Activity 13

Population vs Sample

Question 1: After reading some of the documentation, what is the population of interest for the GSS?

The target population of the GSS is adults (18+) living in households in the United States.

Question 2: What undercoverage might exist?

From 1972 to 2004 GSS was further restricted to those able to do the survey in English. From 2006 to present it has included those able to do the survey in English or Spanish Those unable to do the survey in either English or Spanish are out-of-scope. Residents of institutions and group quarters are out-of-scope — for example: college students and military members.

Question 3: Can our survey dataset (survey you took at the beginning of the quarter) be considered a random sample of NU students? What are some differences between all NU students and our sample (i.e. the people in STAT 202)? How might these differences affect the results?

No! Students are not randomly assigned to 202 they choose to take it. Some differences include year in school (predominantly freshman), major/minor, etc…

Question 4: Compare the responses to 'how happy are you' between the GSS and NU students (our survey data).

Sampling Methods

Question 5: Using the 'movies' dataset simulate a simple random sample of size n = 300 to fill in the table.

	mean	standard deviation	count
Population	5.75	1.51	39123
Sample	5.93	1.47	300

Question 6: Simulate stratified sampling selecting 25 observations from each group.

Stratified sampling works best when a heterogeneous population is split into fairly homogeneous groups, meaning each group represents some characteristic. If the sizes in each group are not similar, then it is often more appropriate to sample x% from each group rather than a fixed number.

Are there any potential issues with our stratified sampling design?

We need to consider if each decade adequately represents the population or if each decade is homogeneous. Without more information this is slightly opinionated as long as you can defend your answer. I would say each decade is most likely homogeneous because movies released within each decade are probably similar to each other and have changed over the decades (meaning one decade cannot represent the population).

So the only potential issue is if the number of total movies in each subgroup is not the same it would be better to sample a set percent from each group rather than a count.

Question 7: Simulate cluster sampling with 2 clusters.

Each subgroup should be heterogeneous when using cluster sampling, meaning each group should be a mini representation of the entire population. If the groups are not similar then the results could be biased or inaccurate.

What decades did your simulation select?

1950 and 1980

Are there any potential issues with our cluster sampling design?

We need to consider if each decade adequately represents the population or if each decade is homogeneous. Without more information this is slightly opinionated as long as you can defend your answer. I would say each decade is most likely homogeneous because movies released within each decade are probably similar to each other and have changed over the decades (meaning one decade cannot represent the population).

This means cluster sampling is NOT appropriate because each subgroup does not adequately represent the population.

Feel free to explore the other datasets and sampling methods in the app!

Random Sampling vs. Random Assignment

Question 8: Classify whether each statement is generalizable and whether the conclusions are causal.

a) The Illinois state board of education wants to determine if a new standardized math curriculum will improve student test scores. Each teacher chooses whether they use the old or new curriculum and you randomly select 40 schools to evaluate.

Teacher chooses curriculum -> NOT random assignment = NO causal conclusion Randomly select 40 schools -> Random sample = generalizable claims

No causal conclusion if curriculum improves test scores we can only conclude correlation. However, we can generalize these correlation conclusions to the whole population.

b) Seventy percent of voluntary respondents to an on-line poll indicated that they want the Boston Red Sox to win the World Series.

No random sampling/selection (no generalizability); No random assignment (correlation/observational or no causal claim)

c) Unfortunately, the antidepressant drug caused an increase in the mean depression score for patients in our study.

No random sampling/selection (no generalizability); Random assignment (causation or causal claim)

I chose to assume we only provided medication to patients that reported being depressed and that the doctor conducted this as a clinical trial.

d) While reviewing their hospital's medical records, doctors observed that their patients who drank coffee daily had a higher rate of cancer than those who did not drink coffee daily.

No random sampling/selection (no generalizability); No random assignment (correlation/observational or no causal claim)

In this case the population is the hospitals medical records and that each doctor only observed their own personal patient's habits. Since he most likely did not tell patients to drink coffee or not too this would be observational.