Mr Murphy **HW Pg. 42 #2.15, 2.18, 2.29, 2.33, 2.34**
AP Statistics
3.1 The Data Analysis Process and Sampling

Goals: 1. Determine types of bias.
2. Determine methods of sampling.

Why do we take samples of populations to conduct studies?
Would it be wise to test the breaking point of all coke bottles? to run a census on cars to test their safety ratings?

How do we put a sample together? Will we always get a sample that perfectly represents our population? Much can go wrong when selecting your sample. Statisticians have a word for that - **bias**.

Bias - A sample is **biased** if it systematically over-represents or under-represents a segment of our population of interest.

There is no way to recover from a biased sample or a survey that asks biased questions.

Types of Bias

- **Selection Bias**
  Tendency for samples to differ from the corresponding population as a result of systematic exclusion of some part of the population.

  Selection bias occurs when a part of the population is systematically excluded.
  Consider telephone surveys? volunteers for a study?

- **Undercoverage Bias**
  A sampling scheme that fails to sample from some part of the population or that gives part of the population less representation than it has in the population.

  An example of **undercoverage** is the Literary Digest voter survey, which predicted that Alfred Landon would beat Franklin Roosevelt in the 1936 presidential election. The survey sample suffered from **undercoverage** of low-income voters, who tended to be Democrats.

- **Response Bias from measurement**
  Response bias refers to the bias that results from problems in the measurement process. Social desirability and leading questions are common examples of response bias.

  Consider the following question -

  “It is estimated that disposable diapers account for less than 2 percent of the trash in today’s landfills. In contrast, beverage containers, third-class mail, and yard waste are estimated to account for about 21 percent of trash in landfills. Given this, in your opinion, would it be fair to tax or ban disposable diapers?”

- **Nonresponse Bias**
  Nonresponse Bias occurs in a sample design when individuals selected from the sample fail to respond, cannot be contacted, or decline to participate.

  Recall phone interviews? Consider mail surveys?

It is important to note that bias is introduced by the way in which a samples is selected or by the way in which the data are collected from the sample.

Increasing the size of the sample, although possibly desirable for other reasons, DOES NOTHING TO REDUCE BIAS!

Ex1 According to the article “Effect of Preparation Methods on Total Fat Content, Moisture Content, and Sensory Characteristics of Breaded Chicken Nuggets and Beef Steak Fingers”, sensory tests were conducted using 40 college volunteers at Texas Women’s university. Give three reasons, apart from the relatively small sample size (although really it’s not that small), why this sample may not be ideal as the basis for generalizing to the population of all college students.

Methods of Sampling

- **Simple Random Sample (SRS)**

  A **simple random sample of size n** is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.

The definition of a simple random sample implies that every individual member of the population has an equal chance of being selected AND every group of size n is possible.

Ex2 A small private college has 4500 students enrolled. Assume that the university can provide a list of the students with the students numbered from 1 to 4500. Describe the procedure you will use to select a simple random sample of 20 students, and then identify (by number) which students from the list are included in your sample.

We have 3 methods to find an SRS:
1. Slips of paper
2. Calculator
3. Random Digit Table

Start with line 27 from the random number table 1 on p. 860 of your book.

Other Methods of Sampling

- **Stratified Sampling - usually easier and less costly than an SRS, provides additional information as well**
  When the entire population can be divided into a set of non-overlapping, homogeneous subgroups, a method known as **stratified sampling** often proves easier to implement and more cost-effective than an SRS. In stratified sampling, separate random samples are independently selected from each subgroup (called **strata**).

  If you wanted to poll the entire student body at SI you could break us up into Freshmen, Sophomores, Juniors, and Seniors…then take an SRS inside each of the **strata**.

- **Cluster Sampling**
  Cluster sampling involves dividing the population of interest into non-overlapping heterogeneous subgroups, called clusters. Clusters are then selected at random, and all individuals in the selected clusters are included in the sample.

  Consider a 10-story college dorm building. We could get a list of each student living in the dorm, number them, get our random sample and track each kid down for an interview. It would be easier to pick a random floor in the dorm building and interview each student on that entire floor.

- **Systematic Sampling**
  Systematic sampling is a procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement.

  Let’s say you want to TP your teachers’ houses but you want to randomly choose which teachers to TP so that no teacher will take the attack personally. There are 120 faculty members at SI. You want to nab ten teachers in one awesome night. Break the teachers into 12 groups of ten, randomly select a number between 1 and 10…let’s say you randomly select 7, then from each group of 12 TP the 7th faculty member on the list. :)

- **Convenience Sampling**
  Convenience sampling relies on using an easily available or convenient group to form a sample.

  If you want to know how many hours the average students at SI studies, will asking just your friends give data that represents the entire student body?

Ex3 If we took the 500 people attending a school in New York City, divided them by gender, and then took a random sample of the males and a random sampling of the females. This is an example of what type of sampling method?

(a) Simple Random Sample
(b) Stratified Random Sample
(c) Systematic Random Sample
(d) Cluster Sample
(e) Convenience Sample

Ex4 Say the target population in a study was church members in the United States. There is no list of all church members in the country. The researcher could, however, create a list of churches in the United States, choose a sample of churches, and then obtain lists of members from those churches to interview.
This is an example of what type of sampling method?

(a) Simple Random Sample
(b) Stratified Random Sample
(c) Systematic Random Sample
(d) Cluster Sample
(e) Convenience Sample

Ex5 Determining the sample interval (represented by *k*), randomly selecting a number between 1 and *k*, and including each *k*th element in your sample are the steps for which form of sampling?

(a) Simple Random Sample
(b) Stratified Random Sample
(c) Systematic Random Sample
(d) Cluster Sample
(e) Convenience Sample

Ex6 Each of the 29 NBA teams has 12 players. A sample of 58 players is to be chosen as follows. Each team will be asked to place 12 cards with their players' names into a hat and randomly draw out two names. The two names from each team will be combined to make up the sample. Will this method result in a simple random sample of the 348 basketball players?

(a) Yes, because each player has the same chance of being selected.
(b) Yes, because each team is equally represented.
(c) Yes, because this is an example of stratified sampling, which is a special case of simple random sampling.
(d) No, because the teams are not chosen randomly.
(e) No, because not each group of 58 players has the same chance of being selected (i.e. not every group is possible).

Ex7 To survey the opinions of bleacher fans at Wrigley Field, a surveyor plans to select every one-hundredth fan entering the bleachers one afternoon. Will this result in a simple random sample of Cub fans who sit in the bleachers?

(a) Yes, because each bleacher fan has the same chance of being selected.
(b) Yes, but only if there is a single entrance to the bleachers.
(c) Yes, because the 99 out of 100 bleacher fans who are not selected will form a control group.
(d) Yes, because this is and example of systematic sampling, which is a special case of simple random sampling.
(e) No, because not every sample of the intended size has an equal chance of being selected.

Ex8 (Not a multiple choice problem) The financial aid officers of a university wish to estimate the average amount of money that students spend on textbooks each term. They are considering taking a stratified sample. For each of the following proposed stratification schemes, discuss whether you think it would be worthwhile to stratify the university students in this manner.

(a) Strata corresponding to class standing (freshman, sophomore, junior, senior, graduate student)
(b) Strata corresponding to field of study, using the following categories : engineering, architecture, business, other
(c) Strata corresponding to the first letter of the last name : A –E, F – K, etc.

Checkpoint:
Multiple Choice Questions

1. Which of the following are true statements?

I. If bias is present in a sampling procedure, it can be overcome by dramatically increasing the sample size.
II. There is no such thing as a bad sample.
III. Cluster sampling is an SRS.

(a) I only (b) II only (c) III only (d) I and II only (e) None of the above

2. A researcher planning a survey of heads of households in a particular state has census lists for each of the 23 counties in that state. The procedure will be to obtain a random sample of heads of households from each of the counties rather than grouping all the census lists together and obtaining a sample from the entire group. Which of the following is a true statement about the resulting stratified sample?

I. It is not a SRS.
II. It is easier and less costly to obtain than a SRS.
III. It gives comparative information that a SRS wouldn’t give.

(a) I only
(b) I and II
(c) I and III
(d) I, II and III
(e) None of the above gives the complete set of true responses.