MTH 245 Statistics 1 DE: 2022 summer assignment

(due Friday 9/16/22 in Canvas, counts as an assignment grade)

Welcome to my statistics course! I look forward to getting to know each of you in the fall. Please take some time this summer to show me what you know, and a little bit about you as a reader and a writer. I am particularly interested in seeing work that is clear, concise and well-formatted. -- Ms Maskelony (geraldine.maskelony@apsva.us)

Part 1: Vocabulary and Introduction to Statistics

Statistics is the science of learning from data and using that data to make informed decisions. For example, suppose we want to know the mean height of high school students in the United States. Would it be feasible to conduct a census of these students? NO! That would be very costly and time-consuming. Instead, we take small samples from the large population of high school students and use these to infer (estimate as best we can) the true mean height of all high school students.

Important definitions:

- Population: the entire group about which we want to make an estimate
- Sample: a subset of the population from which we actually collect the data.
- Census: data collected from every individual in the population.

State whether the data set is a population or a sample. Explain your answer.

dataset	Population or sample? Explain.
the age of every fourth person entering a department store	
the age of each employee at a local grocery store	

Identify the population AND the sample.

- A survey of 1212 American households found that 52% of the households own a computer.
    - population:
    - sample:

- When 1348 American households were surveyed, it was found that 57% of them owned two cars.
    - population:
    - sample:

- A survey of 2625 elementary school children found that 28% of the children could be classified as obese.
    - population:
    - sample:

What is the best way to collect data? We'll go through the following example to answer that question.

Example: A farmer has a plot of land in which she plants and harvests crops. This plot is divided into 16 equal regions as shown below:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

The farmer would like to know which area of her field yields the most crops. In this example, it MIGHT be feasible for her to take a census of her entire field, counting the total number of crops in each of the 16 sections. This would certainly allow her to determine which sections are best for planting her crops. However, perhaps the farmer does not have time to conduct a census and must instead sample only four plots. How does she choose which four to sample?

Method #1: Simple Random Sample (SRS).

The farmer could simply take a simple random sample (SRS) and randomly choose four of the 16 plots. To do this, we will generate four random numbers. In the course we will explore other ways to generate random numbers, but for the summer assignment just use your graphing calculator to choose four random plots by using: math > PROB > 5: randInt(Lower = 1,Upper = 16 n = 4 [randint(1, 16, 4)] Press "enter" as many times as it takes to get four unique numbers. List your four numbers below:

Plots selected with simple random sampling (SRS).	

Method #2: Stratified Random Sample

One disadvantage to choosing randomly is that the farmer may not get a truly representative sample of her plot. For example, perhaps one corner of her plot floods every time it rains and does not produce a lot of crops. If, by chance, she randomly selected this corner and used that data to describe her crop growth, it would not be representative of the entire plot. Her data would not be useful. Instead, she could use a Stratified Random Sample. To do this, the farmer would first classify the population into groups called strata. Then, she would choose a SRS from each stratum and combine these individual SRSs to form the sample. In this example, she could stratify by row or column. You are going to simulate a stratified random sample in your graphing calculator by row. To do this, follow the same steps as above, except instead of having your lower and upper bounds be 1 and 16, they should represent the row or column you are looking at. So, for example, your first randomly selected number should be between 1 and 4, the second should be between 5 and 8, and so on. Also note that n = 1 (n represents the sample size, in this case the number of plots you are selecting).

Plots selected with stratified random sampling.	

Unbeknownst to the farmer, here are the actual numbers for the yield of crops for each section of her plot:

4	7	6	5
29	31	27	32
94	98	92	97
150	153	148	147

Using this chart, answer the following questions:

Find the average of all 16 of these numbers.	
Find the average of the four sections that were chosen from your simple random sample above.	
Find the average of the four sections that were chosen from your stratified random sample above	
Which sample average was closest to the true average?	

There is one more type of sampling we will conduct in this course, called **cluster sampling**. Students often confuse cluster sampling with stratified random sampling, so I will try to address that here:

Stratified Random Sample

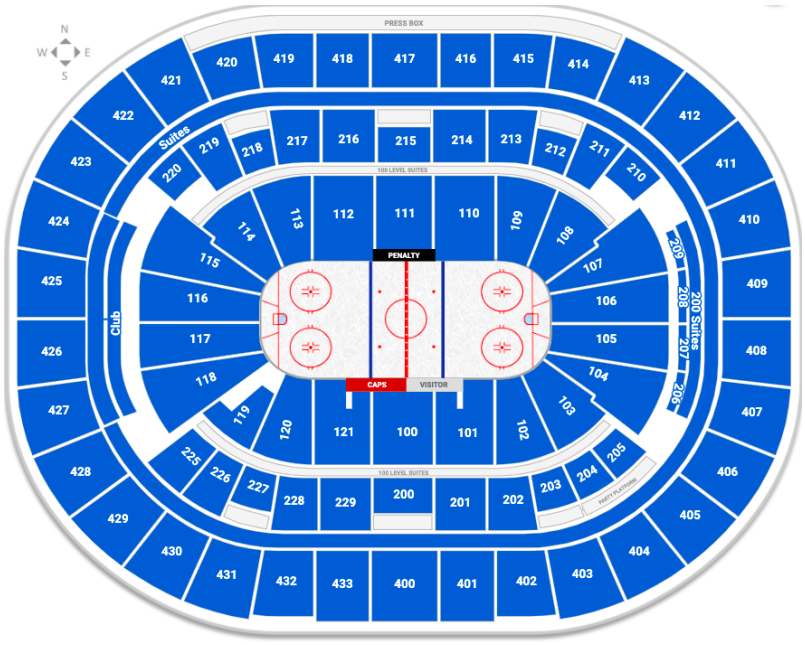
- Population is classified into groups, called strata.
- Strata are usually chosen so that they are DIFFERENT from one another.
    - For example, a high school may be stratified by GRADE LEVEL.
- A simple random sample (SRS) is taken from EACH strata. All of these "mini-samples" are combined to create our final sample.

Cluster Sample

- Population is classified into groups, called clusters.
- Clusters are usually chosen to be the SAME as each other.
    - For example, a high school may be clustered by first letter of last name
    - each cluster would include all types of students from each grade level
- Choose a simple random sample of clusters (NOT WITHIN EACH CLUSTER).
- All individuals within each cluster are included in the sample.

Exercise

Shown below is a seating chart of the Capital One Arena in DC, home of the Washington Capitals hockey team. The owner would like to survey fans at a Caps game about their experience. It would obviously be impossible to talk to all 18,573 fans during a game (so a census is impossible).



Describe how the owner, Ted Leonsis, could collect a sample of fans using each method. Also, describe an advantage and disadvantage for each method.

Simple Random Sample	
Describe how a sample might be collected using this method.	
Possible advantages:	
Possible disadvantages:	
Stratified Random Sample	
Describe how a sample might be collected using this method.	
Possible advantages:	
Possible disadvantages:	
Cluster Sample	
Describe how a sample might be collected using this method.	
Possible advantages:	
Possible disadvantages:	

Sample Surveys: What Can Go Wrong?

Not all samples produce reliable data. There are a few things that can go wrong:

Undercoverage: Undercoverage occurs when some members of the population cannot be chosen in a sample. For example, surveys delivered via email results in undercoverage for people who do not have an email address. What is another example of undercoverage bias?

Nonresponse Bias: Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate. For example, surveys done via telephone will result in nonresponse if the person does not pick up their phone or hangs up because they are not interested in the survey. What is another example of nonresponse bias?

Wording Bias: Wording bias occurs when the wording of a question intentionally attempts to sway the results of the study in one direction. What is an example of wording bias?

There are two types of sampling that often lead to substantial bias:

- Convenience sample: a convenience sample consists of choosing individuals from the population who are easy to reach. For example, if I am at a math department meeting and conduct a survey about everybody's favorite subject, my results will likely be biased.
- Voluntary sample: a voluntary response sample consists of people who choose themselves by responding to a general invitation. For example, if a survey was sent out asking about a proposed new tax, perhaps only the people who feel most strongly about it will respond.

Part 2: Basic Data Skills

You are going to find the mean, median, and standard deviation of a set of data. The data set is shown below:

12 10 08 10 11 07 15 08 09 15 11 05 13 13 13 08 06 11 12 10

Show two different ways to find this information. You may use any technology or pencil you like. Paste a photo or a screenshot of your work below with the answer clearly marked. If you do not know or remember how to find mean, median and standard deviation, please consult with Professor Google or Professor YouTube.

Part 3: Basic Probability

A die consists of six sides (numbered 1, 2, 3, 4, 5, and 6). When a fair die is rolled, all six sides have an equal chance of being rolled. Find each probability (expressed as a fraction):

The probability of rolling a 2	
The probability of rolling an odd number	
The probability of rolling a 1.5	
The probability of rolling a 1 or a 3	
The probability of rolling a 2 GIVEN that an even number is rolled	
The probability of rolling an even number GIVEN that a 2 is rolled	

Part 4: Basic Algebra

Solve for "n" in each problem. You may use a calculator. Share your written work as a photo.

$$0.05 = 2.145\left(\frac{1.362}{\sqrt{n}}\right)$$

$$0.025 = 1.96\sqrt{\frac{(0.3)(0.7)}{n}}$$

Part 5: Ambiguity

Here is an example of the type of question that will be answered throughout Statistics I. Do your best to answer each question. Unlike questions asked in other math courses, there is no “right” or “wrong” answer here -- for now, use your intuition and answer each question thoroughly but concisely. (This part shouldn’t take long to complete, but put some thought into each answer).

Dr. Thomas wants to know what proportion of her geometry students would call her their favorite teacher. To find out, she creates a survey. The survey consists of one question: “I am about to give you a free day in class with no homework and unlimited Skittles. Who is your favorite teacher?” After collecting the results, Dr. Thomas is astonished and flattered to see that 85% of her students call her their favorite teacher. Can Dr. Thomas trust this data? Why or why not? If not, what could she have done to get better data?

Adapted from this source.