Prof. David Draper
Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

AMS 206: Quiz 2 *[45 total points]*

Name: _____

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 $T =$ (16 Jan 2019). The most recent U.S. census, extrapolated to the beginning of 2019, estimates the total population of the city of Santa Cruz at that time as 64,465, and data from the website `suburbanstats.org` lead to an estimate of $N \doteq 54{,}342$ as the 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 54,342 people who would have answered *yes* to this question, if You had been able to survey the entire population, and let s (an integer between 0 and n, inclusive) be the number of people in Your sample who actually do answer *yes*.

(1) In class we agreed that the simplest method for obtaining a *representative* sample from a (finite) population is *random sampling*. 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. *[5 points]* Describe (e.g., on another sheet of paper) how You personally would attempt to obtain an arguably representative sample from $\mathcal{P}$. *[5 points]*

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 random sampling, and that Your results were as follows: $n_{yes} = s = 830$ people said *yes*, $n_{no} = 72$ 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. *[5 points]*

(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 four types — *description*, *inference*, *prediction*, and *decision-making* — making a total of five classes of methods relevant to AMS 206. For each of the following (*[5 point each]*), identify the activity or calculation as one of these five 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 2020 would contain about $\hat{n}_{yes} = 553$ *yes* responses.

(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.

(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 2020 budget to be distributed to winning grant proposals for ways to reduce unemployment in the city.

(d) After the data set has been collected, You estimate θ to be about $\hat{\theta} = \frac{s}{n} = \frac{830}{921} \doteq 90.1\%$, with a give-or take of about 1.0% and a 95% interval estimate of about $(88.2\%, 92.0\%)$.

(e) You summarize Your data set with the vector $(n_{yes}, n_{no}, n_{other}) = (830, 72, 19)$.

(4) 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*. One possible approach is *sensitivity analysis*: at one extreme You could imagine that all 19 of those people would have answered *yes* if they had given a *yes*/*no* answer, and at the other extreme You could imagine them all answering *no*. This defines a range of possible unemployment rate estimates, and if this range is narrow enough You've demonstrated that it doesn't matter much what You do with the *other* people. Compute the lower and upper endpoints of this range with the data set in this problem. Would You say that the effect of the *other* people is negligible here? Explain briefly. *[5 points]*