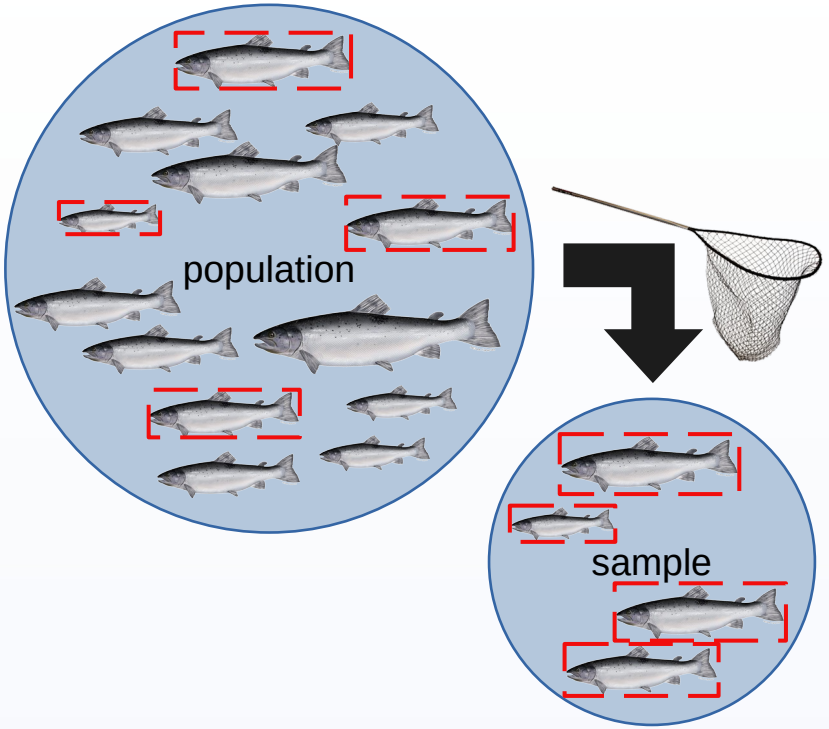
EGM101 – Skills Toolbox

Week 5, Part 3: Collecting Data

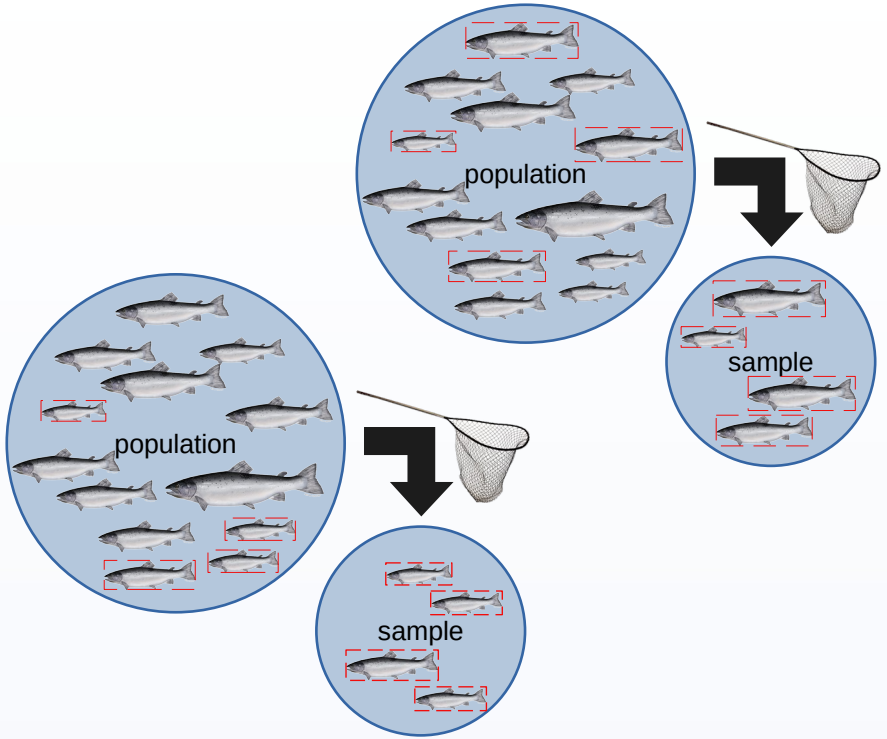
Need to Sample

- Studying an entire population is usually not feasible/possible:
  - Limited resources
  - Population may not be finite
- Use information about the sample to *infer* information about the population (inferential statistics)



Sampling Bias

- Sample should have similar characteristics as population (representative)
- Sample bias: some members of population are more likely to end up in sample
- Important to consider *how* we sample
- Random sample (ideal case):
  - All members of population have equal chance of selection

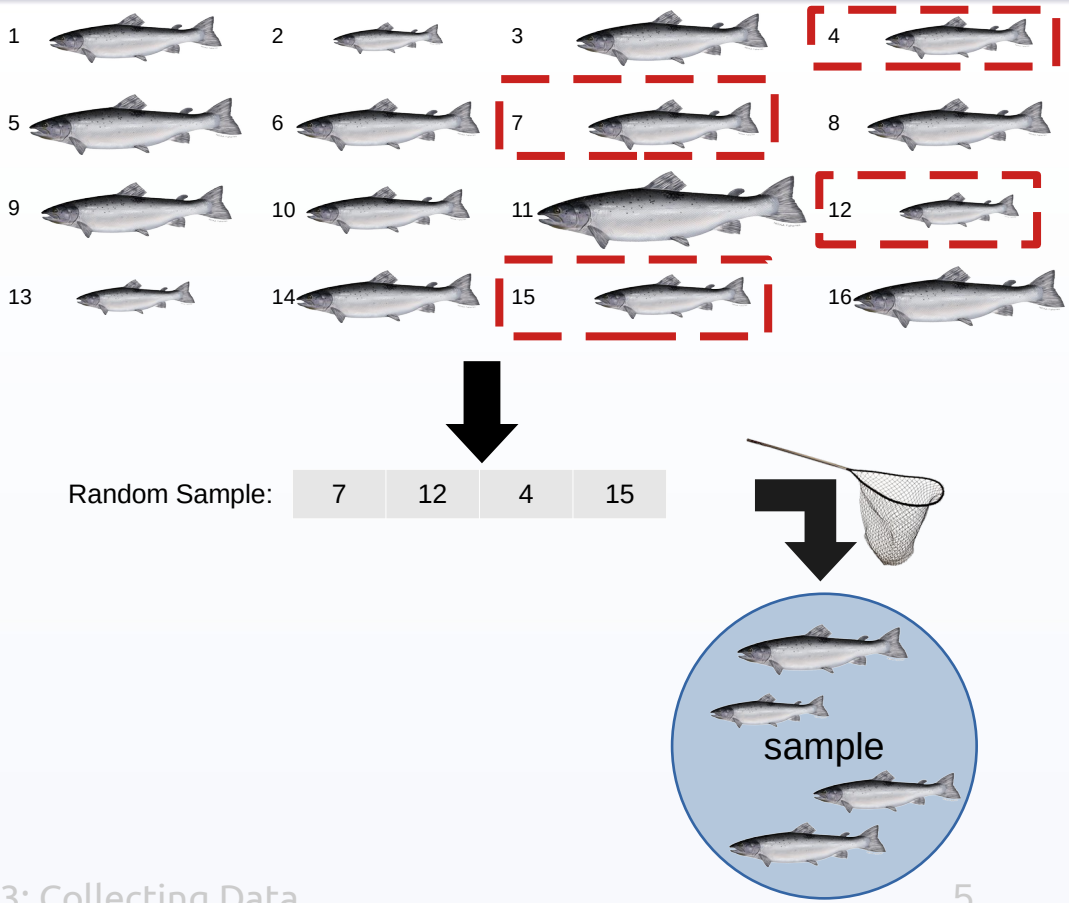


Sampling Error

- Remember:
  - Statistic ↔ sample
  - Parameter ↔ population
- Difference between statistic, parameter: sampling error
- In general: larger (random) samples decrease sampling error
- With random samples:
  - Reduce sampling bias
  - Can estimate the sampling error (more on this later...)

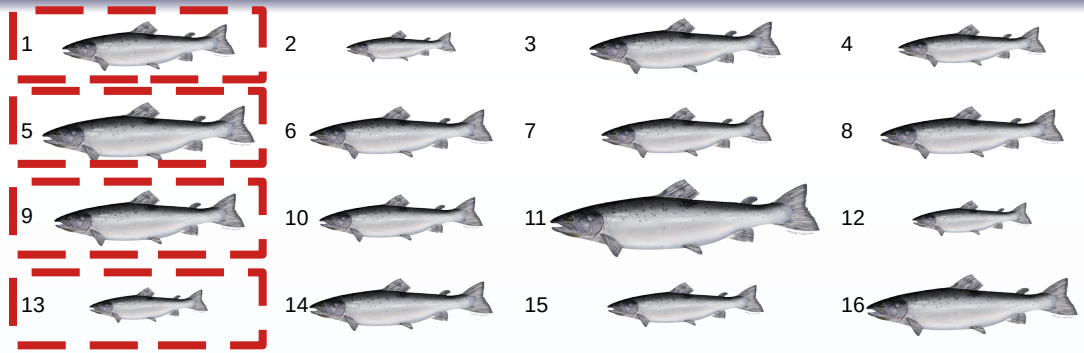
Simple Random Sampling

- "Lottery method"
- Steps:
  - Choose the target population, sample size
  - Assign each member of population a number
  - Select numbers at random
- Pros:
  - Minimizes selection bias
- Cons:
  - Can be time-consuming
  - Access to subjects/respondents
  - No guarantee that sample isn't biased

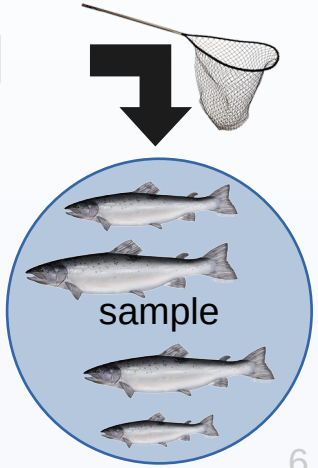


Systematic Random Sampling

- "Constant Skip" method
- Steps:
  - Choose the target population, sample size
  - Assign each member of population a number
  - Randomly select a starting point
  - Choose every n^{th} member after the starting point
- Pros:
  - Simple to do (depending on application)
  - Only have to choose one random number
- Cons:
  - Sampling frame could have periodicity (e.g., rooms in a building)

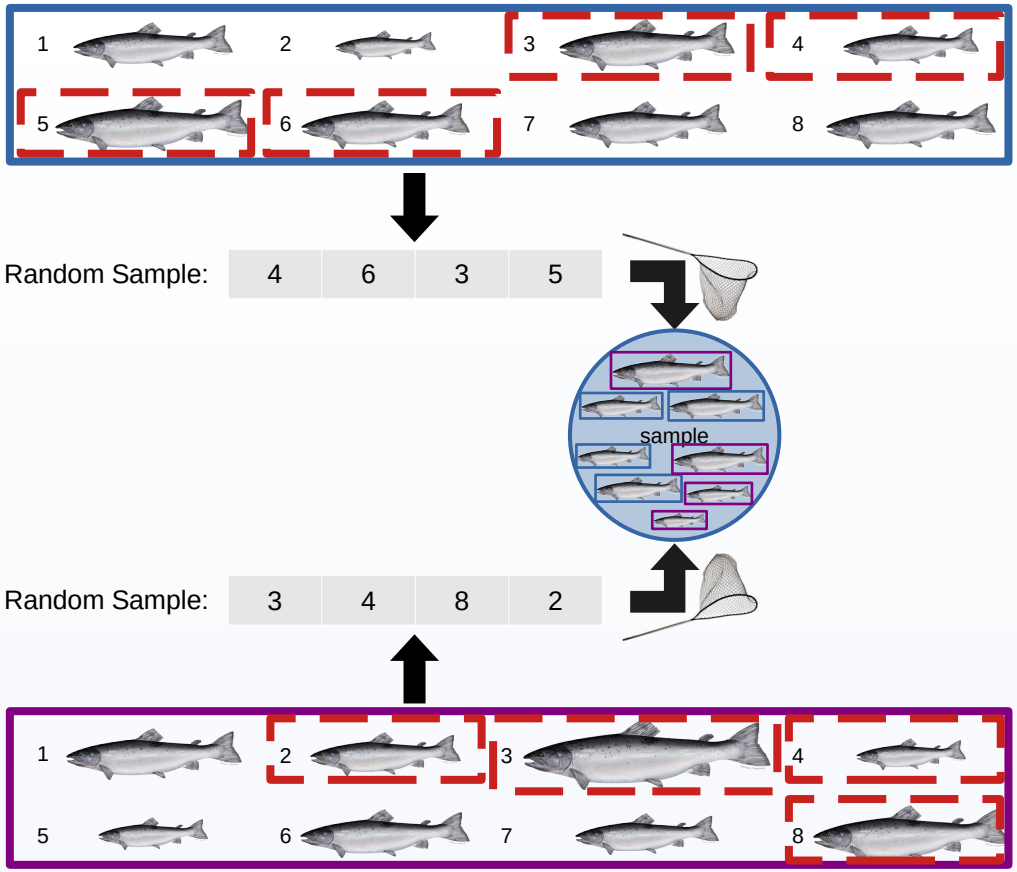


Random Sample: 1 5 9 13



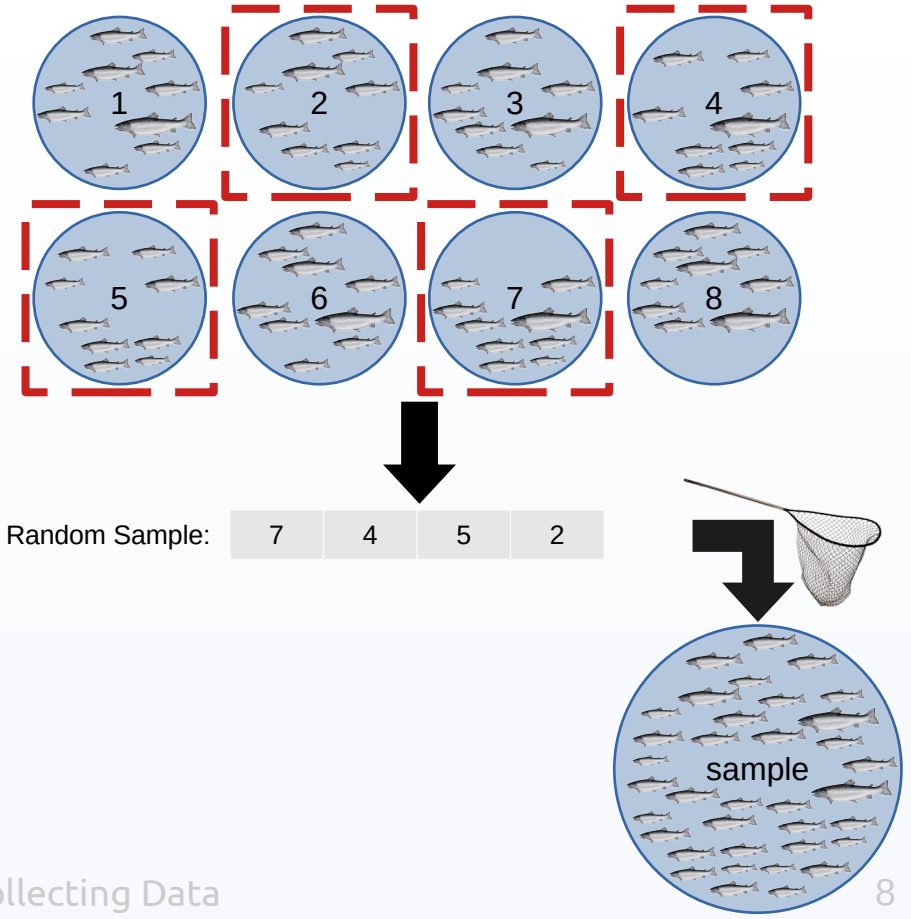
Stratified Random Sampling

- Steps:
  - Divide population into groups based on similar characteristics (strata)
  - **Important**: all members must belong to a single group (exhaustive, mutually exclusive)
  - Choose a random sample from each group
- Pros:
  - Works well when population has obvious strata
  - Gives representation to each strata (more representative sample)
- Cons:
  - Not all datasets can be stratified efficiently
  - Strata sizes are not always sufficient



Cluster Sampling

- Steps:
  - Divide population into smaller groups (clusters)
  - **Important**: clusters must be exhaustive, mutually exclusive
  - Choose a random sample of the clusters
- Pros:
  - Efficient (time, cost)
  - Don't need a list of the entire population (just the groups)
  - Useful for sampling based on location (geographic areas, streets/neighborhoods, etc.)
- Cons:
  - Sample may not be representative (bigger error)
  - No information about unsampled clusters → many small clusters vs few large clusters



Summary

- Studying entire populations is often infeasible or impossible → have to sample
- To effectively study population from sample, have to avoid sampling bias
- Most frequently, use some form of random sampling
  - Choice depends on study design, application

Additional Resources

- Illowsky and Dean, Chapter 1.2
- Caswell, Chapter 2
- Weiss, Chapter 1.2 – 1.4
- Fowler, Cohen and Jarvis, Chapter 2
- Bergstrom and West, “Selection Bias” (Chapter 6)
- Techniques for random sampling and avoiding bias [Khan Academy]