

## **Powerful Data Collection: Understanding Stratified Sampling and Blocked Experiments**

*In statistics, students learn about stratified random sampling and experiments that use blocking. However, students often don't understand the benefits of these methods and when to use them. In this session, we will use activities to show how blocking can improve an experimental design and stratified sampling can produce more precise results.*

**Objectives:** At the end of the session, participants will:

- Obtain classroom-tested examples and activities that show how stratified sampling produces more precise estimates.
- Obtain classroom-tested examples and activities that show how using blocking in experiments helps account for variation in the response.
- Understand when stratifying and blocking are appropriate and how they increase power.

## Stratified Random Sampling: The River Problem

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 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

| river

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

| river

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

| river

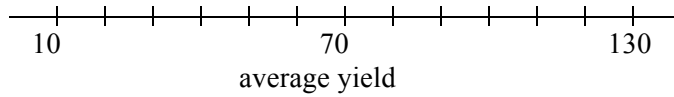
Which method do you think will work the best? Explain.

It's 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.

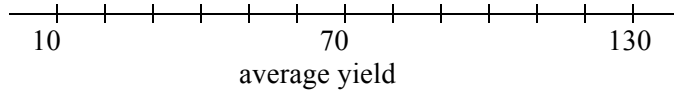
4	29	94	150
7	31	98	153
6	27	92	148
5	32	97	147

**Graphing the results:**

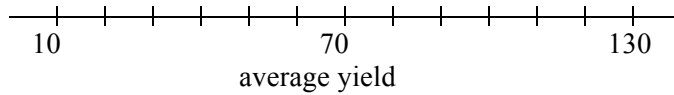
Simple Random Sample:



Stratified by Row:



Stratified by Column:



What is the benefit of using a stratified random sample? What is the best way to choose a variable to stratify by?

## 2012 AP<sup>®</sup> Exam #6: The Soda Problem

6. Two students at a large high school, Peter and Rania, wanted to estimate  $\mu$ , the mean number of soft drinks that a student at their school consumes in a week. A complete roster of the names and genders for the 2,000 students at their school was available. Peter selected a simple random sample of 100 students. Rania, knowing that 60 percent of the students at the school are female, selected a simple random sample of 60 females and an independent simple random sample of 40 males. Both asked all of the students in their samples how many soft drinks they typically consume in a week.

(a) Describe a method Peter could have used to select a simple random sample of 100 students from the school.

Peter and Rania conducted their studies as described. Peter used the sample mean  $\bar{X}$  as a point estimator for  $\mu$ . Rania used  $\bar{X}_{\text{overall}} = (0.6)\bar{X}_{\text{female}} + (0.4)\bar{X}_{\text{male}}$  as a point estimator for  $\mu$ , where  $\bar{X}_{\text{female}}$  is the mean of the sample of 60 females and  $\bar{X}_{\text{male}}$  is the mean of the sample of 40 males.

Summary statistics for Peter's data are shown in the table below.

Variable	N	Mean	Standard Deviation
Number of soft drinks	100	5.32	4.13

(b) Based on the summary statistics, calculate the estimated standard deviation of the sampling distribution (sometimes called the standard error) of Peter's point estimator  $\bar{X}$ .

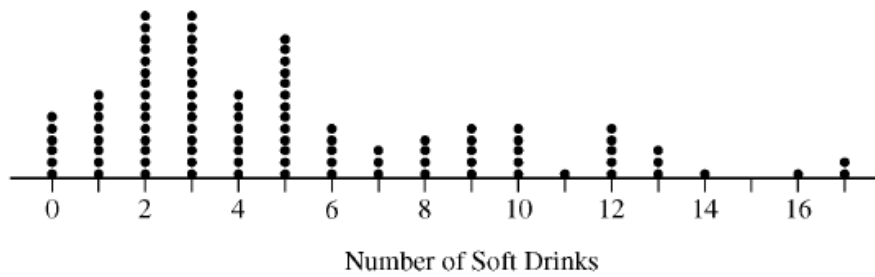
Summary statistics for Rania's data are shown in the table below.

Variable	Gender	N	Mean	Standard Deviation
Number of soft drinks	Female	60	2.90	1.80
	Male	40	7.45	2.22

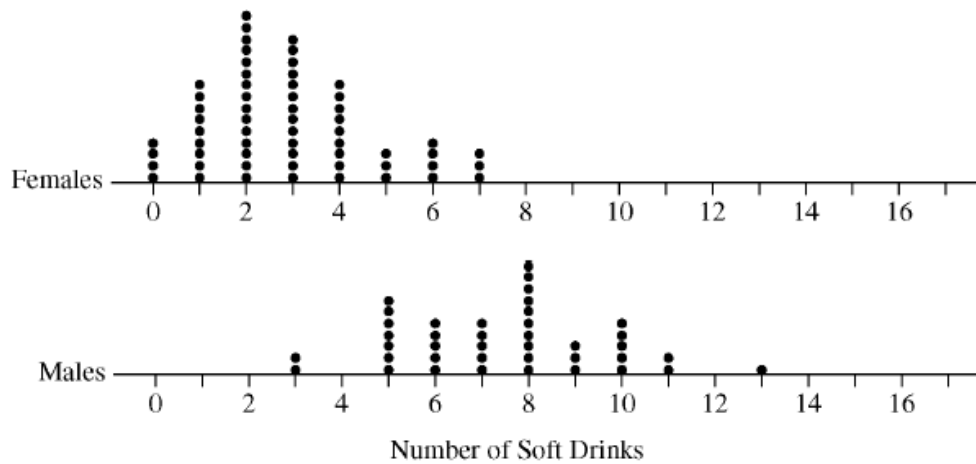
(c) Based on the summary statistics, calculate the estimated standard deviation of the sampling distribution of Rania's point estimator  $\bar{X}_{\text{overall}} = (0.6)\bar{X}_{\text{female}} + (0.4)\bar{X}_{\text{male}}$ .

## 2012 AP<sup>®</sup> Exam #6: The Soda Problem, continued

A dotplot of Peter's sample data is given below.



Comparative dotplots of Rania's sample data are given below.



- (d) Using the dotplots above, explain why Rania's point estimator has a smaller estimated standard deviation than the estimated standard deviation of Peter's point estimator.

## 2013 AP<sup>®</sup> Exam #2: Appearance of University Buildings and Grounds

2. 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?

## **Blocking in Experiments: Matched Pairs and Cholesterol**

A pharmaceutical company has developed a new drug to lower cholesterol and wants to compare it to the drug that is currently being recommended by most doctors. To do this, the company recruits 20 male volunteers, all similar in age. Half of the subjects will be assigned the new drug and the other half will be assigned the current drug.

Method 1: A completely randomized design

Method 2: A matched pairs design

What is the benefit of using blocking in an experiment? What is the best way to choose a blocking variable?

## Blocking in Experiments: Online SAT schools

Many students enroll in prep courses to improve their SAT scores. Twenty students who have taken the SAT once volunteered to participate in an experiment comparing online and classroom prep courses. How can we account for the fact that 10 of these volunteers are in a Precalculus course, 6 are in Algebra 2, and 4 are in Geometry?

Here are the results of the experiment. Make dotplots to compare the improvements of the students in the Online course and the improvements of students in the Classroom course. Based on the dotplots, does there appear to be convincing evidence that the Online course is better?

Class	Treatment	Improvement
P	Online	100
P	Online	100
P	Online	90
P	Online	90
P	Online	100
P	Classroom	70
P	Classroom	70
P	Classroom	80
P	Classroom	80
P	Classroom	80
A	Online	50
A	Online	60
A	Online	40
A	Classroom	30
A	Classroom	40
A	Classroom	20
G	Online	30
G	Online	30
G	Classroom	0
G	Classroom	20

The average improvement for students in Precalculus was  $\bar{x}_P = 86$ , the average improvement for students in Algebra 2 was  $\bar{x}_A = 40$ , and the average improvement for students in Geometry was  $\bar{x}_G = 20$ . How can we use this information to account for the variability created by differences in class level?



## 2007 AP<sup>®</sup> Exam #2: Blocking with Dogs

2. As dogs age, diminished joint and hip health may lead to joint pain and thus reduce a dog's activity level. Such a reduction in activity can lead to other health concerns such as weight gain and lethargy due to lack of exercise. A study is to be conducted to see which of two dietary supplements, glucosamine or chondroitin, is more effective in promoting joint and hip health and reducing the onset of canine osteoarthritis. Researchers will randomly select a total of 300 dogs from ten different large veterinary practices around the country. All of the dogs are more than 6 years old, and their owners have given consent to participate in the study. Changes in joint and hip health will be evaluated after 6 months of treatment.
- (a) What would be an advantage to adding a control group in the design of this study?
  - (b) Assuming a control group is added to the other two groups in the study, explain how you would assign the 300 dogs to these three groups for a completely randomized design.
  - (c) Rather than using a completely randomized design, one group of researchers proposes blocking on clinics, and another group of researchers proposes blocking on breed of dog. How would you decide which one of these two variables to use as a blocking variable?

## What is the Difference Between Stratifying and Blocking?