Data 8 Spring 2016 hw07, Due: 8pm Tuesday, April 5

Please write your answers in the space provided. You can write on a printed copy or fill in the blanks with a PDF editor such as Acrobat Reader or Apple Preview. (Beware: some people have found that editing a PDF in a browser doesn't work.) When you're done, upload a scanned copy to Gradescope (`gradescope.com`). This assignment is due 8pm Tuesday, April 5. You will receive an early submission bonus point if you turn it in by 5pm Monday, April 4.

You are welcome to use `ds8.berkeley.edu` to try out Python expressions. Directly sharing answers is not okay, but discussing problems with course staff or students is encouraged.

Collaborators:

Problem 1 (Stents)

A stent is a kind of medical device – a tube implanted in a person's artery to shore it up when it is in danger of becoming blocked. A certain stent is designed to have a very precise thickness to match an artery's width, but the stent's manufacturing process makes small errors, so every stent has a slightly different thickness. These errors in thickness can make the device work less well. You're evaluating the manufacturing process, and you want to know the average amount of error in thickness among all the stents the manufacturer has made so far.

You have a machine for measuring the thickness of a stent, but when it measures a stent, it renders it unsuitable for implantation. So you can't just measure all the stents. Instead, you choose 10,000 stents uniformly at random from among all the manufactured stents, and you measure the errors in their thicknesses. (Error is measured as the absolute value of the difference between each stent's actual thickness and its designer-specified thickness.) Then you compute the average of those 10,000 numbers. For the purpose of the questions below, that average is the *statistic* you're working with in this scenario.

- What population parameter are you trying to estimate with this statistic?
  1(a):

- What is the population?
  1(b):

- You're worried that this statistic isn't a good enough estimate of the population parameter. Describe how you would use an inferential technique you've learned to make a quantitative claim about the population parameter *that somehow conveys your uncertainty.* Assume you have access only to the 10,000 measurements you've made.

  1(c):

Answer:

(a) The average error in the thickness of all the stents manufactured so far.

(b) All the stents manufactured so far.

(c) We approximate the sampling distribution of our statistic by resampling from our sample, then report an approximate confidence interval computed from the distribution of the means of those samples.

Here's a more concrete answer. We could take 4,000 resamples (samples with replacement, each with size 10,000) from our sample and compute the mean error for each resample. Then, to find an approximate 98% confidence interval for the population parameter, we could take the 1st and 99th percentiles of this distribution of resample means. (Other choices of the confidence level are justifiable.)

Problem 2 (Pollution Confusion)

Air pollution is a serious health concern in many cities around the world. Suppose that last year, before you took this course, you lived in Beijing and wanted to measure the average amount of fine particulate matter (known as $\texttt{PM}_{2.5}$) across the city on March 30. We'll call that the "average $\texttt{PM}_{2.5}$" for short. You couldn't get to every place in the city, so you measured the $\texttt{PM}_{2.5}$ on March 30 at the 40 street corners nearest your apartment. You decided to use the average of those measurements as an estimate of the average $\texttt{PM}_{2.5}$.

You knew that your sample didn't include all the locations in the city, so your estimate was prone to error. To reflect your uncertainty, you decided to compute an interval of estimates you might reasonably have seen instead. You took 10,000 resamples (uniform random samples with replacement) of size 40 from your sample, computed the average of each resample, and claimed that the 2.5th and 97.5th percentiles of those resample averages formed a 95% confidence interval for the average $\texttt{PM}_{2.5}$.

- Did you use a random sample or some other kind of sample in this study?
  2(a):

- If you had repeated this study many times, would your reported interval have contained the true average $\texttt{PM}_{2.5}$ in roughly 95% of the repetitions? Why or why not?

  2(b):

- Describe a problem with the design of the study and recommend a fix.

  2(c):

Answer:

(a) Some other kind of sample. This one might be called a *convenience sample*, since we sampled locations in Beijing where it was convenient for us to take measurements.

(b) No. There are different reasonable interpretations of what it could mean to "repeat" this study, but it is clear that we are not sampling randomly from all the locations in the city. Suppose there is a location far away from our apartment that has an extremely high $\texttt{PM}_{2.5}$ level, enough to dramatically raise the city's average. We had no chance of sampling that location, so our intervals might *never* include the city's true average. This is why the standard methods for generating confidence intervals are only justified when you have a random sample from the population you're interested in studying. It wasn't appropriate to describe our interval as a confidence interval in this case.

(c) We didn't use a random sample, so our estimate of the average $\texttt{PM}_{2.5}$ may be biased. Perhaps we live in an unusually unpolluted part of the city, in which case we will consistently underestimate the average $\texttt{PM}_{2.5}$. As noted in (b), our "confidence intervals" might also mislead us about how accurate our estimate is!

The simplest fix is to take a uniform random sample from all the locations in the city.

We could also reduce the scope of our study: give up on measuring pollution throughout the city and be satisfied with measuring only local pollution. If we decided to measure the average pollution at the 40 nearest street corners, our estimate of that quantity is exact, and there is no need for a confidence interval.

(This space is left intentionally blank. There's one more problem on the next page.)

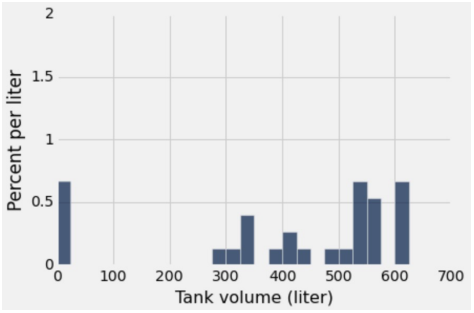
Problem 3 (The Statistician)

Astronaut Mark Watney is stranded on Mars with 1000 tanks of oxygen, each with a different unknown amount of oxygen remaining. He needs to know how much oxygen he has left, so he opens 30 randomly-selected tanks to measure their oxygen levels. He finds that the average oxygen level among the 30 sampled tanks is 409 liters. In fact, the average oxygen level among the population of 1000 tanks is 432 liters, though of course Mark doesn't know that.

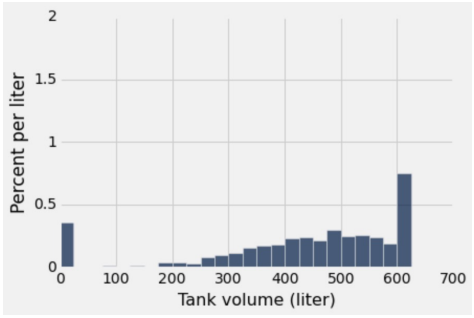
Below are four histograms displaying four distributions involved in this question. Match each item below to the histogram it most plausibly describes, **using each histogram exactly once**. (Just write i, ii, iii, or iv for each part.)

- The distribution of oxygen levels among all 1000 tanks. (Mark doesn't know this.)
  3(a):
- The distribution of oxygen levels among the 30 sampled tanks.
  3(b):
- The sampling distribution of the average oxygen level in 30 tanks sampled uniformly at random from all 1000 tanks. (Not knowing the population, Mark couldn't compute this.)
  3(c):
- An approximation to the distribution in (c), computed by resampling Mark's 30 sampled tanks 10,000 times (that is, taking 10,000 samples of size 30 with replacement from Mark's sample of 30 tanks) and taking the average oxygen level of each resample.
  3(d):

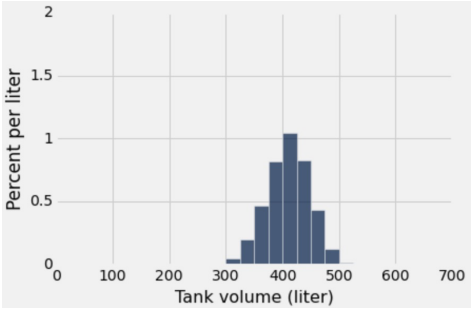
(i)



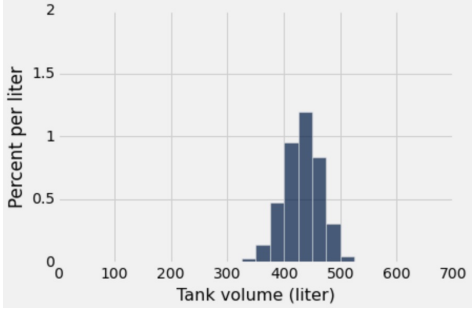
(iii)



(ii)



(iv)



Answer:

(a) iii.

(b) i.

(c) iv.

(d) ii.

Here's our reasoning (which you didn't need to provide). The Central Limit Theorem says that a sampling distribution of an average should be roughly bell-shaped, so (c) and (d) are likely (ii) and (iv) in some order. The sampling distribution of the average should be centered at the population's average, while the distribution of resample averages should be centered at the sample's average. The population average is a little bigger than the sample average, so (iv) must match (c) and (ii) must match (d). Finally, (i) is more jagged than (iii), so (i) is probably the sample (b) and (iii) the population (a).

This is a pretty standard situation when you're computing the average of something using a sample. (The sample average could have been higher or lower than the population average, though.) It may be helpful to keep these 4 pictures in mind when you're thinking about how confidence intervals are computed.