Prof. David Draper
Department of Statistics
University of California, Santa Cruz
Winter 2024

STAT 206: Quiz 1 *[190 total points]*

Name: Devanathan Nallur Gandamani (2086936)

You're an econometrician interested in patterns of employment and 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 living in Santa Cruz (city, not county) as of time $T =$ (1 Jan 2024) who were eligible to work. Employment is a somewhat complicated variable to measure: one way is to treat it as qualitative with the following categories, which form a (mutually exclusive and exhaustive) partition set.

- (I) not employed (not eligible for paid work: too young [0–13 years old]);
- (II) not employed (eligible, not working, not looking for work);
- (III) not employed (eligible, not working, looking for work);
- (IV) partially employed (eligible, working, but person states that they wish to work more hours per year than they currently do); and
- (V) fully employed (eligible, working, and person states that they're working as much as they want to).

A variety of web searches reveals that early in the current decade (i.e., 2020–2023) the percentages of people in Santa Cruz county (not city) in these five categories were approximately $(15, 5, 10, 10, 60)\%$.

The most recent U.S. census, extrapolated to the beginning of 2024, estimates the total population size of the city of Santa Cruz at that time as approximately 63,900, with approximately $N \doteq$ 54,300 as the approximate size of $\mathcal{P}$ (the total number of Santa Cruz city residents who were eligible to work). 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 [category (V)], no [categories (II)–(IV)], 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*.

(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 **frame** (a 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 a difficult task to generate a random sample like which is representing the population. Sanat Cruz county is a diverse area with different socio-economic class, employment categories. In order to precisely take in account of the variability into a sample, complex stratification is required. Practical limitations also is a factor like Reaching out to people, coordination of data which could be resource heavy.

Certain key aspects that influence this:

(i) **Selection Bias ::** If one has to gather information, it would be done so by trying to fetch from my close circles or who someone who actually come into bucket of employment. Hence, it is difficult to categorize people just by looking at them.

(ii) **Diversity ::** Gathering input from diverse set of target population is difficult. For instance, we could go to different hotspots of Santa Cruz county but that doesn't guarantee a diverse pick.

(iii) **Privacy intrusion ::** Everybody in the county would not be very comfortable sharing their private and sensitive information outside. This is a complicated process where getting consent is really an important aspect to keep in mind for recording their information.

(iv) **Accessibility ::** Manual collection of information by approaching people directly is a tedious and cumbersome process and is very hard.

Given these limitations strategies that I could adopt are:

(i) Understanding demographic composition from public available data from authentic source such as US Census. This might aid in dividing the population into different sections based on key features (age, gender etc).

(ii) What could be a tricky challenge is how we handle Homeless people. They do not have permanent address. Collaborating with local agencies and NGOs can be a step to solve this problem.

(iii) Approaching and reaching out to people via Online / Digital Mediums might be a feasible solution when considering reach and manual efficiency.

(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 several parts of the U.S. government devoted to representing the entire population, e.g., the `Census Bureau` ; how could public data from those agencies help here? ***[10 points]***

Solution:

A legitmate / authentic source of information like United State Census or in general a multifaceted way that fetches from available data to obtain a representative sample from Population P in Santa Cruz County by:

(i) In order to produce statistics that define populations and their characteristics—such as age, education, housing, and income—the United States Census Bureau collects data through demographic surveys. The ability to create strata based on factors like age, gender, work status, and ethnicity is made possible by the data that has been gathered, which is crucial for understanding Santa Cruz's demographic composition. When certain groups—such as the homeless or people living in unconventional housing—are underrepresented in Census data, I would look for other sources of information or work with neighborhood organizations to obtain a more thorough understanding and inclusion of these communities.

(ii) It is possible to research employment trends in Santa Cruz City. The Bureau of Labor Statistics' data is one source used in this research. It can provide a strong foundation for comprehending local employment circumstances, which is essential for creating a stratified sample that accurately represents Santa Cruz residents' employment status. This information is used to determine the representativeness of the sample and to guide stratification.

(iii) Since a proper list of people isn't available, it's best to use different resources, such as voter registration lists and utility customer databases, to create a proxy list.

I'll make an effort to use the data I've gathered to ensure that the people I hire are diverse in terms of age, profession, skill level, race, etc. While it is still not a random sample, it is now more closely related to a likely random sample.

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 from $\mathcal{P}$, and that Your results were as follows: $n_{yes} = s = 662$ people said *yes*, $n_{no} = 240$ 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? Then answer both questions at a moment in time *after* Your sample data has arrived. ***[10 points]***

Solution:

(i) **Before Obtaining Sample Data** - It is very hard to determine the logical status of θ. The percentage of Population P that would respond "yes" is **Status**: Unknown. This represents the portion of the population that believes they are fully employed. This proportion is unknown prior to survey completion because we are trying to estimate it through sampling. The sampled population's "yes" response rate is .Status: Not specified. - s is the proportion of sample participants who responded "yes" when asked if they were fully employed. Before the survey is sent out and responses are gathered, we have no idea how many people will answer "yes."

(ii) **After Obtaining Sample Data** - Even after collecting the sample data , status is still unknown. The sample provides an estimate, but the precise value of full Population P remains unknown.s - **Status**: Known. - After gathering the data, s is known. According to the data, 662 persons out of a total of 921 answered 'yes'. So, s is no longer an unknown quantity.

$$\theta = \frac{n_{yes}}{n_{yes} + n_{no} + n_{other}} = \frac{662}{662 + 240 + 19} \approx \mathbf{0.7187}$$

(3) In class we saw that calculations relevant to uncertainty quantification were of six types (the ***7 Pillars of Statistical Data Science***, omitting the first item ***Problem Formulation***):

- ***Probability***;
- ***Design of data-gathering activities***, including ***sample size determination***;
- ***Data curation***, including ***description of existing data sets*** and ***treatment of missing data***;
- ***Inference***;
- ***Prediction***; and
- ***Decision-making***.

For each of the following (***[10 points each]***), identify the activity or calculation as one of these six 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 2025 would contain about $\hat{n}_{yes} = 441$ *yes* responses.

Solution: This can be put under **Prediction**, since we are predicting on an unobserved aspect in the future dataset based on the current dataset. Utilizing past data, predictive statistics forecast future events with confidence. In early 2025, a future sample survey with a size of n[future] = 614 from P is anticipated to yield about 441 yes answers. It involves projecting future outcomes using models or data that already exists.

(b) Before the data set arrives, and temporarily pretending that θ is known, under IID random sampling the sampling model $[SM]$ (in this case, probability mass function) of S given θ (and n) is $(S \,|\, n\,[SM : \mathbb{B}]\,\theta\,\mathcal{B}) \sim \text{Binomial}(n, \theta)$, where $\mathcal{B}$ summarizes the background context of Your sample survey and $\mathbb{B}$ identifies the Binomial $[SM]$.

Solution: This relates to **Probability**; specifically, we're attempting to ascertain which probability approach—Classical, Frequentist, or Bayesian—to employ along with its distribution (or mass function). It is a probability exercise to describe the sampling procedure under IID random sampling using a binomial model prior to the arrival of the data set. This step involves modeling the generation of data by learning and using probability principles.

(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 \$61,900 in the fiscal year 2025–2026 budget to be distributed to winning grant proposals for ways to reduce unemployment in the city.

Solution: This falls under **Decision-making**. The council decides to set aside a portion of the budget to reduce unemployment in light of our conclusion. The decision-making process was satisfied by the Santa Cruz City Council's vote to distribute funds based on your survey. The survey results are used by the City Council to inform policy decisions regarding budget allocation. This illustration shows how statistical information can influence and directly educate real decision-making processes in business and government.

(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 (rather than, e.g., inputs to a laptop), to be transcribed later into a `.csv` or `.txt` 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 is **Data Curation**. It is necessary to preserve the integrity and quality of the data. It entails handling and resolving problems with data quality, such as unclear or missing information and illegible responses. Monitoring, maintaining,

and guaranteeing data quality throughout its life cycle are all part of proper data curation. In other words, since we're attempting to decide how best to use the data that has already been gathered.

(e) After the data set has been collected, assuming no bias in Your sampling method and using frequentist reasoning, You estimate that θ is about $\hat{\theta} = \frac{s}{n} = \frac{662}{921} \doteq 71.9\%$, with a give-or take of about 1.5% and a 99.9% confidence interval of about $(67.0\%, 76.8\%)$.

Solution: This falls under the category of **statistical inference**, in which the goal is to extrapolate the findings from a sampled dataset to the global populace. Inference is the process of estimating with a confidence interval after data collection, assuming no bias, and applying frequentist reasoning. It comprises using information from a sample to draw conclusions about a population parameter.

(f) You summarize Your data set with the vector $(n_{yes}, n_{no}, n_{other}) = (662, 240, 19)$.

Solution: This falls under the category of **Data Curation**. It includes summarizing your data set with a vector $(n_{yes}, n_{no}, n_{other})$. This entails describing and summarizing the data set, which is an important step before further research.

(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 might be a part of **Design of data-gathering activities**, where we're trying to determine the sample size that would be required for the survey. Developing the procedures for gathering data, including figuring out the sample size The process of estimating with minimal uncertainty involves designing data-gathering activities and using a randomly selected sample size of 921 individuals. This involves determining the sample size needed for the survey to generate estimates with a specific degree of accuracy.

Table 1: *Four imputation methods in the full employment 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}	$\frac{n_{yes}+n_{other}}{n_{total}}$	0.7394	921
(B)	n_{yes}	$(n_{no} + n_{other})$	**0**	n_{total}	$\frac{n_{yes}}{n_{total}}$	**0.7187**	921
(C)	n_{yes}	n_{no}	0	$n_{yes} + n_{no}$	$\frac{n_{yes}}{n_{yes}+n_{no}}$	0.7339	**902**
(D)	$\hat{n}_{yes}^D$	$\hat{n}_{no}^D$	**0**	$(\hat{n}_{yes}^D + \hat{n}_{no}^D)$	$\frac{\hat{n}_{yes}^D}{\hat{n}_{yes}^D+\hat{n}_{no}^D}$	**0.7339**	**921**

In estimating the full employment 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 proportionally 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 the missing (blank) entries in Table 1, briefly explaining your reasoning in each case. ***[40 points]***

Solution: In all cases we are trying to eliminate n_{other}. Hence it is 0 in all 4 cases.
Method (A): We take other answers as yes. Hence, **no**. of yes is increased by n_{other}. Total remains same, so is ratio of new no. of yes and total.
Method (B): We take other answers as **no**. Hence, no. of yes is unchanged but no. of nos is increased by n_{other}. Total remains same, so is ratio of old no. of yes and total = 0.7187.
Method (C): We ignore other values. Hence, no. of yes is unchanged as well as no. of nos. Total remains same, so is ratio of old no. of yes and total.
Method (D): We compute new values by distributing values of others based on the

distribution we get from the collected values into yes and no sets. From the formula given, $n_{yes} = 676$, $n_{no} = 245$. So θ is ratio of new no. of yes and new total ≈ 0.7339.

(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: The value of $\hat{\theta}_I$ in 3(e) appears to be nearly identical to that found in Method (B). It appears that other $\hat{\theta}_I$ values differ by 0.02, or 2%, if we convert it. Compared to the original $\hat{\theta}$ estimate from problem 3(e), which was 0.719 (71.9%), we see that methods (A) and (C)/(D) provide a higher estimate, while method (B) gives a lower one. This suggests that the way missing data is handled can significantly impact the estimated proportion, potentially leading to overestimation or underestimation of the true rate of full employment in the population. Given that the other values are smaller than the yes and no values, this indicates that they have less of an influence on the estimation.

(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: Yes, I agree with the statement given in the question. As the table suggests, the impact is less significant since it creates a 2% fluctuation in the result. On the contrary, Considering the range of $\hat{\theta}_I$ values across the imputation methods, the statement that the 'other' responses don't matter is not entirely accurate. The variation in $\hat{\theta}_I$ values shows that even a small number of responses, when imputed differently, can affect the overall estimate. This is particularly relevant in statistics where precise estimates are important, and even small changes can impact conclusions.

(d) In what ways, if any, do the results from imputation methods (C) and (D) differ?

Explain briefly. ***[10 points]***

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 information and 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: While method (D) distributes the "other" category proportionately based on the observed prevalences of "yes" and "no" responses, method (C) eliminates the category entirely. The $\hat{\theta}_I$ numerical estimate came out to be 0.7339, and the results from both approaches were identical. This may not always be the case with different data due to the particular numbers involved and the rounding used in this instance. I believe that under MCAR, method (c), which entails ignoring the undefined values, makes more sense. There was a choice between (c) and (d), but selecting (d) could cause the distribution's variance to significantly decrease.

(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: The fact that a better fitted model is obtained for the given sample space when the undefined values are ignored is, to be honest, not surprising. A result that differs greatly from the initial one could be obtained by treating the other values as yes. The same holds true for treating all values as false. (d) appears to be a good choice, but only if the distribution's mean is a concern. We must consider additional statistical measures when projecting the result for a future space, and it is obvious that only choice (c) is appropriate. But, The conclusion in (e) may come as a surprise since it implies that the handling of the missing data can affect the estimate even in cases of random missing. After all, though, it seems reasonable that using all of the information at hand—including the known response distribution—will yield a more precise estimate when operating under the MCAR assumption.