4.1 Sampling and Surveys

Learning Objectives

1. Identify the population and sample in a statistical study.
2. Identify voluntary response samples and convenience samples. Explain how these sampling methods can lead to bias.
3. Describe how to obtain a random sample using slips of paper, technology, or a table of random digits.
4. Distinguish a simple random sample from a stratified random sample or cluster sample. Give the advantages and disadvantages of each sampling method.
5. Explain how undercoverage, nonresponse, question wording, and other aspects of a sample survey can lead to bias.

Vocabulary: population, census, sample, bias, convenience sample, voluntary response sample, random sampling, simple random sample, stratified random sample, strata, cluster sample, clusters, inference, undercoverage, nonresponse

Activity: Sampling from *The Federalist Papers*

The Federalist Papers are a series of 85 essays supporting the ratification of the U.S. Constitution. At the time they were published, the identity of the authors was a secret known to just a few people. Over time, however, the authors were identified as Alexander Hamilton, James Madison, and John Jay. The authorship of 73 of the essays is fairly certain, leaving 12 in dispute. However, thanks in some part to statistical analysis[1], most scholars now believe that the 12 disputed essays were written by Madison alone or in collaboration with Hamilton[2].

There are several ways to use statistics to help determine the authorship of a disputed text. One example is to estimate the average word length in a disputed text and compare it to the average word lengths of works where the authorship is not in dispute.

Directions: The following passage is the opening paragraph of *Federalist Paper #51*[3], one of the disputed essays. The theme of this essay is the separation of powers between the three branches of government. Choose 5 words from this passage, count the number of letters in each of the words you selected and find the average word length. Share your estimate with the class and create a class dotplot.

> To what expedient, then, shall we finally resort, for maintaining in practice the necessary partition of power among the several departments, as laid down in the Constitution? The only answer that can be given is, that as all these exterior provisions are found to be inadequate, the defect must be supplied, by so contriving the interior structure of the government as that its several constituent parts may, by their mutual relations, be the means of keeping each other in their proper places. Without presuming to undertake a full development of this important idea, I will hazard a few general observations, which may perhaps place it in a clearer light, and enable us to form a more correct judgment of the principles and structure of the government planned by the convention.

[1] Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Mass., 1964.

[2] http://en.wikipedia.org/wiki/Federalist_papers

[3] http://www.constitution.org/fed/federa51.htm

Directions: Use a table of random digits or a random number generator to select a simple random sample (SRS) of 5 words from the opening passage to the *Federalist Paper #51*. Once you have chosen the words, count the number of letters in each of the words you selected and find the average word length. Share your estimate with the class and create a class dotplot. How does this dotplot compare to the first one? Can you think of any reasons why they might be different?

Number	Word	Number	Word	Number	Word
1	To	44	To	87	A
2	What	45	Be	88	Full
3	Expedient	46	Inadequate	89	Development
4	Then	47	The	90	Of
5	Shall	48	Defect	91	This
6	We	49	Must	92	Important
7	Finally	50	Be	93	Idea
8	Resort	51	Supplied	94	I
9	For	52	By	95	Will
10	Maintaining	53	So	96	Hazard
11	In	54	Contriving	97	A
12	Practice	55	The	98	Few
13	The	56	Interior	99	General
14	Necessary	57	Structure	100	Observations
15	Partition	58	Of	101	Which
16	Of	59	The	102	May
17	Power	60	Government	103	Perhaps
18	Among	61	As	104	Place
19	The	62	That	105	It
20	Several	63	Its	106	In
21	Departments	64	Several	107	A
22	As	65	Constituent	108	Clearer
23	Laid	66	Parts	109	Light
24	Down	67	May	110	And
25	In	68	By	111	Enable
26	The	69	Their	112	Us
27	Constitution	70	Mutual	113	To
28	The	71	Relations	114	Form
29	Only	72	Be	115	A
30	Answer	73	The	116	More
31	That	74	Means	117	Correct
32	Can	75	Of	118	Judgment
33	Be	76	Keeping	119	Of
34	Given	77	Each	120	The
35	Is	78	Other	121	Principles
36	That	79	In	122	And
37	As	80	Their	123	Structure
38	All	81	Proper	124	Of
39	These	82	Places	125	The
40	Exterior	83	Without	126	Government
41	Provisions	84	Presuming	127	Planned
42	Are	85	To	128	By
43	Found	86	Undertake	129	The
				130	Convention

Read Article: How Forensic Linguistics outed J.K. Rowling

Read 209–211

The ______________________________ in a statistical study is the entire group of __________________________ we want information about.

A ____________________ is a subset of ____________________________ in the population from which we actually collect data.

A ______________________ collects data from every individual in the population.

Example: *Identify the population and sample in each of the following settings.*
(a) The student government at a high school surveys 100 students to get their opinions about a change to the bell schedule.

(b) The quality control manager at a bottling company selects 10 cans from the production line every hour to see whether the volume of soda is within acceptable limits.

What is that icon in the top-right corner of the example on page 210?

Read 211–213 (How to Sample Badly)
Choosing individuals from the population who are easy to reach results in a ________________________sample.

What's the problem with convenience samples?

.

What is bias?

AP Exam Tip: If you're asked to describe how the design of a study leads to bias, you're expected to do two things:
(1) identify a problem with the design, and
(2) explain how this problem would lead to an underestimate or overestimate

What's a voluntary response sample? Is this a good method for obtaining a sample?

Alternate Example: To estimate the proportion of families that oppose budget cuts to the athletic department, the principal surveys families as they enter the football stadium on Friday night. Explain how this plan will result in bias and how the bias will affect the estimated proportion.

Alternate Example: A recent online poll posed the question "Should female athletes be paid the same as men for the work they do?" In all, 13, 147 (44%) said, "Yes." 15, 182 (50%) said, "No." The remaining 1, 448 said, "Don't know." In spite of the large sample size for this survey, we can't trust the results. Why not?

HW page 229 (1, 3, 6, 7, 9, 10) – Remember to check your answers in the back of the book & make corrections in pen!

4.1 Random Sampling Methods

Read 213–217

What's a simple random sample (SRS)? How can you choose a SRS?

What's the difference between sampling *with* replacement and sampling *without* replacement? How should you account for this difference when using a table of random digits or other random number generator?

Alternate Example: Mall Hours
The management company of a local mall plans to survey a random sample of 3 stores to determine the hours they would like to stay open during the holiday season. Use Table D at line 101 to select an SRS of size 3 stores.

Aeropostale
All American Burger
Arby's
Barnes & Noble
Carter's for Kids
Destination Tan
Famous Footwear
Forever 21
GameStop
Gymboree
Haggar
Just Sports
Mrs. Fields
Nike Factory Store
Old Navy
Pac Sun
Panda Express
Payless Shoes
Star Jewelers
Vitamin World
Zales Diamond Store

From Table D: 19223 95034 05756 28713 96409 12531 42544

Suppose we wanted to estimate the yield of our corn field. The field is square and divided into 16 equally sized plots (4 rows x 4 columns). A river runs along the eastern edge (right) of the field. We want to take a sample of 4 plots.

Using a random number generator, pick a **simple random sample (SRS)** of 4 plots. Place an X in the 4 plots that you choose.

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

Now, randomly choose one plot from each horizontal row. This is called a **stratified random sample**.

1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4

Finally, randomly choose one plot from each vertical
column. This is also a **stratified random sample**.

1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4

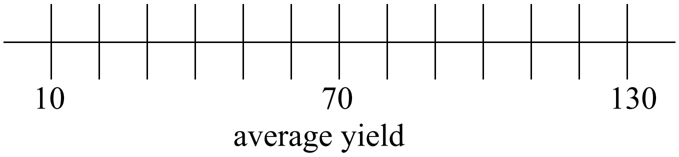
Which method do you think will work the best? Explain.

Now, its time for the harvest! The numbers below are the yield for each of the 16 plots. For each of your three samples above, calculate the average yield.

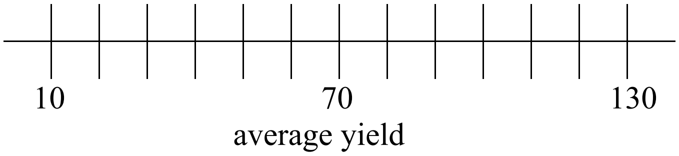
4	29	94	150
7	31	98	153
6	27	92	148
5	32	97	147

Graphing the results:

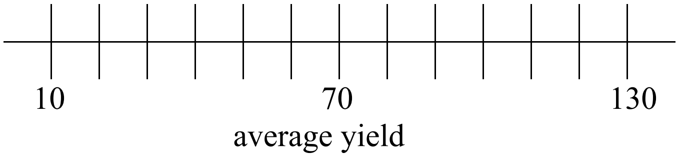
Simple Random Sample:



Stratified by Row:



Stratified by Column:



Read 219–220

What is a stratified random sample?

How is it different than a simple random sample?

When is it beneficial to use a stratified random sample? What is the benefit? How do you choose a variable to stratify by?

HW: page 230 (11, 13, 15, 17, 18, 19) - Remember to check your answers in the back of the book & make corrections in pen!

4.1 More about Sampling & Inference

Read 221–222

What is a cluster sample? Why do we use a cluster sample? How is it different than a stratified sample?

Alternate Example: A Good Read
A school librarian wants to know the average number of pages in all the books in the library. The library has 20,000 books, arranged by type (fiction, biography, history, and so on) in shelves that hold about 50 books each.
(a) Explain how to select a simple random sample of 500 books

(b) Explain how to select a stratified random sample of 500 books. Explain your choice of strata and one reason why this method might be chosen.

(c) Explain how to select a cluster sample of 500 books. Explain your choice of cluster and one reason why this method might be chosen.

(d) Discuss a potential drawback with each of the methods described above.

Read 223–225

What is inference?

What is a margin of error?

What is the benefit of increasing the sample size?

Read 225–227

What is a sampling frame?

What is undercoverage and what problems might undercoverage cause?

What is nonresponse and what problems might nonresponse cause? How is it different than voluntary response?

What is response bias and what problems might response bias cause?

HW: page 231 (21, 23, 25, 27, 30, 31, 33, 35) Remember to check your answers in the back of the book & make corrections in pen!