\sigma_{\hat{p}}$, which is standard error, $SE_{\hat{p}}$.

Step 3:

We can now construct the 96% confidence interval for p, population proportion of winning lottery tickets.

$$\hat{p} \pm m = 0.275 \pm 0.145 = (0.13, 0.42)$$

This means that we are 96% certain/sure/confident that the interval (0.13, 0.42) contains p, population proportion of winning lottery tickets.

Example 3.3 *Mintaek was not satisfied with how wide his confidence interval was. He wants to figure out how many lottery tickets he would need to purchase if he wants the margin of error to be 0.01 for the 96% confidence interval.*

We can find the sample size for a given level of confidence and margin of error using

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1-p^*)$$

where (1) $m = 0.01$ is the desired margin of error, (2) $z^* = 2.054$ is the critical value for the desired level of confidence, and (3) p^* is the guessed population proportion.

Notice that we do not use $\hat{p}$ as we did in Question 2 here. Instead, we use p^*. It is because we won't know $\hat{p}$ until we collect all samples, and we can't collect samples until we determine the sample size, n. For this reason, we usually use the given population proportion, p_0, as our p^*. If it is not given, it is conventional to just use $p^* = 0.5$. For more information on this, see the additional comments on the next page.

However, remember that we are given that the population proportion is $p_0 = 0.35$ from the state lottery commission's claim. So, we can just use $p^* = 0.35$ here.

Then we would get

$$\begin{aligned} n &= \left(\frac{z^*}{m}\right)^2 p^*(1-p^*) \\ &= \left(\frac{2.054}{0.01}\right)^2 0.35\ (1-0.35) \\ &= 9598.0339 \end{aligned}$$

Remember that 9598.0339 is the **minimum** number of lottery tickets Mintaek has to buy to get the desired margin of error. If you round it to 9598, you won't necessarily be able to get the desired margin of error. That is why you always **round up** to get your minimum required sample size.

Mintaek would need to buy at least 9599 lottery tickets to get the desired margin of error. (To buy that much, he would need to collect almost 20 years worth of paychecks!)

Additional Comments: Notice that there are four different terms for proportions: p, $\hat{p}$, p_0, and p^*. Make sure that you can clearly distinguish between them.

- p is the 'true' population value which is rarely known. This is the value that we try to estimate by constructing the confidence interval.
- $\hat{p}$ is an estimation of p from your sample. You divide the number of success (x) by the total number of trials or sample size (n) to obtain $\hat{p} = \dfrac{x}{n}$.
- p_0 is the hypothesized population proportion, or just the null hypothesis value. This is the value that you make judgments with in hypothesis testing. For example, suppose that you want to test if your sample supports the claim that population proportion is larger than 70%. Then, your null and alternative hypotheses would be: $H_0 : p = 0.7$ vs. $H_A : p > 0.7$. Note that 0.7 is actually p_0 values, the hypothesized population proportion. You will conduct appropriate hypothesis testing to see if your evidence supports H_A or not (or H_0 can be rejected or not).
  Lastly, p_0 can come from the following sources:
  1. Values given from the context: For example, in Problem 1 from workbook page 121, the researcher states that they want to test if population proportion is larger than 70% or not. So you use $p_0 = 0.7$ to make inferences.
  2. Equal chance between two options, i.e. $p_0 = 0.5$. For example, in Problem 2 from workbook page 122, they want to see if majority of people get better results from their product. Since you have the majority of people if you have larger than 50%, you would set $p_0 = 0.5$ and make inferences.
  3. Population value: For example, in Problem 3 from workbook page 123, you are given the population proportions, and you make inferences from it.
- p^* is the estimated value of the proportion to use when you are calculating the desired sample size for a certain margin of error. See the bottom of workbook page 132 for more information. We can't use $\hat{p}$ because you need to determine n in advance to collect the sample (we don't know n because that's what we are trying to find here). Also, you can't use p_0 because you are not doing a hypothesis testing here. That's why when you are calculating the desired sample size for a certain margin of error, you need to find p^*. We usually use the given population value from the context for it. If none is given, just use $p^* = 0.5$.