Learning Goals:

- Define statistical bias.
- Recognize biased sampling plans.
- Explain the purpose of random sampling in an observational study.

Introduction:

We now focus on how to collect reliable and accurate data for an observational study.

In an observational study, we analyze a sample data and draw a conclusion about a population. For example, we want to use an exit poll to predict who will win an election. The sample needs to be a subset of the population. In addition, the responses of the sample should be representative of the responses of the population.

A *sampling plan* describes exactly how we will choose the sample. A sampling plan is *biased* if it systematically favors certain outcomes.

Bias often occurs in situations where individuals self-select into the sample. For example, a poll conducted by a conservative call-in radio program may overestimate opposition to proposed gun legislation because only conservatives with strong opinions against gun control will take the time to participate.

Statisticians use random sampling in an attempt to eliminate bias. In random sampling all individuals in the population have an equal chance of being selected. In general, larger random samples produce more reliable results than smaller ones.

Random sampling also guarantees that the sample results do not change haphazardly from sample to sample. When we use random selection, the variability we see in sample results is due to chance and the results obey the mathematical laws of probability. This is important in statistical inference.

Check your understanding:

1) Suppose that we want to estimate the mean number of text messages sent by LMC students each day. Which sampling plan will produce the most reliable estimate? Why do you think so?

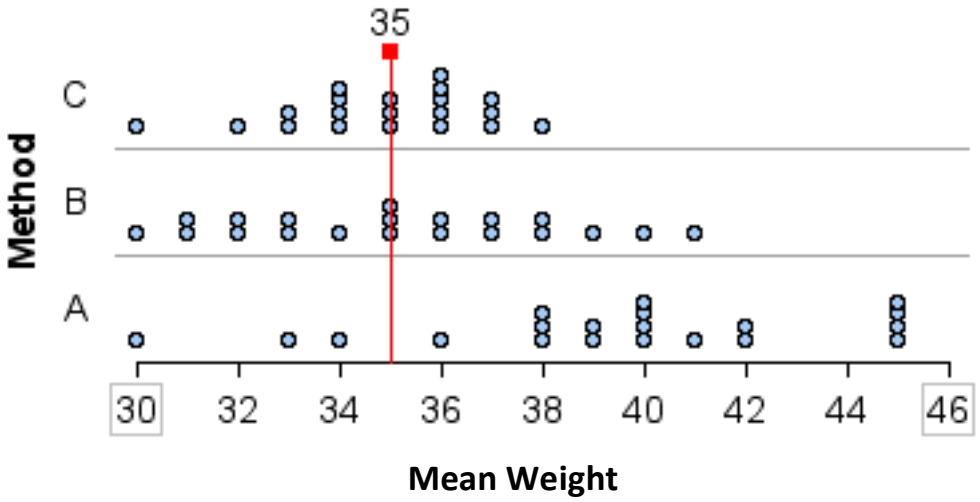
    a) Select 50 students at random from the list of students' Insite email addresses. (All LMC students have Insite email addresses.)
    b) Select 100 students at random from the list of students' Insite email addresses.
    c) Select the first 200 students who you see texting in the LMC Quad.
    d) Select 300 students at random who follow LMC on twitter.

Group work:

2) A 5th grade class conducts a study to estimate the mean weight of a snail. They have a box containing 100 snails; each snail has a number on its shell. The 20 children do the study using three different sampling methods:

Method A: Each child picks 5 average looking snails, weighs each one, then averages the 5 weights.
Method B: Each child draws 5 numbers out of a hat, locates the 5 snails labeled with those numbers, weighs each one, then averages the 5 weights.
Method C: Each child draws 10 numbers out of a hat, locates the 10 snails labeled with those numbers, weighs each one, then averages the 10 weights.

Here are the results:



a) The mean weight of the 100 snails is 35 grams. Which of the sampling method is producing the most biased estimates of the overall mean weight? How can you tell by looking at the graphs?

b) How does the graph produced by Method B and C illustrate that random samples produce unbiased weight estimates?

c) How does the graph produced by Method C illustrate that larger random samples produce more accurate estimates than smaller random samples?