Prof. David Draper
Department of Statistics
University of California, Santa Cruz Winter 2022

STAT 206: Quiz 2 *[145 total points]*

Name: Kevin Guillen

You're an economist interested in patterns of unemployment of U.S. adults over time. As part of this interest, You decide to take a sample from the population $\mathcal{P}$ of people 18 years of age or older who were living in Santa Cruz (city, not county) as of time T = (1 Jan 2022). The most recent U.S. census, extrapolated to the beginning of 2022, estimates the total population of the city of Santa Cruz at that time as approximately 64,500, and data sets from the website `suburbanstats.org` lead to an estimate of $N \doteq 54{,}300$ as the approximate total number of those people whose age was at least 18 at time T. You decide to take a representative sample of $n = 921$ people from $\mathcal{P}$ and ask each sampled person "Do you consider yourself fully employed at the time of this survey?", with possible responses {*yes, no, other* (e.g., refuse to answer)}.

Let θ be the proportion of the 54,300 people who would have answered *yes* to this question, if You had been able to survey the entire population $\mathcal{P}$, and let s (an integer between 0 and n, inclusive) be the number of people in Your sample who actually do answer *yes*. By way of context, the latest national estimate of the percentage of people 18 years old or older who regard themselves as fully employed is about 75% (lots of American adults who want full employment can only find part-time jobs).

(1) ***How should You choose Your sample?***

(a) In class we agreed that the simplest method for obtaining a *representative* sample from a (finite) population is *random sampling*. However, given that there's no list of {all N people, with their addresses and other contact information} from which You could draw a random sample (which is true; for one thing, what about homeless people?), in practice would it be easy, hard, or in between for You to construct a sample that You and other reasonable people would agree is representative (*like* a random sample) from the population $\mathcal{P}$? Explain briefly. *[10 points]*

> ***Solution***. It would be hard to construct a sample that is representative of from the population. Since it would be hard to take a like IID random sample since we have no list (as defined in the problem statement) in which we draw a random sample from. Even if we did have said list, it would still be hard because there can be cases where we repeatedly try to contact individuals drawn from the list and they repeatedly don't reply to us. □

(b) Describe (e.g., continuing on another sheet of paper, if You're writing Your answers in longhand) how You personally would attempt to obtain an arguably representative sample from $\mathcal{P}$. *Hint:* There's a part of the U.S. government devoted to representing the entire population; how could public data from that agency help here? *[10 points]*

> ***Solution***. The most recent data avaiable from the US Census isn't too helpful here since their latest data is from 2020, while we are trying to work with a sample from the population in 2022. One thing we might be able to do is take the trends from older data up to 2022 and predict how the population might look now in 2022 to help obtain a represntative sample from $\mathcal{P}$, but this would then be another thing we would have to try to calculate/assume. □

For the rest of this problem, let's assume that You have indeed been able to create a sample that's similar to what You would have obtained with IID random sampling, and that Your results were as follows: $n_{yes} = s = 728$ people said *yes*, $n_{no} = 174$ said *no*, and $n_{other} = 19$ were recorded as *other*.

(2) Before You get Your sampled data, is the logical status of θ known or unknown? What about s? Answer both questions at a moment in time after Your sample data has arrived. *[10 points]*

> ***Solution***. The logical status of θ and s before we get our sampled data is unknown. The reason for θ being unknown is because that would imply we took a complete census of the entire population which we haven't. Since s depends on our data we of course can't calculate it before our data arrives.
> After our data, θ is still unknown since our sample, n, is less than N. However we do know s since we can now perform that calculation since it depended on our sample. □

(3) In class we saw that calculations relevant to uncertainty quantification were of two types — *probability* and *statistics* — and that statistical activities in turn were of six types — *optimal design of data-gathering*, *data curation*, *description*, *inference*, *prediction*, and *decision-making* — making a total of seven classes of methods (***Pillars of Statistical Data Science***) relevant to STAT 206. For each of the following (*[5 point each]*), identify the activity or calculation as one of these seven classes, and briefly explain Your choice.

(a) After the data are available, You estimate that a future sample survey of size $n_{future} = 614$ from $\mathcal{P}$ in early 2023 would contain about $\hat{n}_{yes} = 485$ *yes* responses.

Solution. This would be *prediction* since it is an estimation of new data. □

(b) Before the data set arrives, and temporarily pretending that θ is known, under IID random sampling the sampling distribution (probability mass function) of s given θ (and n) is $(s \,|\, n\, \theta\, \mathcal{B}) \sim \text{Binomial}(n, \theta)$, where $\mathcal{B}$ summarizes the background context of Your sample survey.

Solution. This would be *probability* since we're working in a moment in time before the data is known and pretending we know θ. We know from Stats 131 that a sum of IID bernoulli θ's has a binomial distribution. We know they are bernoulli since we have our data as 1 or 0 for desired employment and not desired employment. □

(c) In consultation with You and on the basis of Your survey, the Santa Cruz City Council votes (5 in favor, 2 opposed) to allocate \$57,300 in the fiscal year 2022–2023 budget to be distributed to winning grant proposals for ways to reduce unemployment in the city.

Solution. This would be *decision*, since an action is being taken based on an analysis. □

(d) You and Your assistants have been interviewing the sampled people to obtain Your data set, and You've found that the people You interview are less intimidated by You if You record their answers on paper, to be transcribed later into a computer file for analysis. On the resulting sheets of paper, during the transcription process You discover that 11 of the "data" responses are illegible. You're now thinking about what to do about this problem.

Solution. This would be *data curation* since we are figuring out how to cope with the errors in our data, and imputation problem. □

(e) After the data set has been collected, assuming no bias in Your sampling method and using frequentist inference, You estimate that θ is about $\hat{\theta} = \frac{s}{n} = \frac{728}{921} \doteq 79.0\%$, with a give-or take of about 1.3% and a 99.9% interval estimate of about $(74.6\%, 83.5\%)$.

Solution. This would be *inference* since we are infering about θ based on $\hat{theta}$, and after some calculations it is indeed a correct inference. □

(f) You summarize Your data set with the vector $(n_{yes}, n_{no}, n_{other}) = (728, 174, 19)$.

Solution. This would be *description* because it a numerical summary of existing data. □

(g) Before the survey is conducted, You work out that $n = 921$ people sampled in a like-at-random manner will be sufficient to estimate θ with a small enough level of uncertainty to support good decisions about how to decrease unemployment in Santa Cruz.

Solution. This would be a *optimal design of data-gathering* since it is under sample size determination. □

Table 1: *Four imputation methods in the unemployment case study; see text for definitions of* n_{total}, $\hat{n}_{total}$, $\hat{\theta}_I$, *and* $\hat{n}_{no}^D$.

	Imputed					Numerical Value	
Method	$\hat{n}_{yes}$	$\hat{n}_{no}$	$\hat{n}_{other}$	$\hat{n}_{total}$	$\hat{\theta}_I$	$\hat{\theta}_I$	$\hat{n}_{total}$
(A)	$(n_{yes} + n_{other})$	n_{no}	0	n_{total}	$\frac{\hat{n}_{yes}}{\hat{n}_{total}}$	0.8111	921
(B)	n_{yes}	$(n_{no} + n_{other})$	0	n_{total}	$\frac{n_{yes}}{n_{total}}$	0.7904	921
(C)	n_{yes}	n_{no}	0	902	$\frac{n_{yes}}{902}$	0.8071	902
(D)	$\hat{n}_{yes}^D$	$\hat{n}_{no}^D$	0	$(\hat{n}_{yes}^D + \hat{n}_{no}^D)$	$\frac{\hat{n}_{yes}^D}{\hat{n}_{yes}^D + \hat{n}_{no}^D}$	0.8071	921

In estimating the unemployment rate in $\mathcal{P}$ at time T, You have to decide what to do about the $n_{other} = 19$ people who answered *other* (their lack of yes–no responses plays the role of *missing data* in this problem). During the ***data curation*** step of Your analysis, the standard approach to coping with missing data is something called *imputation*; this is an attempt to *predict* what the missing data values would have been if they had not been missing. Here are four natural ways to perform the imputation in this problem.

(A) At one extreme You could imagine that all 19 of those people would have answered *yes* if they had given a *yes*/*no* answer;

(B) At the other extreme You could imagine them all answering *no*;

(C) You could just remove them from the data set (some statistical computing packages make this choice for You without necessarily telling You that they did so); or

(D) You could spread the *other* people out proportionately across the *yes* and *no* categories, based on the observed *yes* and *no* prevalences.

Table 1 summarizes these four imputation methods in this case study; in the table, $\hat{\theta}_I$ is the imputed estimate of θ with the indicated method, $n_{total} = (n_{yes} + n_{no} + n_{other}) = n$, $\hat{n}_{total} = (\hat{n}_{yes} + \hat{n}_{no} + \hat{n}_{other})$, and the $\hat{n}_{no}$ value with method (D) is as follows:

$$\hat{n}_{no}^D \triangleq n_{no} + \left(\frac{n_{no}}{n_{yes} + n_{no}} \right) \cdot n_{other} \,. \tag{1}$$

(4) (a) Complete (impute?) the missing (blank) entries in Table 1, briefly explaining your reasoning in each case. *[30 points]*

Solution. Under method (A) the responses with *other* get treated as *yes* responses. Which is why $\hat{n}_{yes}$ is equal to $(n_{yes} + n_{other})$ which means we have to recalculate $\hat{\theta}_A$ as $\frac{\hat{n}_{yes}}{\hat{n}_{total}}$ since we have a different number n_{yes} now.
Under method (B) $\hat{n}_{yes} = n_{yes}$ since in this case we imagine all the responses of *other* to be changed to *no* leaving the total responses of yes unchanged. Our numerical value for $\hat{\theta_B}$ is simply calculated by $\frac{728}{921}$.
Under method (C) since we are simply removing the *other* responses the number of *no* responses remains unchanged, therefore $\hat{n}_{no} = n_{no} = 174$. Because though we are removing the *other* responses our total number of responses decreases, $\hat{n}_{total} = n_{total} - n_{other} = 921 - 19 = 902$.
Under method (D) we can calculate $\hat{n}_{yes}$ in a similar fashion to $\hat{n}_{no}$ by the given equation,

$$\hat{n}_{yes} = \hat{n}^D_{yes} = n_{yes} + \frac{n_{yes}}{n_{yes} + n_{no}} \cdot n_{other} = 743.3348$$

this is because we are spreading the *other* responses out proportionally across the *yes* and *no* responses, based on their prevalences. Now for $\hat{\theta}_I$ we have the following,

$$\hat{\theta}_I = \frac{743.3348}{743.3348 + 177.6652} = 0.8071$$

and our new total we simply be $\hat{n}^D_{yes} + \hat{n}^D_{no} = 921$. □

(b) How do the $\hat{\theta}_I$ numerical values compare with each other and with Your $\hat{\theta}$ in problem 3(e) above? Explain briefly. *[10 points]*

Solution. Well right away we see,

$$\hat{\theta} = 0.7904 = \hat{\theta}_B < \hat{\theta}_C = 0.8071 = \hat{\theta}_D < \hat{\theta}_A = 0.8111$$

which means under method (B), simply changing all the *other* responses to *no* yields the same $\hat{\theta}$ as our estimate from 3(e).
We also see that two different methods, (C) and (D), will yield the same result as one another. Specifically disregarding the *other* responses or spreading out the *other* responses in a proptional manner will give the same value for $\hat{\theta}$.
Finally we see the largest $\hat{\theta}_i$ is when we change all the *other* responses to *yes* responses, which could likely mean there could be a bias in favor of *yes* here. By similar reasoning method (B) might be bias towards the *no* responses.
Meaning method (D) and (C) are a good compromise, but method (D) exaggerates effective sample size. Meaning method (C) seems to be the best, dropping the *other* responses. □

(c) Someone says, "There were only 19 *other* responses out of almost 1,000 people in Your sample, so it doesn't matter what You do with them in Your analysis." Looking at the range of $\hat{\theta}_I$ values, from smallest to largest, across the four imputation methods, would you agree with that statement? Explain briefly. *[10 points]*

Solution. Well we see from $\hat{\theta}_A - \hat{\theta}_B = .0207$ or about 2 percentage point difference. Recall though our standard error from 3(e) is about 1.3 percentage points, meaning 2 percentage points is about 1.5 standard deviations. So I would disagree with this statement, since how we treat the *other* responses can yield big differences. □

(d) In what ways, if any, do the results from imputation methods (C) and (D) differ? Explain briefly. *[10 points]*

Solution. Well firstly we as noted before we see that $\hat{\theta}_C = \hat{\theta}_D$, but we see,

$$902 = \hat{n}^C_{total} < \hat{n}^D_{total} = 921$$

but the total for (D) is too high since it is claiming the same sample size as if we have seen all the 19 people, while (C)'s total is just disregarding the responses we don't know. □

To figure out which of the imputation methods is best, it turns out (this should make good sense to You, based on the *No Free Lunch* principle) that You need a probability model for the missing data that quantifies Your judgments about *why* the missing data values are missing. In the imputation literature the simplest such model is called ***Missing Completely At Random (MCAR)***; this model assumes that the missing subjects are themselves like an IID random sample from the population of interest to You.

(e) Considering the set of five estimates of θ formed by collecting together the four $\hat{\theta}_I$ values and $\hat{\theta}$ in problem 3(e), choose the estimation method that is best under MCAR, and briefly justify Your choice. *[10 points]*

Solution. Under MCAR the missing 19 people are like IID just in the same way the non-missing 902 people are like IID from the population, meaning there it is unbiased. Which means it is okay to drop the 19 missing people from our sample, which matches with method (C) □

(f) Some people find the conclusion in (e) initially surprising. Is it still surprising after a bit of careful thought? Explain briefly. *[10 points]*

Solution. After careful thought you realize after you assume MCAR, then it's the best thing to do. Method (D) may seem nice because it assums the proportion of *yes* and *no*'s in the missing peoples responses will be similar to that of the non-missing people, but it assumes that about data we don't actually know. While under MCAR since the missing people are like IID the sameway the non-missing people are, then it's safe to just drop them and thereby assuming nothing about data we don't know. □