

Introduction to Statistics

Welcome to Statistics. For those of you who thought that you were signing up for a math class, I feel that I should make you aware that I really do not consider statistics a math class. I consider statistics a class that uses math much like some of the science classes utilize math. In statistics, like science, we test hypotheses through the design and completion of observational studies and experiments.

Okay so you are asking yourself, if statistics is not really a math course what is it and why is it important?

First, I will address the question: **Why is statistics important?** It is important because I said it is. I am guessing that my answer didn't satisfy many or any of you. Okay, so how about this: How do you think Amazon grew to be so huge and powerful when it started out as online book store? The answer their use of statistics. How do you think that Japan became a major manufacturing power known for high quality automobiles when their country and industry was virtually destroyed by a world war? The answer is statistics. How do you think a baseball team with the lowest salary base became a divisional champion? The answer is statistics. How did a group of MIT nerds get banned from Las Vegas? The answer is statistics. How was an effective HIV screening test created from one that was completely unreliable? The answer is statistics. How are advertisements selected for your computer screen or smartphone? The answer is statistics. How did Target know that a teenage girl that they had never seen was pregnant before her father did? The answer is statistics.

Let's explore the story about Target and the teenage pregnancy in a little more detail.

The background: To make better use of their advertising budget, Target began utilizing statistics to better identify the purchasing patterns of their customers so that they could send advertisements that might be of greater interest to each customer. In other words, based on what a person purchased, Target was sending advertisements and coupons to them for things that Target was predicting that the customer would want to purchase.

To that end a teenage girl was receiving advertisements for baby items. As a result an angry father walked into a Target outside of Minneapolis demanding to talk to the manager: "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August.

The Complete article can be found: <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/>

A closely related article and interview of the Target advertising Mastermind:

<https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=6&r=1&hp>

What is Statistics?

So now that you are a little curious, let's answer the question: **What is Statistics?**

Statistics is the study of how to collect, organize, display, describe, and analyze the data from a sample in order to make generalizations about a population.

I am certain that some or many of you are thinking, great another definition that my teacher believes explains everything, but in reality it leaves more questions than answers. To which I must respond, fair enough, but give me a few more minutes and I believe that I will be able to break down the definition and give you an overview of the course at the same time.

Statistics is divided into 4 major sections:

- Collecting Data,
- Displaying and Describing the Collected data
- Analyzing the Collected Data (Probability)
- Making inferences/decision based on the data collected.

The First Quarter will be devoted to Collecting Data and Displaying and Describing data. These two sections are the easiest sections of the course and it is imperative that you achieve really good grades in both of these sections for two major reasons. First, the collection of data is foundational to all other sections of the course and secondly the 2nd quarter which is comprised almost entirely of data analysis, **(the probability section)** is the most challenging for students. Yes, most students say probability is the hardest and their second quarter grades tend to be lower than that of their first quarter.

The final section of the course, making inferences, is where you actually get to make decisions based on data. Should we purchase the new machine, is the new drug more effective than the old one, can blank predict blank and with how much confidence.

I love the section on inferences and I feel it is empowering, because it is from this section that I can use all of the collected data to make an informed decision and not just a guess. Now I want to be perfectly clear about the word guess. **Nothing in statistics is certain**, however, with statistics I am able to determine the chance that I could be wrong. Typically I will choose to be correct 95% of the time. Which means that I accept the probability that I will be wrong about 5% of the time—but I am getting ahead of myself as these discussions will not occur in detail until the 3rd quarter.

As we begin this course, I should warn you that there are a significant number of definitions that must be learned in order to understand what we are trying to accomplish. So how important are the definitions? Very important—put it this way, there will be some tests where you will not need to perform any mathematical calculations. This isn't to say that those tests are easy, it is just a recognition that statistics is different from all of the math classes you have taken thus far.

Incidentally, the first few test are challenging for most students, not because the topics are difficult, but because you are being asked to think and reason in a different manner. In previous math classes, you followed an explicit set of directions to find a solution to a problem. In statistics, you will be expected to not only know and apply the definitions and methods, but be able to relate them to the outside world. In fact, the more you know about the world and other subjects, the more connections you will be able to make and the more successful you will be in this course.

Basic Vocabulary

Statistics is the study of how to collect, organize, display, describe, and analyze the data from a sample in order to make generalizations about a population.

To fully understand what this means, we need to understand the concept and importance of both a **population** and a **sample**. (we will briefly define and discuss sample shortly and in more detail next week)

Population: the entire group of subjects or individuals that is the subject of interest.

Parameter-a numerical measurement of a population (*rarely known*).

Census-the collection of data from each unit in the population.

(Difficult if not impossible with a large population)

In statistics, a population is the entire group of subjects or individuals that is the subject of interest while a sample is a subset of the population. For example: we may be interested in the percentage of students in your statistics class that have brown eyes. In that case, our population would be the students in your statistics class and it would be fairly easy to count and figure out the percentage of brown eyed students in the class. The percentage of brown-eyed students would be the parameter. Because we are checking the eye-color of the entire population, every student in the classroom, we are conducting a census.

The great thing about being able to conduct a census, is that we know the true population parameter. In this case, we would know the true proportion or percentage of students with brown eyes in the classroom. According to some of my students, the best part of being able to conduct a census knowing the true population parameter is that there is no need for statistical methods and no need for this class.

So why not perform a census all of the time? To that answer that question let's consider the percentage of students with brown-eye color scenario. What if my population of interest is not the students in my classroom but all of the students at Reagan? While counting that many students might be a challenge, and I might creep a few people out when I asked them for their eye color, it is probably still doable. However what if we wanted to know the percentage of people in San Antonio or the world who had brown eyes? Well then our population would be all of San Antonio or the world and it would be unrealistic to conduct a census count all of the brown eyed people in San Antonio or the world.

Why don't we always collect data using a census?

1. Sometimes it is impossible to conduct a census.
 - I cannot find the average size of bass in a particular lake without draining the lake at which point the bass is dead and there is no lake.
 - I cannot count the number of red blood cells in your body, without removing all of the blood from you and sucking your life away.
2. A census can be extremely expensive
The 2010 U.S census cost **\$13 billion, approximately \$42** person—your tax dollars at work.
3. Due to the difficulty in execution, a census can very inaccurate
 - In the 2000 census, it is estimated that 6.4 million Americans were missed, however, there was an overcount of 36,000. Now how did that happen?
 - In 2010, it is estimated that in Hildalgo county in Texas 225,000 persons were not counted.

Basic Vocabulary

So while an accurate census is great, it is not always possible and may be cost prohibitive. In the event that it is not plausible to conduct a census, we are forced to settle for an estimate. So the questions is how do we make an accurate estimate

Samples: are a subset of the population. In non-math terms, a sample is just a smaller group of the entire population.

For instance: If the population of interest is the students in a high school a sample might be the group of students in a particular classroom or at a lunch table. Any subset (small group) of the population is a sample. It is possible to have samples of size 1 or massive samples of thousands. Any subset or small group of the population is a sample. However, **not all samples are created equally**. Some samples are good and some are not.

Good Samples: To be a good sample, the sample needs to look like the population. I like to say that a sample is a microcosm of the population. In other words, a good sample should look just like the population but smaller. It is a mirror image or photograph of the population. It looks the same only smaller.

So, what is a bad sample? A bad sample is a sample that does not proportionally represent the elements of the population. A sample that does not proportionally represent all of the elements of a population is said to be biased.

Biased Sample: A sample that over or under-represents a part of the population. We will spend a great deal amount of time discussing bias in detail in about a week.

Randomness: One of the key ingredients to collecting a good sample is randomness. In other words we need to randomly select the sample from the population because we when we design studies we are unable to control for all of the different variables, but because we all have hidden biases we cannot use our best judgement. Yup, I just said you were biased. Don't believe me let's look at our opening activity.

So when you entered class today, I had you do something very simple, and had you select one number from a display like this: Please do so no if you have not already

1 2 3 4

Randomness Continued

1 2 3 4

How many of you chose: 1. _____ 2. _____ 3. _____ 4. _____

According to psychologists, typically, about 75% of the population chooses 3, 20% are divided between 2 and 4 and 5% choose 1.

Because there are 4 numbers, each number should be chosen 25% of the time. So why didn't that happen. I don't know why and several theories have been published. However, regardless of the reason, this shows that in general we make biased selections and in statistics biased samples result in samples that are not representative of the population and are therefore worthless.

However, if I had rolled a 4 sided die a million times about each number would have been selected about 25% of the time.

Believe it or not this actually has very practical applications to you especially when you return to campus. You don't think so?

Consider this: You go to the bathroom and there are three stalls. Assuming the bathroom is empty, which stall do you normally choose? For the vast majority of people, the answer is the second stall or middle stall which makes it the grossest and least clean stall of all. The next most chosen stall is the third stall which is the furthest one from the door. Assuming we are concerned about germs, hygiene etc. we should be choosing the first stall. Who knew that stats class would be so helpful with toilet training?

In the scenarios above we exhibited a bias. We over-represented the number 3 and the 2nd stall and we significantly under-represented the number 1 and the 1st stall. To have samples that are useful, we must eliminate bias through the design and execution of our studies and experiments.

Summary: The key point is that we must design studies in a manner that selects samples that are representative of the population. We must not over or under-represent segments of the population and to account for variables beyond the control of the study we must employ randomness.

Notes: Simulations

As mentioned earlier an accurate census is great, however, it is not always possible and may be cost prohibitive. In the event that it is not plausible to conduct a census, we are forced to settle for an estimate. So the question is, **how do we make an accurate estimate**. Essentially we have two choices, we can take a sample or we can run a simulation. We will discuss in great detail how to utilize sample data, but we are going to begin with learning how to conduct a simulation.

Simulation: a way to model random events, such that simulated outcomes closely match real-world outcomes. By observing simulated outcomes, researchers gain insight on the real world.

- Chance/randomness must be employed
- Often Used because it is more economical than running a true experiment
- Typically a probability model can be used to generate the same information
- We Want things to be fair and without bias or prejudice in our method of selection

Random: An event in which we know what possible outcomes can occur but do not which outcome actually will take place.

Component: The most basic **situation** in which something happens at random—(Hint: the singular event that you are repeating)

Outcome: the result of a single component

Trial: The number of components necessary to occur to model a situation. A single run of sequence of events being simulated

Response Variable: variable that measures the outcome of each trial; **outputs (Y's)**

Question 1: How many heads would I get if I flipped a fair coin 8 times? $(\frac{1}{2}) \times 8 = 4$

Are you certain? **Maybe** Will that happen every time? **Not very likely** Because of uncertainty we need to conduct studies which we will do this time as a simulation.

While there are several steps to running a simulation let's begin with just part of the process to answer the above question

Assign 2 digit numbers in proportion to the chance of success and failure.

- 00 = 0 or 100; 01=1; 02 = 2; 03 = 3...09 = 9
- 10 = 10; 11 = 11... 99 = 99
- Assign the numbers to be skipped or ignored
- We will flip the coin until we have 4 heads.
- We will repeat the simulation 3 times.

I will assign 2 digit numbers such that 00, 01... through 49 equals heads.

50, 51 to 99 equals tails. There are no skips and repeats are allowed

52822 48990 ~~03648~~ 34861 ~~54680~~ 64791

~~31645~~ 45552 ~~78255~~ 64794 21228 69707

~~38804~~ 45687 ~~85320~~ 54654 76156 01853

Trial #	# Coin flips
Trial 1	13
Trial 2	11
Trial 3	8
Total	32
Average	10.333

Based on our simulation we would expect to have to flip the coin 10.33 times to get 4 heads

Recipe for Success: Simulations

1. Read the entire problem

What is being asked?

2. Identify the Question

Explain what the question is asking in your own words

3. Identify and define a success and component in context

- **Success:** What we want to happen
- **Component:** What is being repeated

4. Identify a Trial

How many successes are required?

5. Model the simulation

Assign 1 or 2 digit numbers in proportion to the chance of success and failure.

- 00 = 0 or 100; 01=1; 02 = 2; 03 = 3...09 = 9
- 10 = 10; 11 = 11... 99 = 99
- **Assign the numbers to be skipped or ignored**

6. Address Duplications

Are repeats permitted: Can something occur twice?

- Percentages-usually can be duplicated
- Specific items-usually cannot be duplicated
(Occurs when the quantities of items are known)

7. Explain how to run the simulation

Explain how to run a trial

- Beginning from left to right I would select 1 or 2 digit numbers until there were ____ number of successes.
- Count how many 1 or 2 digit values that were not skipped.

Tell how many trials are going to be run

Find the average/mean of all the trials

8. Run the simulation & Make a Table

Trial Number	Number of 2 Digit Values Counted (successes)
1	
2	
3	
	Total

- Draw a line through the values that represent failures
- Circle the values that represent successes
(Do not forget about duplicates are they permitted or not)
- Mark an X through Skips
(These are numbers that are not possible- for instance when duplicates are not permitted)
- Draw a vertical line at the end of a trial
- Count the number of 2 digit numbers in the trial
- Record the values in a table
- Repeat for all necessary trials to complete the simulation

9. Calculate the Expected number

Take the average/Mean

Sum the number of successes counted for each trial

- Divide by the number of trials

Conclusion:

Based on the simulation above, on average we would expect to have

_____ in order to find _____

Give the number & definition of component

Give the number & definition of successes

Notes: Simulations Scenarios

Scenario 1 (Coupons): Every Monday a local radio station gives coupons away to 50 people who correctly answer a question about a news fact from the previous day's newspaper. The coupons given away are numbered from 1 to 50, with the first person receiving coupon 1, the second person receiving coupon 2, and so on, until all 50 coupons are given away. On the following Saturday, the radio station randomly draws numbers from 1 to 50 and awards cash prizes to the holders of the coupons with these numbers. Numbers continue to be drawn without replacement until the total amount awarded first equals or exceeds \$300. If selected, coupons 1 through 5 each have a cash value of \$200, coupons 6 through 20 each have a cash value of \$100, and coupons 21 through 50 each have a cash value of \$50.

(a) Explain how you would conduct a simulation using the random number table provided below to estimate the distribution of the number of prizewinners each week.

I would assign 01-05 a \$200 valuation, and 06-20 a \$100 valuation and 21-50 a \$50 valuation. I would skip 00 and 51-99. Working from left to right I would select 2 digits at a time, skipping any duplicates, until the sum of the valuation was greater than or equal to \$300. I would then count the number of selected 2 digit numbers between 01 and 50 which would be the number of prize winners for that trial. To increase accuracy, I would run the simulation 20 times and find the average number of winners and that would be the expected number of prize winners.

(b) Perform your simulation 3 times. (That is run 3 trials of your simulation.) Start at the leftmost digit in the first row of the table and move across. Make your procedure clear so that someone can follow what you did. You must do this by marking directly on or above the table. Report the number of winners in each of your 3 trials.

72749	^{\$50} ^{\$50} ^{\$200} 1347	65030	26128	49067	02904	49953	74674	94617	73317
81638	36566	^{\$50} ^{\$50} ^{\$100} 42709	38717	^{\$50} ^{\$100} 59943	^{\$50} ^{\$100} 12027	46547	61303	46690	76423
^{\$50} ^{\$50} 38449	46438	91579	^{\$100} ^{\$100} 01907	72146	05764	22400	94490	49890	09258

Trial 1	3 winners	\$300
Trial 2	5 winners	\$350
Trial 3	4 winners	\$300
Average	4 winners	\$317

Based on the simulation above I would expect there to be 4 prize winners a week and I would expect my weekly payout to be about 317 dollars,

Simulations (Running & Assigning)

A cereal manufacturer puts pictures of famous athletes on cards in boxes of cereal. 20% of the boxes contain Katie Ladecky, 30% of the boxes contain Michael Phelps and 50% of the boxes contain Simone Biles.

Assign 2 digit numbers in proportion to the chance of success and failure.

- 00 = 0 or 100; 01=1; 02 = 2; 03 = 3...09 = 9
- 10 = 10; 11 = 11... 99 = 99
- Assign the numbers to be skipped or ignored

01 to 20 Represents Katie Ladecky

~~21~~ to 50 Represents Michael Phelps

51 to 99 and 00 Represent Simone Biles

There are no skips and repeats are allowed.



How many boxes of cereal would you expect to buy in order to get 3 Katie Ladecky Cards?

~~52822~~ ~~48990~~ 03648 ~~34861~~ ~~54680~~ 64791
~~31645~~ ~~45552~~ ~~78255~~ ~~64794~~ ~~21228~~ 69707

Based on this simulation I would have to open 30 boxes of Cereal to get the (3) Katie Ladecky cards. This would be a stalker like obsession.

How many boxes of cereal would you expect to buy in order to get 4 Michael Phelps Cards?

~~52822~~ 48990 ~~03648~~ 34861 ~~54680~~ 64791
31645 45552 78255 64794 21228 69707

Based on this simulation I would have to open 16 boxes of Cereal to get the (4) Michael Phelps Cards.

How many boxes of cereal would you expect to buy in order to get 5 Simone Biles Cards?

52822 ~~48990~~ ~~03648~~ 34861 54680 64791
 31645 45552 78255 64794 21228 69707

Based on this simulation I would have to open 7 boxes of Cereal to get the (5) Simone Biles Cards.

Recipe for Success: Simulations

1. **Read the entire problem** What is being asked?
2. **Identify the Question** Explain what the question is asking in your own words
3. **Identify and define a success and component in context**
 - **Success:** What we want to happen
 - **Component:** What is being repeated
4. **Identify a Trial** How many successes are required?
5. **Model the simulation**

Assign 1 or 2 digit numbers in proportion to the chance of success and failure.

 - 00 = 0 or 100; 01=1; 02 = 2; 03 = 3...09 = 9
 - 10 = 10; 11 = 11... 99 = 99
 - **Assign the numbers to be skipped or ignored**
6. **Address Duplications**

Are repeats permitted: Can something occur twice?

 - Percentages-usually can be duplicated
 - Specific items-usually cannot be duplicated
(Occurs when the quantities of items are known)
7. **Explain how to run the simulation**

Explain how to run a trial

 - Beginning from left to right I would select 1 or 2 digit numbers until there were ____ number of successes.
 - Count how many 1 or 2 digit values that were not skipped.

Tell how many trials are going to be run
Find the average/mean of all the trials
8. **Run the simulation & Make a Table**

Trial Number	Number of 2 Digit Values Counted (successes)
1	
2	
3	
	Total

 - Draw a line through the values that represent failures
 - Circle the values that represent successes
(Do not forget about duplicates are they permitted or not)
 - Mark an X through Skips
(These are numbers that are not possible- for instance when duplicates are not permitted)
 - Draw a vertical line at the end of a trial
 - Count the number of 2 digit numbers in the trial
 - Record the values in a table
 - Repeat for all necessary trials to complete the simulation
9. **Calculate the Expected number**

Take the average/Mean
Sum the number of successes counted for each trial

 - Divide by the number of trials

Conclusion:

Based on the simulation above, on average we would expect to have

_____ in order to find _____
Give the number & definition of component *Give the number & definition of successes*

Notes: Simulations Scenarios

Scenario 2 (Oil Wells): : Suppose the probability that an exploratory oil well will strike oil is about 0.2. Conduct a simulation to answer the following questions. Assume that the outcome (oil or no oil) for any one exploratory well is independent of outcomes from other wells. Use the random number table below and conduct 20 trials. Clearly identify each trial on the table.

(a) Estimate the average number of wells that need to be drilled in order to strike oil.

I will work left to right choosing single digit numbers. 0 and 1 represent oil strikes & 2-9 represent wells that did not strike oil. There are no skips & repeats are allowed 20 trials were run and 98 wells were drilled.

Based on the simulation below, we would expect to have to drill an average of 4.9 wells to strike oil.

(b) What is the probability that it will take fewer than 3 attempts to strike oil?

8 out of 20 times or 40% of the time oil was struck in less than 3 times.

(c) What is the probability that it will take exactly 6 wells to strike oil?

2 out of 20 times or 10% of the time oil was found on the 6th well.

47169	80410	03333	73856	85627	54351
36653	55390	20439	48605	45513	05458
76361	47409	14914	55280	70533	52960
20579	87054	59998	90071	67554	91237
96994	65965	73235	49260	45309	24660

Trial	# of Wells	Trial	# of Wells
1	3	11	2
2	4	12	9
3	2	13	4
4	1	14	2
5	1	15	3
6	19	16	6
7	10	17	2
8	2	18	8
9	7	19	2
10	5	20	6

Notes: Samples

While economical, easy to run, and unbiased, simulations have 2 major limitations.

- 1. To run a simulation we must know something about the population.** We must know the population parameter of interest. For instance, in the oil well simulation, we needed to know the percentage of time that we expected to strike oil. Unfortunately, much of the time we often don't know anything about the population. We don't know the percent of deformed blood cells in your body; we don't know what percent of the population supports a candidate.
- 2. Simulations do not allow us to test whether or not the claimed percentage of successes is correct.** When running a simulation we are forced to assume that the percentage of successes is correct. The oil well simulation problem claimed that we struck oil twenty percent of the time, but what if the percentage had been different than 20 percent and we the drilling company were unaware of the true percentage of successful wells? We assumed that the drilling company would strike oil twenty percent of the time and used that value in our simulation. However, simulations do not provide a mechanism to determine whether or not this was actually the case.

So the question is: what do we do when we don't know a population parameter and conducting a census is not appropriate? **The answer is:** take a sample. Well that sounds simple enough, but leads us to the next question: **What is a sample?** A sample is merely a subset of the population.

Scenario: If I wanted suggestions for a theme song for this year's Reagan students, instead of performing a census and asking all of the students at Reagan for their suggestions, I could take a sample or subset of Reagan students. Since a sample is a subset of the population there are several ways to go about finding a subset.

1. I could ask twenty-five students sitting together in the lunch room.
2. I could ask my freshmen advisory and ask for their idea of the theme song.
3. I could go into the Teacher's lounge and get suggestions for the song there.
4. I could attend the PTA meeting and ask for their suggestions.

So which of those methods would you suggest I use and how do you feel about them?

I am going to go out on a limb here and guess that you don't really like any of those methods. I am betting that you are concerned that your opinion for the Reagan Theme Song may not be represented.

While samples can be an excellent way to make determinations about a population, **samples must be representative of the entire population.** So the question is: how do we take a "good sample."

First, I want to stress again that **a good sample is one that is representative of the population.** While that may seem obvious, it isn't as easy to do as one may think. If you are having trouble believing me, consider who I should ask about choosing a theme song for Reagan and give me a list of those people.

Summary:

- A sample is a portion of a population which is studied to draw conclusions & about the characteristics of the whole population.
- A sample must be representative of the population to be useful.
- No conclusions should be drawn from poorly collected data or badly designed experiments or studies.

Notes: Bias

Bias is the result of a bad study design and is sometimes referred to as **selection bias**, because the sample we **selected is biased**, that is to say that the selected sample does not represent the population.

As we stated earlier, biased data is absolutely useless and we have no methods to fix data that is not representative of the population of interest. If we happen to collect a biased sample, then we must discard it, start over and collect another sample.

In order to avoid collecting a biased sample, it is important to recognize the types of bias that exist so that we can create data collection methods that prevent the introduction of biased data into our sample.

Common types of statistical bias are:

Under-coverage Bias-Occurs when members of a population are not able to be selected or chosen for a sample because they are not able to be accounted for (*homeless, inmates, no land-line, college students*). It is desirable to have a listing of the population and then randomly select individuals from the listing for the sample. Obviously, the sample is only as good as the listing of the population and in this case the above mentioned groups are often left off of population lists.

Non-response Bias-Occurs when a member in the population is randomly identified and selected but refuses to participate in the poll or chooses not to return the survey or they were called but chose not to answer the questions in the poll. In the non-response bias situation, the sample group does mirror the population, but a person or persons in the sample group refused to participate.

Voluntary Response Bias occurs when anyone is permitted to choose to respond to a general invitation such is the case with radio call-in surveys; write in surveys; internet polls. The problem with this type of data collection is that it tends to over-represent people with strong opinions because they are the only ones who care enough to answer the poll. The person in the survey chooses to participate rather than be randomly selected from a listing of the entire population. There is no method to choosing the sample group. The respondents choose themselves.

Response Bias results from questions that are embarrassing/sensitive or are Non-neutral or poorly worded questions which due to their phrasing tend to lead people to a particular response. We minimize the chance of response bias by carefully wording and field testing questions. If a question is sensitive or embarrassing the respondent may choose to lie. Poorly worded questions tend to influence the responses of individuals.

Example of a sensitive question where someone might be prone to lie:

Coach Ham asks the team if anyone one of them had stayed out partying that weekend. Of course, none of the athletes were.

Examples where the wording of the questions influenced the response:

Are you in favor of providing school lunches for children that do not have enough to eat so that they have a chance to have a normal life?

Are you in favor of the government raising your taxes and taking your money away to pay for food for people who don't work? Obviously, the two questions will result in different responses.

Notes: Bias Scenarios

Scenario 1: Interested in determining the percent of the population who believed in God, a surveyor stood outside a church on Sunday morning and asked all of the congregates a neutrally worded question about whether or not they believe in God. Will this survey produce biased results? Explain.

It is reasonable to assume that most people who attend church believe in God. Consequently, the percentage of people who believe in God is probably over-represented and the people who do not believe in God are under-represented. This case of SELECTION BIAS—selecting a really poor sample is subject to the following biases:

Undercoverage Bias: Those that do not believe in God are not likely to be at church.

Response bias: If I am outside of a church I may feel pressured to say that I believe in God because of all of those around me

2008 Question 2: A local school board plans to conduct a survey of parents' opinions about year-round schooling in elementary schools. The school board contacts 500 of the families by mail. The survey question is provided below.

A proposal has been submitted that would require students in elementary schools to attend school on a year-round basis. Do you support this proposal? (Yes or No)

The school board received responses from 98 of the families, with 76 of the responses indicating support for year-round schools. Based on this outcome, the local school board concludes that most of the families with at least one child in elementary school prefer year-round schooling.

- a. What type of bias is included in this survey?
- b. Could we just contact another 500 families and add their data to the original results?
- c. Name 2 ways to address the bias in this survey?
 - a. **Non-response bias:** Only 98 out of 500 families or 19.6% responded to the survey. It is likely that the opinions of parents who are most in favor of year round schools were the ones who responded and there are over-represented in the survey.
 - b. **NO, this would not be reasonable.** Even after increasing the sample size the non-response bias would still be present because there are still 402 of 500 families that did not respond from the first survey. Remember you cannot fix biased data.
 - c. **1st** The school board could contact the 402 families who did not respond and try to get their responses.
 - 2nd** The school board could take a completely new survey using phone calls or person to person contact to try to eliminate the amount of non-response bias.

Notes: Bias Scenarios

2004 Form B Question 2: At a certain university, students who live in the dormitories eat at a common dining hall. Recently, some students have been complaining about the quality of the food served there. The dining hall manager decided to do a survey to estimate the proportion of students living in the dormitories who think that the quality of the food should be improved. One evening, the manager asked the first 100 students entering the dining hall to answer the following question.

Many students believe that the food served in the dining hall needs improvement. Do you think that the quality of food served here needs improvement, even though that would increase the cost of the meal plan?

_____ Yes _____ No _____ No opinion

- (a) In this setting, explain how bias may have been introduced based on the way this convenience sample was selected.
- (b) In this setting, explain how bias may have been introduced based on the way the question was worded
- (c) How could the question have been worded differently to avoid that bias?

(a) The first 100 students who eat at the dining hall may have very different opinions than those that come later. It is likely that the food is fresher when the first 100 students arrive and it is likely that the first 100 students have a more positive attitude about the food as evidenced by the fact they are willing to arrive early.

The first 100 are probably hungrier and food always tastes better when you are really hungry.

The first 100 are actually excited about the food (freshmen) and rush to get to the cafeteria because they really enjoy it.

(b) Bias is introduced by including the phrase, "Many students believe that the food served in the dining hall needs improvement" which could cause many students to go along with the norm and say yes rather than giving their own opinion.

Bias is introduced by including the phrase, "even though that would increase the cost of the meal plan" a cost consideration is included and students may think the food is awful but are unwilling or unable to pay more for their meals.

(c) If they were wanting to know the opinion of the students about the food quality, the question could be worded "Do you think that the quality of food served here needs improvement?"

Notes: Sampling Error

Recall that we defined **statistics** as the study of how to collect, organize, display, describe, and analyze the data from a sample in order to make generalizations about a population parameter of interest. Also remember that the parameter is a numerical measurement of a population which could be the mean, median, variance, range, percent...etc. From the sample, or subset of the population, we find estimates of population parameters. **The estimate of a population parameter that comes from a sample is known as a sample statistic and is often referred to as a statistic.** Thus, we talk about the sample mean, or sample median or sample variance...etc. being estimates or approximations for the corresponding population parameters.

For instance, if I wanted to know the true number of hours that Reagan students sleep on average per week, I could find an estimate by taking a sample 50 Reagan students. If I took a "good sample" one that represents the population at Reagan, my findings should be pretty close to the actual number of hours that are slept per night. However, the value I calculated from my sample is likely to differ somewhat from the actual number of hours slept. The difference between my sample statistic and the population parameter is known as sampling error or sampling variation.

More specifically, **Sampling Error or Sampling Variation** is variation due to random chance and can be described by a probability model and can be minimized by increasing the size of the sample. Sampling error is the recognition that every randomly selected sample from a population is likely to be different. Because there are so many variables or factors within a population, it is nearly impossible to design studies that account for every variable without drawing a census, which in itself is difficult, if not impossible. Because we cannot account for all of the variables, we involve chance in the selection of our sample to equalize the effects of variables that could not be accounted for. Even with a well-designed study or experiment, **the sample statistic is only an estimate of the population parameter.** The difference between the two is known as sampling error.

Recall that when we ran our simulations, there were multiple approaches that were all considered acceptable and led to different responses. However as we increased the sample size, the amount of variation in our responses decreased. It makes sense that if we increase our sample size that our result would be closer to the true population parameter. We can say that as we increase our sample size that the difference between the population parameter and the sample statistic will be less and that leads us to say that **the amount of variation diminishes as our sample size increases.** Obviously, our observations may be more varied as the sample size increases, but on average the sample statistics will vary less.

Caution: Please **do not confuse sampling error with Bias.** Sampling error is a term used to describe properly collected data that didn't actually match the population. In other words, given a proper method of data collection was employed, I would expect that my sample will vary or differ from the population the amount of variation or difference of a sample from the population is error. Error is normal and is predictable and therefore can be controlled. **The best way to minimize sampling error/sampling variation is by increasing the sampling size.**

Unlike sampling error, bias is problem in the design and/or method of collecting data. Increasing sample size will not fix biased data. Biased data cannot be fixed and must be discarded. To eliminate bias, we must carefully plan and execute how we collect our data.

Notes: Observational Studies & Randomness

Obviously, it is important to collect samples that are representative of the population and by now you should be asking yourself: “**how do we do that?**” One primary method used to collect samples is the Observational Study. Observational studies are studies in which we gather data about subjects doing what they are already doing. The research of historical records or medical records and surveys are examples of observational studies. Most data is collected through observational studies. Observational studies do NOT assign treatments. **Because treatments are not assigned, causation cannot be proved.** However, observational studies can be statistically significant and can demonstrate a relationship, association or a correlation.

**Note: Only Experiments should be used to establish cause and effect relationships.
Most data is collected by observational study and not by controlled experiments.**

A common type of observational study is the survey of which there are several types including:

- The Census
- The Convenience Sample
- The Simple Random Sample
- The Systematic Sample
- The Cluster Sample
- The Stratified Sample

The Census: As we discussed earlier, a census involves the collection of data from each unit in the population. Unfortunately, a census is often impractical if not impossible to perform. On the positive side, if we perform an accurate census, we will have the population parameter of interest and inference procedures will not be necessary.

The Convenience Sample: The convenience sample is a sampling technique which selects subjects because of their convenient accessibility and proximity to the researcher. Convenience samples are rarely representative of the population and not every subject in the population has an equal chance of being selected. As a consequence, convenience sample findings are most often of no value, however, they are easy to perform if you are interested in collecting worthless data. In other words, **avoid them.**

The purpose of Randomness.

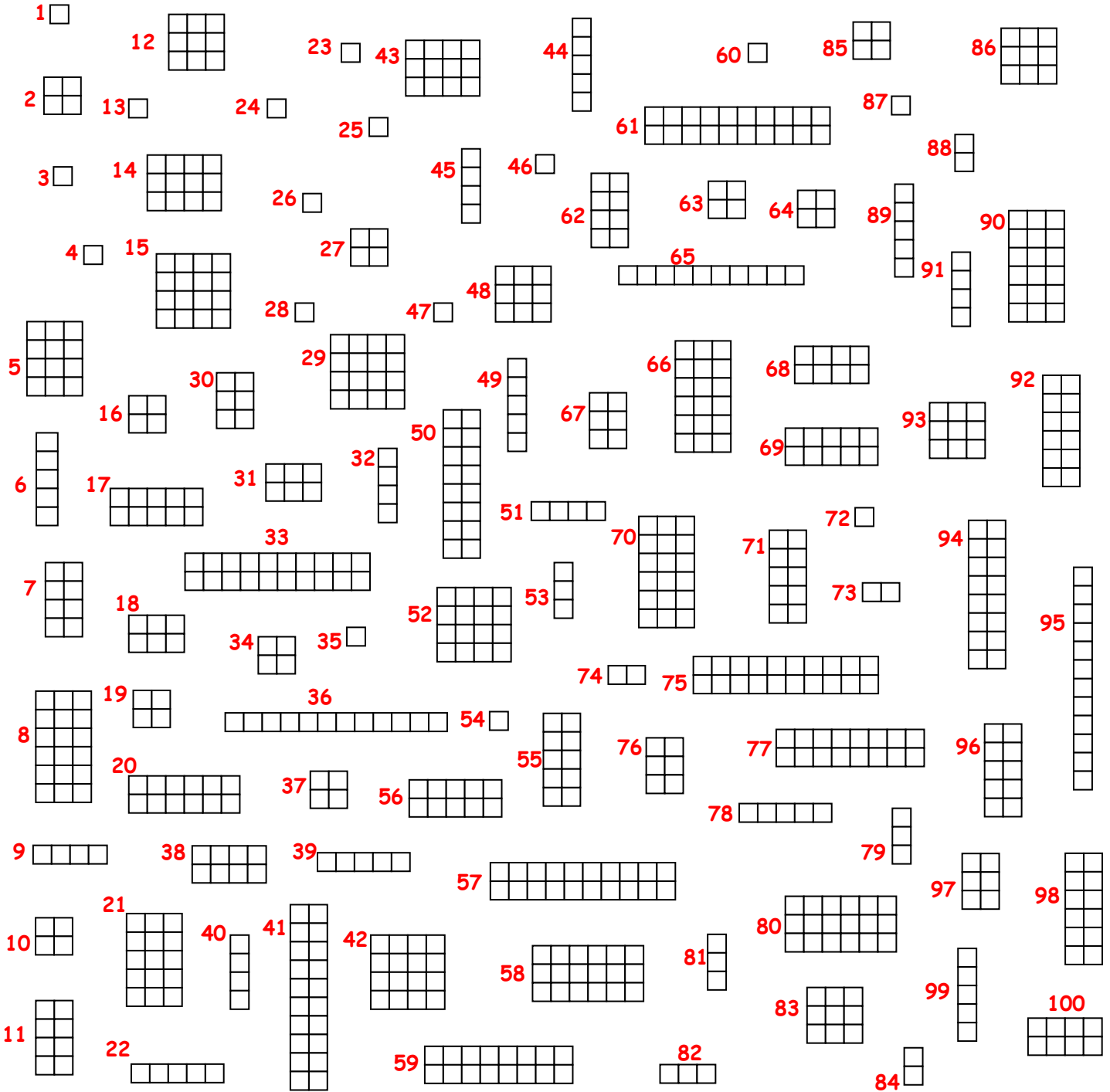
Recall that we want our samples to be representative of the population and that means we would desire that we would like every combination of traits to be proportionally represented in our sample. Unfortunately, if we really break it down, we are all unique individuals. Thus if we were to represent every combination of traits, we would need to perform a census and sample everyone. As we discussed previously, this is impractical if not impossible.

So how do we address the problem? The answer is the introduction of chance or randomness. Statisticians and mathematicians have found that randomness, tends to equalize the variations that we can't account for in our studies.

So you don't believe me. Try the exercise on the next page.

Notes: Sample Surveys & Random Rectangles

So let's begin with your best guess as to the mean (average) number of squares in each figure? _____
Please use whole numbers.



Judgment Sample - Select a sample of 5 rectangles that, in your judgment, re representative of the rectangles. Record the size of your five rectangles and find their mean.

_____ mean = _____

Number each figure 1 to 100. The 100 figures make up our population of rectangles.

Notes: Sample Surveys & Random Rectangles

So how do we determine the average number of rectangles per figure? Obviously, we could conduct a census and count the squares in each figure, and take an average. However, I am guessing that many of you would refuse to do the assignment and I am guessing that among those that did the assignment we would still have varying answers. So instead of conducting a census or just using our best guess or our judgment, let's utilize some sampling techniques that employ randomness. We are going to use samples of size 5 which are pretty small. We will begin with the simple random sample.

The Simple Random Sample: This is the basis for all statistical sampling and differs from census and the convenience sample in that randomness is utilized. In a simple random sample all subjects have an equal chance of being selected and every possible combinations of subjects has an equal chance of being selected. That sentence is so important, that I am going to say it again: **In a simple random sample all subjects have an equal chance of being selected and every possible combinations of subjects has an equal chance of being selected.**

So how do we conduct a Simple Random Sample? Typically a number is assigned to all subjects in the population and then a sample of numbers is randomly selected. Acceptable methods of selecting the numbers include: random number generators; mixed numbers in a hat; random number tables. If a **random number table** is used, it is necessary to describe the **selection process**, provide the **stopping point** and define how **duplicate values** are to be addressed. Unfortunately, it is often difficult to obtain a list of the entire population and there is a chance that our sample will not be representative of the population.

Use the random number generator on your calculator to select 5 numbers for our sample.

Calculator Commands:

Press **MATH** → Highlight **PROB**
 ↓ Highlight **8: randIntNoRep(**
 Press **ENTER**
Lower: 1 Upper: 100 n: 5 Press **ENTER 2x's**

Locate the corresponding rectangles and record their size and find the mean.

10	12	1	16	18	mean = <u>11.4</u>
69	95	26	50	58	

The Systematic Random Sample: The systematic sample also involves randomness and every individual has an equal opportunity of being selected, however every group is not possible. To perform a systematic random sample begin by dividing the population by your desired sample size which gives us n . We will be selecting every n th value. Systematic sampling is fast and easy and is a good method if there is no pattern in the population. We often use it when we want to survey people coming into a building (selecting every n th person) or every n th thing in a list.

In our case, we will divide 100 by 5 and get 20. Thus we will be selecting every 20th figure in the list. We will also use the n to find our starting point. We do this by using the random number generator and selecting one value 1 to 20. Please perform a systematic random sample.

4	5	16	3	5	mean = <u>6.6</u>
19	39	59	79	99	

Notes: Sample Surveys & Random Rectangles

<p>Cluster 1</p>	<p>Cluster 2</p>	<p>Cluster 3</p>	<p>Cluster 4</p>	
<p>Cluster 5</p>	<p>Cluster 6</p>	<p>Cluster 7</p>	<p>Cluster 8</p>	
<p>Cluster 9</p>	<p>Cluster 10</p>	<p>Cluster 11</p>	<p>Cluster 12</p>	
<p>Cluster 13</p>	<p>Cluster 14</p>	<p>Cluster 15</p>	<p>Cluster 16</p>	<p>Cluster 17</p>
<p>Cluster 18</p>	<p>Cluster 19</p>	<p>Cluster 20</p>		

Notes: Sample Surveys & Random Rectangles

The Cluster Random Sample: To perform a cluster sample, the population is broken into smaller groups or clusters that are representative of the population. In other words, each cluster is a microcosm of the population. Once the clusters are identified, we number them and then randomly select one or more clusters. Once a cluster is selected, we then perform a census within the cluster, which is to say, that within each selected cluster we survey or sample every subject.

Cluster sampling tends to be difficult to perform with humans. For example, how would you divide Reagan up if you wanted to perform a cluster sample? Consider how you might divide the state of Texas or San Antonio into clusters. Keep in mind that each cluster needs to mirror the population, which means that each cluster is very similar to the other clusters.

You might be thinking that cluster sampling isn't useful or important. However, there are several instances in which cluster sampling is important. Here are a three scenarios:

Drawing Blood: When a lab needs to test a person's blood for a disease. They draw a small sample and then place it under the microscope. That small sample of blood under the microscope is a cluster and is representative of all of the blood in someone's body.

Baking: When you make cookies or cakes or brownies, you mix all of the ingredients together thoroughly and then you take a spoonful and taste it and make adjustments as necessary. That spoonful is a cluster and is representative of all the batter that is being made. I certainly could use a chocolate chip cookie about now. Hint. Hint...

Forrest Gump: "Life is like a box of chocolates. You never know what you're going to get." However, my expectation is that every box of Russell Stover's Chocolate Truffles is the same. So if I wanted to perform a cluster sample, I would randomly select a box, (this would be my cluster). Now remember to perform a cluster sample, I must perform a census within the chosen group. To perform a census in this scenario would mean that I would need to take a bite out of every truffle. Oh how wonderful. Incidentally I may need to increase my sample size which would mean that I would need more boxes of chocolate and would need to eat all of the chocolate. It's a tough job, but someone's got to do it.

The rectangles have been broken into 20 clusters of 5 which are representative of the population of rectangles. Number each cluster 1 to 20 and randomly select one of the clusters using the random number generator on your calculator

Calculator Commands:

Press **MATH** → Highlight **PROB**

↓ Highlight **8: randIntNoRep(**

Press **ENTER**

Lower: 1 Upper: 20 n: 1 Press **ENTER 2x's**

Locate the corresponding cluster and record the number of rectangles in each figure and find the mean of your sample

4	6	2	12	18	
Cluster 16	Cluster 16	Cluster 16	Cluster 16	Cluster 16	mean = <u>8.4</u>

Notes: Sample Surveys & Random Rectangles

Label each of the 5 sections of rectangles below based on the number of rectangles in each figure.

Strata 1

Strata 2

Strata 3

Strata 4

Strata 5

Number the figures in each section 1 to 20.

Notes: Sample Surveys & Random Rectangles

The Stratified Random Sample: To perform a stratified sample the population is divided into groups called strata based on a common variable or characteristic. We then select a random sample from within each group. Although it is not required, typically, the size of the random sample within each group is proportional to the variables representation within the population. It is important to note that the strata are not determined randomly. Randomness is introduced through the selection within the strata. Please understand that although we take a simple random sample within each strata, a stratified random sample **IS NOT** a special case of the simple random sample.

We take stratified samples for two important reasons. First, a stratified random sample allows us to ensure that each of the subgroups of interest within a population are represented in the sample. Secondly, we are able to isolate the impact of the common characteristic that determined the groups. Please note that unlike cluster samples (where each cluster is a microcosm of the population), the individual strata or groups are not representative of the population.

Dividing groups by gender is one of the most common ways we stratify and we do this because we believe that in general there are differences in males and females. However, my go to example for stratified sample is our high school. For Instance, let's say that we were trying to choose a theme song for homecoming and instead of performing a census and surveying everyone, we decide to survey the opinions of 100 students. We could do a simple random sample, however we run a risk of having chosen all freshmen in the sample. I don't think the seniors would be very pleased to have freshmen making decisions for them.

However, if we divided Reagan by grade level (freshmen, sophomore, junior and senior) and then performed a simple random sample within each grade level. We would be assured of having every grade level represented. In addition, since we know that 28% of the school is comprised of freshmen, 26% are sophomores, 23% are juniors and 23% are seniors, we could proportionally represent each grade level and our sample would be representative of the Reagan population.

Use the random number generator on your calculator to select 5 numbers for our sample.

Calculator Commands:

Press **MATH** → Highlight **PROB**
 ↓ Highlight **8: randIntNoRep(**
 Press **ENTER**
Lower: 1 Upper: 20 n: 1 Press **ENTER 2x's**

Locate the corresponding figure and record its size.

Note: Repeat the process 5 times (once for each strata)

Calculate the mean.

1	4	8	10	15	
Strata 1 Element 3	Strata 2 Element 11	Strata 3 Element 5	Strata 4 Element 20	Strata 5 Element 6	mean = <u>7.6</u>

Notes: Survey Scenarios

2011 Question 3: An apartment building has nine floors and each floor has four apartments. The building owner wants to install new carpeting in eight apartments to see how well it wears before she decides whether to replace the carpet in the entire building.

The figure below shows the floors of apartments in the building with their apartment numbers. Only the nine apartments indicated with an asterisk (*) have children in the apartment.

<table border="1"> <tr><td>11*</td><td>12</td></tr> <tr><td colspan="2" style="text-align: center;">1st Floor</td></tr> <tr><td>14</td><td>13</td></tr> </table>	11*	12	1st Floor		14	13	<table border="1"> <tr><td>21</td><td>22*</td></tr> <tr><td colspan="2" style="text-align: center;">2nd Floor</td></tr> <tr><td>24</td><td>23*</td></tr> </table>	21	22*	2nd Floor		24	23*	<table border="1"> <tr><td>31</td><td>32</td></tr> <tr><td colspan="2" style="text-align: center;">3rd Floor</td></tr> <tr><td>34</td><td>33</td></tr> </table>	31	32	3rd Floor		34	33			
11*	12																						
1st Floor																							
14	13																						
21	22*																						
2nd Floor																							
24	23*																						
31	32																						
3rd Floor																							
34	33																						
<table border="1"> <tr><td>41</td><td>42</td></tr> <tr><td colspan="2" style="text-align: center;">4th Floor</td></tr> <tr><td>44</td><td>43</td></tr> </table>	41	42	4th Floor		44	43	<table border="1"> <tr><td>51*</td><td>52</td></tr> <tr><td colspan="2" style="text-align: center;">5th Floor</td></tr> <tr><td>54</td><td>53</td></tr> </table>	51*	52	5th Floor		54	53	<table border="1"> <tr><td>61</td><td>62</td></tr> <tr><td colspan="2" style="text-align: center;">6th Floor</td></tr> <tr><td>64</td><td>63</td></tr> </table>	61	62	6th Floor		64	63	<table border="1"> <tr><td colspan="2" style="text-align: center;">* = Children in the Apartment</td></tr> </table>	* = Children in the Apartment	
41	42																						
4th Floor																							
44	43																						
51*	52																						
5th Floor																							
54	53																						
61	62																						
6th Floor																							
64	63																						
* = Children in the Apartment																							
<table border="1"> <tr><td>71</td><td>72</td></tr> <tr><td colspan="2" style="text-align: center;">7th Floor</td></tr> <tr><td>74*</td><td>73*</td></tr> </table>	71	72	7th Floor		74*	73*	<table border="1"> <tr><td>81</td><td>82</td></tr> <tr><td colspan="2" style="text-align: center;">8th Floor</td></tr> <tr><td>84*</td><td>83</td></tr> </table>	81	82	8th Floor		84*	83	<table border="1"> <tr><td>91</td><td>92*</td></tr> <tr><td colspan="2" style="text-align: center;">9th Floor</td></tr> <tr><td>94</td><td>93*</td></tr> </table>	91	92*	9th Floor		94	93*			
71	72																						
7th Floor																							
74*	73*																						
81	82																						
8th Floor																							
84*	83																						
91	92*																						
9th Floor																							
94	93*																						

- (a) For convenience, the apartment building owner wants to use a cluster sampling method, in which the floors are clusters, to select the eight apartments. Describe a process for randomly selecting eight different apartments using this method.

Assign 1 to 1st floor, 2 to 2nd floor... and 9 to 9th floor. Place the numbers 1 through 9 in a hat and mix thoroughly. Draw 2 numbers without replacement. Survey all apartments on the floors represented by the two numbers.

- (b) An alternative sampling method would be to select a stratified random sample of eight apartments, where the strata are apartments with children and apartments with no children. A stratified random sample of size eight might include two randomly selected apartments with children and six randomly selected apartments with no children. In the context of this situation, give one statistical advantage of selecting such a stratified sample as opposed to a cluster sample of eight apartments using the floors as clusters.

With a cluster sample it would be possible to survey apartment clusters which have no children. The 3rd, 4th and 6th floor clusters do not have children. Using a cluster it is also possible to not select apartments that are adult only. Apartments with children are likely to have had more wear than apartments that have adults only. A stratified sample would ensure that apartments with children and those without were sampled.

- (c) Why did we choose to stratify on apartments with and without children?

We always stratify on the variable that we think will most impact our study. We were measuring carpet wear and we believe that having children impacts carpet wear differently than just having adults only

Notes: Survey Scenarios

2010 Form B Question 2: In response to nutrition concerns raised last year about food served in school cafeterias, the Smallville School District entered into a one-year contract with the Healthy Alternative Meals (HAM) company. Under this contract, the company plans and prepares meals for 2,500 elementary, middle, and high school students, with a focus on good nutrition. The school administration would like to survey the students in the district to estimate the proportion of students who are satisfied with the food under this contract.

Two sampling plans for selecting the students to be surveyed are under consideration by the administration. One plan is to take a simple random sample of students in the district and then survey those students. The other plan is to take a stratified random sample of students in the district and then survey those students.

- (a) Describe a simple random sampling procedure that the administrators could use to select 200 students from the 2,500 students in the district.

Assign all students a number from 1 to 2500. Place the 2500 numbers in a bin and mix thoroughly. Draw 200 numbers without replacement and then survey the 200 students whose names correspond with the number drawn.

- (b) If a stratified random sampling procedure is used, give one example of an effective variable on which to stratify in this survey. Explain your reasoning.

Because of potential difference in food preferences as a student ages I would choose to stratify on school type. My three strata would be elementary, middle and high school. As people age their nutritional requirements and tastes often change, consequently food satisfaction is likely to differ by school type.

- (c) Describe one statistical advantage of using a stratified random sample over a simple random sample in the context of this study.

Stratifying based on school type, assures that satisfaction ratings are obtained from elementary, middle and high school other strategies could miss a type of school completely and that would be an issue if food satisfaction varied by age.

Notes: Survey Scenarios

2013 Question 3: An administrator at a large university wants to conduct a survey to estimate the proportion of students who are satisfied with the appearance of the university buildings and grounds. The administrator is considering three methods of obtaining a sample of 500 students from the 70,000 students at the university.

- (a) Because of financial constraints, the first method the administrator is considering consists of taking a convenience sample to keep the expenses low. A very large number of students will attend the first football game of the season, and the first 500 students who enter the football stadium could be used as a sample. Why might such a sampling method be biased in producing an estimate of the proportion of students who are satisfied with the appearance of the buildings and grounds?
- (b) Because of the large number of students at the university, the second method the administrator is considering consists of using a computer with a random number generator to select a simple random sample of 500 students from a list of 70,000 student names. Describe how to implement such a method.
- (c) Because stratification can often provide a more precise estimate than a simple random sample, the third method the administrator is considering consists of selecting a stratified random sample of 500 students. The university has two campuses with male and female students at each campus. Under what circumstance(s) would stratification by campus provide a more precise estimate of the proportion of students who are satisfied with the appearance of the university buildings and grounds than stratification by gender?

- (a) The sample may be biased because it is likely that the first 500 students arriving to the stadium are not representative of the entire population of students at the university. For example the 1st 500 might have substantially more school spirit and would likely give a greater percentage of favorable ratings for the buildings than the student body as a whole.
- (b) Assign each student a number 1-70,000. Use a random number generator to select 500 unique values. Duplicate values are to be ignored. The 500 students who correspond to those numbers will be interviewed.
- (c) We always stratify on the variable that we believe will have the greatest impact on the study. Because there are two separate campuses, it makes sense to stratify by campus rather than by gender if the buildings at each cam