5 - Sampling

You hear many studies:

Gallup Poll questions a sample of about 1500 adult US residents to determine national opinion on many issues.

Population: Adult US residents
Sample: 1500 adult US residents

According to a survey, 11 in 10 Americans know statistics.

Population: Americans
Sample: Not specified

In a random check of several hundred retail stores, the FDA found that 34% of the stores were not storing fish at the proper temperature.

Population: US retail stores
Sample: Several hundred retail stores.

A **population** in a statistical study is the entire group of individuals we want information about.

A **census** collects data from every individual in the population.

A **sample** is a subset of individuals in the population from which we actually collect data.

Big idea: Good sample lets us make inference about the general population.

Bad sampling design introduces **bias**: it consistently under- or over-estimates aspects of the general population.

Good or bad? ...

We conveniently survey students in STAT 2290 on their knowledge of JAVA and R and use this to infer how well all Rowan students understand programming and statistics.

BAD: Convenience sample chooses easy-to-reach individuals from a population.
This sample is biased since STAT 2290 students are all CS majors which is not reflective of general Rowan students.

Student responses on Man's end-of-semester teaching evaluations.

See evaluations.pdf.
Like product reviews, hence why you prioritize 2-4 star reviews over 1&5 star reviews.

BAD: In a voluntary response sample, people choose themselves by responding to a general invitation.
This sample is biased since the most vocal individuals respond.

Assign every Rowan student the same chance of being selected, and randomly select 100 Rowan students to financially audit. We use this to infer the finances of all Rowan students.

BEST: This is a simple random sample (SRS) where each individual has the same chance of being selected.

Assign every Rowan student the same chance of being selected, and randomly select 100 Rowan students to financially audit. We use this to infer the finances of all US students.

BAD: An SRS of Rowan students is not an SRS of all US students.

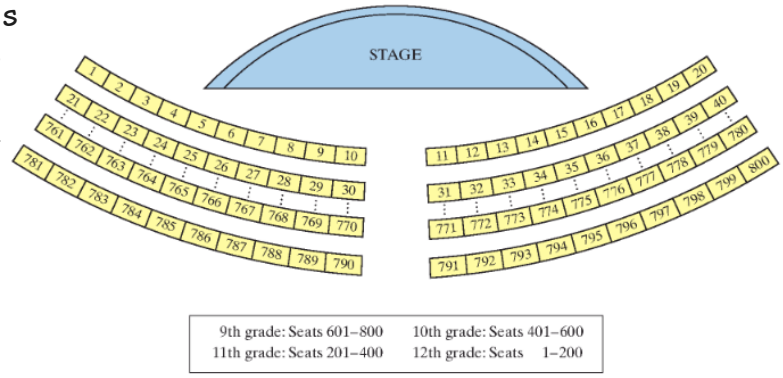
Simple random sampling (SRS) is the most unbiased sampling method, but sometimes impractical or too time-consuming to conduct.

Practical modifications of SRS:

1. Stratified random sample: group similar-trait individuals together into "strata". Do a separate SRS in each stratum and combine these SRS to form the sample. Ideally, each stratum looks homogeneous.

2. Cluster sample: group nearby populations together into "clusters". Do SRS on the clusters, to choose which cluster is completely surveyed. Ideally, each cluster looks like the population, but on a smaller scale.

Example. All 800 high school students are at an auditorium seated in seats 1-800. Describe how to survey 80 students on the topic "use of school library" via:
(a) Simple random sample;
(b) Stratified random sample;
(c) Cluster sample.



Answer:

(a) Use "sample(1:800, 80)" in R to generate 80 random numbers. Survey the selected students in those seat numbers.

(b) Student's library use is likely similar within grade levels but different across grade levels, so we use grade levels as each stratum. We do SRS of 20 seats from each grade: use "sample(1:200, 20)" to select 20 students from 12th grade, "sample(201:400, 20)" to select 20 students from 11th grade, and so on.

(c) Each column (e.g. 1, 21, 41, ..., 761, 781) forms a cluster of 40 seats, with 20 clusters overall. Randomly pick 2 clusters. Survey all individuals in these 2 clusters. Note that each cluster has a good mix of 9th, 10th, 11th, 12th graders.

Note: cluster sampling is more efficient than finding 80 seats scattered across the auditorium.

WHAT GOES WRONG

Undercoverage: some members of population are not reachable (mailed surveys neglect the homeless).
Nonresponse: individuals refuses participation.
Wording of questions: "Should we teach Arabic numerals in school?" vs "Should we teach the symbols 0, 1, 2,...,9 in school?"
Order of question: "How many dates did you had in the last month?", "Are you happy?"