# HSCI 50 LAB 8: Sample Size Estimation ANSWER KEY

**INSTRUCTIONS**

The data set that we will be using for the entire course is resampled data from the MIT COVID-19 Beliefs, Behaviors & Norms Survey (https://covidsurvey.mit.edu/api.html). This is a multi-country, online survey that examined different COVID-19 perceptions across time, from July 6, 2020 to March 28, 2021. We will be using data from the Philippines aged 20 to 60. **Assume the data are randomly sampled.**

In this lab, you will learn how to calculate sample sizes.

You have many options to submit this worksheet. Either you work on this by hand and scan/take a clear photo of your submission and save as PDF, or type your responses in a Word processor or PowerPoint presentation. You do not have to copy the questions again, but please number them accordingly.

We will now assume that you plan to conduct your thesis on COVID-19 vaccine hesitancy among students and employees of Ateneo de Manila University. You plan to estimate sample size based on the results you have gotten from the past 7 labs in this course.

**PART A. One-sample proportion - precision**

You aim to select a simple random sample of currently enrolled Ateneo de Manila University **students**. Upon consultation with the relevant university administration authorities, they request that you give them as precise of an estimate as possible of the level of vaccine hesitancy in the university, with the margin of error set at $\pm 2\%$. For the expected proportion, you take the conditional probability of vaccine hesitancy given 20-30 years old that we calculated in Lab 5, which is 55%. We set $\alpha$ to the typical value of 0.05. There is an estimated total of 10,000 currently enrolled students in the university.

**1. Calculate by hand the sample size required given the estimates above, assuming an infinitely large population, i.e. without applying finite population correction.**

The given values are as follows:

- $z_{\alpha/2} = 1.96$ given $\alpha = 0.05$
- $p = 0.55, 1 - p = 0.45$
- $d = 0.02$

The formula for a sample size of one-sample proportion - precision is:

$$n = \frac{(z_{\alpha/2})^2 (p)(1-p)}{d^2}$$

Plugging in the values we get:

$$n = \frac{(1.96)^2 (0.55)(1-0.55)}{0.02^2} = 2376.99 = 2377$$

Therefore, we would need to recruit 2,377 participants.

**2. Given we have a finite population of students in the university, calculate by hand the sample size with finite population correction.**

**The sample size formula that accounts for finite population correction is:**

$$\text{FPC } n' = \frac{n}{1 + n/N}$$

**NOTE: The formula $FPC = 1 - n/N$ discussed during lecture is applied to the variance. The formula above is the FPC-adjusted sample size that accounts for FPC on the variance**

The given values are as follows:

- *n* = 2377 (sample size calculated from #1)
- *N* = 10000 (total population of Ateneo de Manila University students)

Plugging in the values we get:

$$\text{FPC } n' = \frac{2377}{1 + 2377/10000} = 1920.5 = 1921$$

Therefore the sample size decreases to 1921.

**3. Say you expect that 20% of the population are not going to participate in the survey when reached out, so we have to inflate the sample size by 20%. That is the same as dividing the sample size by 80%. Calculate the final sample size with the finite population correction and the inflation to account for non-response.**

Dividing the sample size by 80%, we get a final sample size of 1921/0.80 = 2401.25 = 2402.

**4. The sample size you have calculated in #3 is not very feasible at all. You calculate a range of possible values instead based on a range of possible values for the margin of error to negotiate with the university administration on a much lower sample size**

| Margin of error | FPC-adjusted n | FPC-adjusted n with 20% non-response |
|---|---|---|
| 0.01 | 4874 | 6093 |
| 0.02 | 1921 | 2402 |
| 0.03 | 956 | 1195 |
| 0.04 | 561 | 702 |
| 0.05 | 367 | 459 |
| 0.06 | 258 | 323 |
| 0.07 | 191 | 239 |
| 0.08 | 147 | 184 |
| 0.09 | 117 | 147 |
| 0.1 | 95 | 119 |

**Based on the table of values above, what do you think is a reasonable compromise to offer the university administration for a feasible sample size while maintaining as high of a precision as possible?**

A margin of error of 0.05 seems feasible while still fairly precise, with a sample size of 459.

**Part B. Two-sample proportions - hypothesis test**

Another objective of your thesis is that you want to determine whether vaccine hesitancy differs between students and university employees.

**1. We take the data from Lab 6 on the vaccine hesitancy proportions between younger and older adults, and assume that the proportion for students apply to younger adults, and the proportion for employees apply to older adults. Recall that the percentages are 0.52 and 0.37 respectively. Assuming the typical $\alpha = 0.05$ and 80% power, equal samples to be recruited in each group, and accounting for 20% non-response, calculate by hand the sample size required.**

The given values are as follows:

- $z_{\alpha/2} = 1.96$ given $\alpha = 0.05$
- $z_{\beta} = 0.84$ given 80$ power (or $\beta = 0.2$)
- $p_1 = 0.52, 1 - p_1 = 1 - 0.52 = 0.48$
- $p_2 = 0.37.1 - p_2 = 1 = 0.37 = 0.63$
- $\Delta = p_1 - p_2 = 0.15$
- $\bar{p} = \frac{p_1+p_2}{2} = \frac{0.52+0.37}{2} = 0.445; 1 - \bar{p} = 0.555$

The formula for a sample size of two-sample proportions - hypothesis test is:

$$n = \left[ \frac{(z_{\alpha/2})\sqrt{2\bar{p}(1-\bar{p})} + (z_{\beta})\sqrt{p_1(1-p_1) + p_2(1-p_2)}}{\Delta} \right]^2$$

Plugging in the values we get:

$$n = \left[ \frac{1.96\sqrt{2(0.445)(0.555)} + (0.84)\sqrt{(0.52)(0.48) + (0.37)(0.63)}}{0.15} \right]^2 = 171.18 = 172$$

To adjust for 20% non-response, we divide the sample size by 172: 172 / 0.8 = 215. Therefore, we recruit 215 participants per group for a total sample size of 430.

**2. You found another study similar to yours, and this time it provided an odds ratio of older adults vs younger adults of 0.9. Given the same assumed proportion among younger adults of 0.52, the same $\alpha$ and $\beta$, and the same 20% adjustment for non-response, what is the new sample size?**

*Hint: The missing value here is the proportion among older adults, which you would need to derive from the given proportion of younger adults and the odds ratio. Recall the following formulas from the Logistic Regression lecture*

$$\text{odds} = \frac{p}{1-p}; p = \frac{\text{odds}}{1+\text{odds}}$$

First, we calculate for the proportion among older adults. To calculate that, we need the odds among older adults. And to calculate that, we need the odds of younger adults. And finally, to calculate that, we need the proportion among younger adults, which is given in the question. Essentially, this is a three step solution.

$$\text{odds}_{\text{younger}} = \frac{p_{\text{younger}}}{1 - p_{\text{younger}}} = \frac{0.52}{1 - 0.52} = 1.0833$$

$$\text{odds}_{\text{older}} = \text{odds}_{\text{younger}} \times \text{odds ratio}_{\text{younger vs. older}} = 1.0833 \times 0.9 = 0.975$$

$$p_{\text{older}} = \frac{\text{odds}_{\text{older}}}{1 + \text{odds}_{\text{older}}} = \frac{0.975}{1 + 0.975} = 0.494$$

The given values now are as follows:

- $z_{\alpha/2} = 1.96$ given $\alpha = 0.05$
- $z_\beta = 0.84$ given $\beta = 0.2$
- $p_1 = 0.52, 1 - p_1 = 1 - 0.52 = 0.48$
- $p_2 = 0.494, 1 - p_2 = 1 - 0.494 = 0.506$
- $\Delta = p_1 - p_2 = 0.026$
- $\bar{p} = \frac{p_1+p_2}{2} = \frac{0.52+0.494}{2} = 0.507; 1 - \bar{p} = 0.493$

Plugging in the values we get:

$$n = \left[\frac{1.96\sqrt{2(0.507)(0.493)} + (0.84)\sqrt{(0.52)(0.48) + (0.494)(0.506)}}{0.026}\right]^2 = 5796.5 = 5797$$

To adjust for 20% non-response, we divide the sample size by 172: 5797 / 0.8 = 7246.25 = 7247. Therefore, we recruit 7247 participants per group for a total sample size of 14494.

**3. Why did the sample size balloon to such a high number from #1 to #2?**

The proportion between the two groups were much smaller in #2 and #1, and therefore, for the same alpha and beta, we need a much larger sample size to detect a smaller difference between the two groups.

**4. Confused between which sample size to use, you decide to explore a range of sample sizes across values of power and odds ratio. You look at power = 0.80 and 0.90, and odds ratios of older vs. younger groups of 0.5 to 0.9 in 0.1 increments. The same $\alpha$, the same assumed proportion among younger adults of 0.52, and the same adjustment for 20% non-response are going to be used.**

| Odds ratio | Total sample size, 80% power | Total sample size, 90% power |
|---|---|---|
| 0.5 | 337 | 449 |
| 0.6 | 610 | 816 |
| 0.7 | 1241 | 1660 |
| 0.8 | 3157 | 4226 |
| 0.9 | 14148 | 18938 |

**Based on the table above, the lowest sample size is achieved with odds ratio = 0.5 and 80% power. Calculate the odds ratio of the proportions in #1 and comment on how the sample size in #1 compares to the table.**

*Hint: Recall the formula for the odds ratio using the proportions:*

$$\text{odds ratio} = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}$$

The given values are:

- $p_1 = 0.52, 1 - p_1 = 1 - 0.52 = 0.48$
- $p_2 = 0.37.1 - p_2 = 1 = 0.37 = 0.63$

Plugging in the values we get:

$$\text{odds ratio} = \frac{\frac{0.37}{0.63}}{\frac{0.52}{0.48}} = 0.54$$

The sample size calculated in #1 (adjusting for 20% non-response) was 430. Given the odds ratio = 0.54, the calculated sample size is between OR = 0.5 and OR = 0.6 on the table above.

END OF LAB