# Stratified and Cluster Sampling

## Lecture 8
## Sections 2.6, 2.8

Robb T. Koether

Hampden-Sydney College

Tue, Jan 26, 2010

# Outline

# Outline

# Introduction

- Suppose we want to measure support for the recent Senate health-care bill in Massachusetts.

# Introduction

- Suppose we want to measure support for the recent Senate health-care bill in Massachusetts.
- Suppose further that we know that the Massachusetts population is 36% Democrat, 12% Republican, and 52% Independent.

# Introduction

- Suppose we want to measure support for the recent Senate health-care bill in Massachusetts.
- Suppose further that we know that the Massachusetts population is 36% Democrat, 12% Republican, and 52% Independent.
- Suppose even further that we suspect that party affiliation is a relevant variable.

# Introduction

- Suppose we want to measure support for the recent Senate health-care bill in Massachusetts.
- Suppose further that we know that the Massachusetts population is 36% Democrat, 12% Republican, and 52% Independent.
- Suppose even further that we suspect that party affiliation is a relevant variable.
- We plan to take a sample of 100 individuals.

# Introduction

- Suppose we want to measure support for the recent Senate health-care bill in Massachusetts.
- Suppose further that we know that the Massachusetts population is 36% Democrat, 12% Republican, and 52% Independent.
- Suppose even further that we suspect that party affiliation is a relevant variable.
- We plan to take a sample of 100 individuals.
- What might go wrong if we take a simple random sample?

# Introduction

- Suppose we want to measure support for the recent Senate health-care bill in Massachusetts.
- Suppose further that we know that the Massachusetts population is 36% Democrat, 12% Republican, and 52% Independent.
- Suppose even further that we suspect that party affiliation is a relevant variable.
- We plan to take a sample of 100 individuals.
- What might go wrong if we take a simple random sample?
- How can we be assured that each party affiliation will be sufficiently represented in our sample?

# Introduction

- We could choose 36 Democrats, 12 Republicans, and 52 Independents.

# Introduction

- We could choose 36 Democrats, 12 Republicans, and 52 Independents.
- What if we chose to survey 25 Democrats, 25 Republicans, and 50 Independents?

# Introduction

- We could choose 36 Democrats, 12 Republicans, and 52 Independents.
- What if we chose to survey 25 Democrats, 25 Republicans, and 50 Independents?
- Could we accurately estimate the population proportion?

# Outline

# Stratified Random Sample

## Definition (Homogeneous)

A group is homogeneous if its member all have similar characteristics with regard to a variable of interest.

## Definition (Stratum)

A stratum is a homogeneous subset of the population.

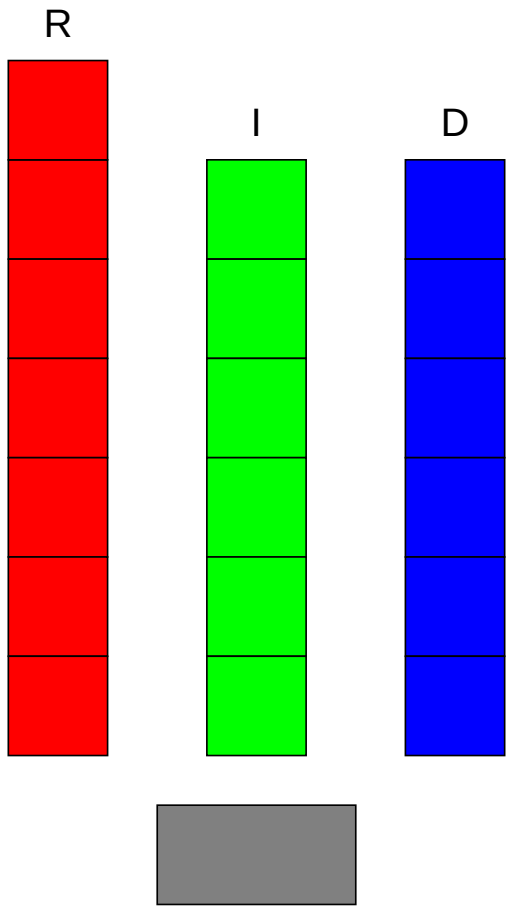
## Definition (Stratified random sampling)

Stratified random sampling is a sampling method in which the population is first divided into strata. Then a simple random sample is taken from each stratum. The combined results constitute the sample.

# Examples

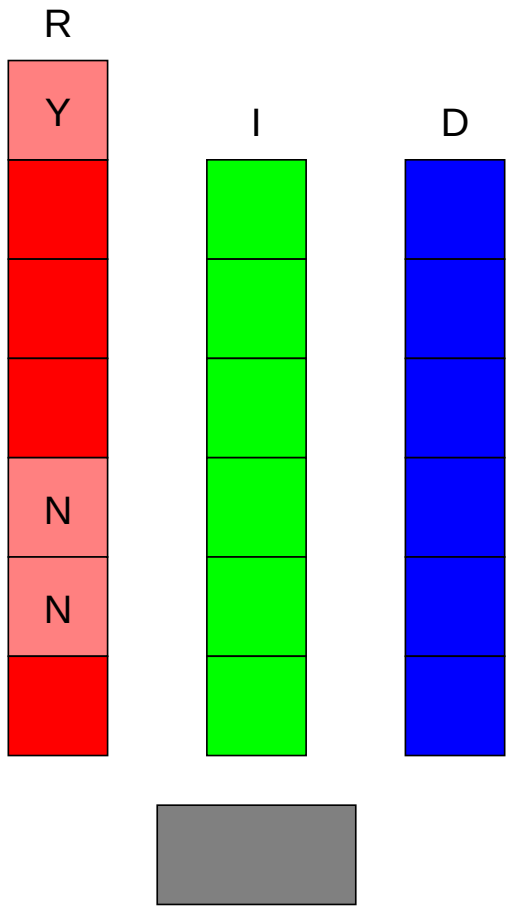
- Possible strata:
  - Male and female strata.
  - Resident and non-resident strata.
  - White, Black, Hispanic, and Asian strata.
  - Protestant, Catholic, Jewish, Muslim, etc., strata.

# Stratified Random Sampling



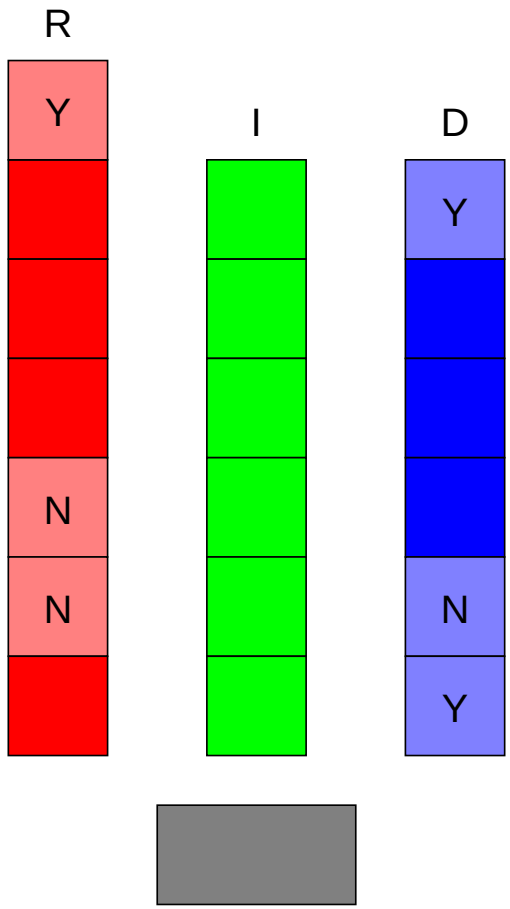

The strata

# Stratified Random Sampling



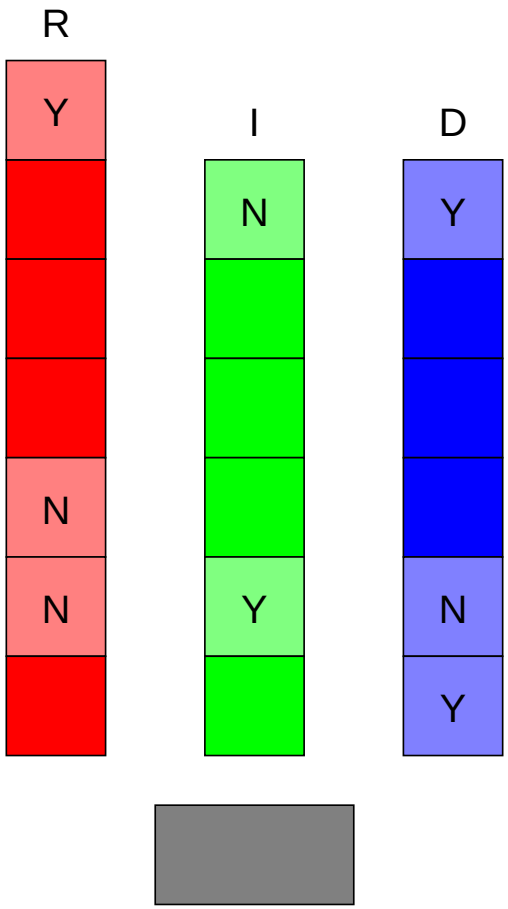

Select 3 Republicans

# Stratified Random Sampling




Select 3 Democrats

# Stratified Random Sampling




Select 2 Independents

# Outline

# Estimating Parameters

- We observed a total of 4 yes's and 4 no's in our sample of 8.
- Our sample proportion is $\frac{4}{8} = 0.50$.
- Can we do better than that?

# Estimating Parameters

- Base on our sample, support is
  - Republicans: $1/3$ of 12%.
  - Democrats: $2/3$ of 36%.
  - Independents: $1/2$ of 52%.

# Estimating Parameters

- We need to compute a *weighted average*.

$$\begin{aligned}\text{wgt. avg.} &= \frac{1}{3}(0.12) + \frac{2}{3}(0.36) + \frac{1}{2}(0.52) \\ &= 0.04 + 0.24 + 0.26 \\ &= 0.54.\end{aligned}$$

# Outline

# Cluster Sampling

## Definition (Heterogeneous)

A group is heterogeneous if is members vary in regard to the variables of interest in the same way that the population varies.

## Definition (Cluster)

A cluster is a heterogeneous subset of the population.

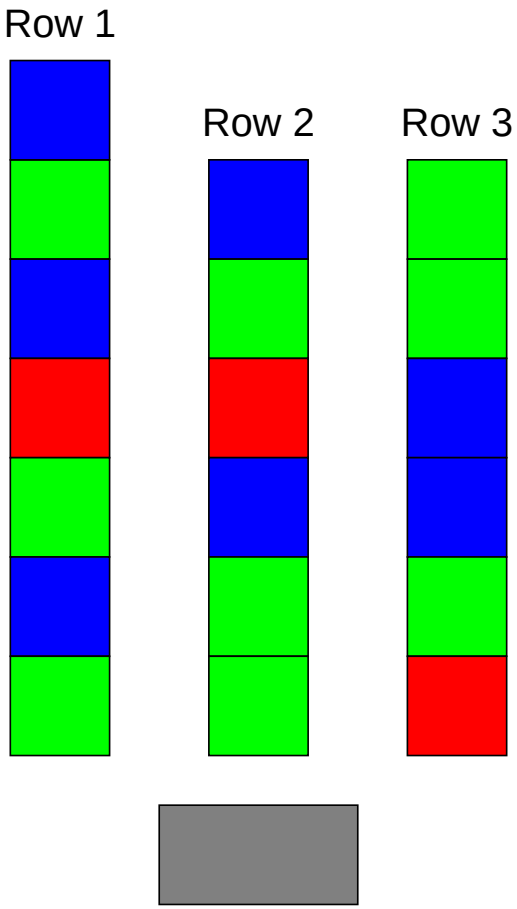
## Definition (Cluster random sampling)

Cluster random sampling is a sampling method in which the population is first divided into clusters. Then a simple random sample of clusters is taken. All the members of the selected clusters together constitute the sample.

# Cluster Sampling

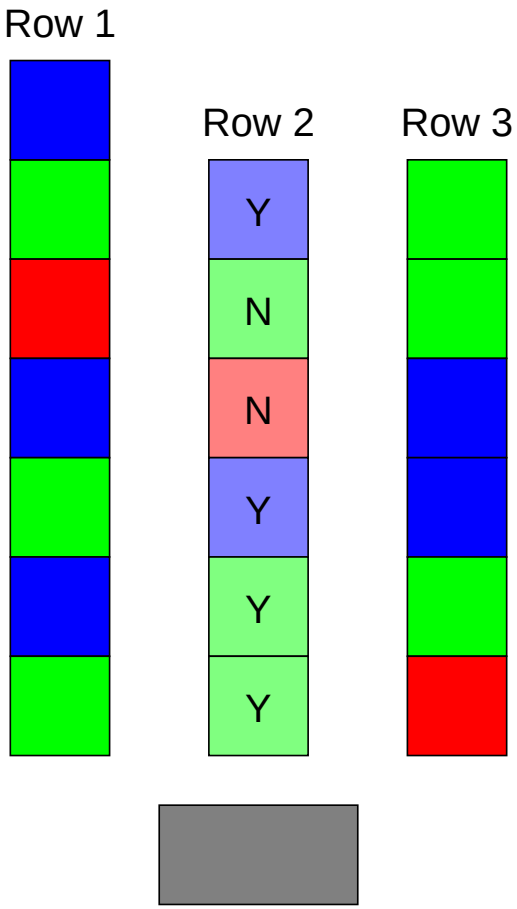
- Note that it is the clusters that are selected at random, not the individuals.
- It is hoped that each cluster by itself is representative of the population, i.e., each cluster is heterogeneous.

# Stratified Random Sampling



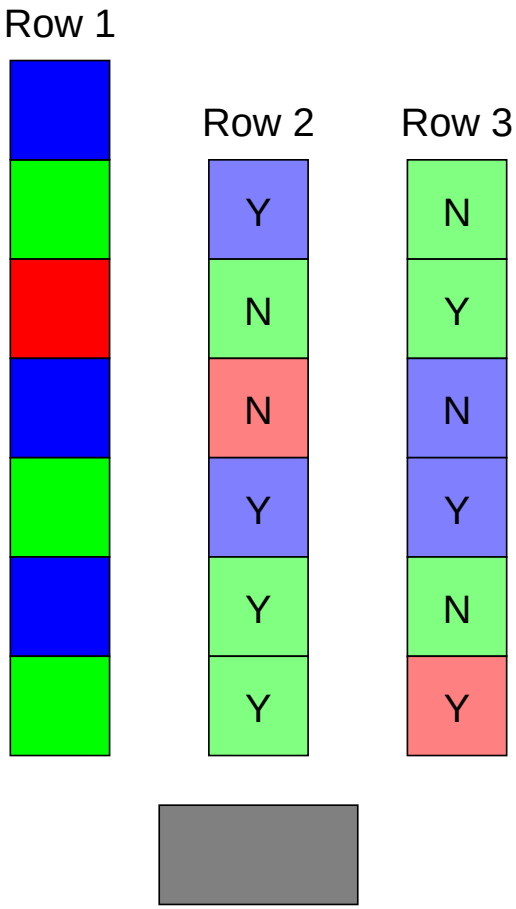

The clusters

# Stratified Random Sampling




Select Row 2

# Stratified Random Sampling




Select Row 3

# Outline

# Stratified Sampling vs. Cluster Sampling

- In stratified sampling
  - From all of the strata we take randomly selected individuals.
- In cluster sampling
  - From randomly selected clusters we take all of the individuals.

# Outline

# Assignment

## Homework

- Read Sections 2.6, 2.8, pages 108 - 115, 122 - 126.
- Let's Do It! 2.6, 2.8.
- Page 115, exercises 19 - 23, 25.
- Page 126, exercises 35 - 38.

# Answers to Even-numbered Problems

## Page 115, Problems 20, 22

2.20 (a) Good. Textbook prices tend to vary dramatically across different majors. A stratified sample will guarantee that all majors are represented.

(b) Not so good. Textbook prices do not vary by the gender of the student, except that females are slightly less likely to be taking science courses, where the textbooks are very expensive. But that takes us back to part (a).

(c) Not so good. Textbook prices may vary somewhat by class rank, but not much.

2.22 (a) $\frac{100}{200}$.

(b) $\frac{100}{1000}$.

(c) 0.10. No. Not all samples are equally likely. Only samples containing 100 males and 20 females are possible.

# Answers to Even-numbered Problems

## Page 126, Problems 36, 38

2.36 (a) Physics and Mathematics.
(b) 60.

2.38 (a) Cluster sampling.
(b) No. The more classes a student takes, the more likely he is to be in the sample.
(c) Poor design together with bad luck. The procedure is not biased, except for the small effect described in part (b). However, the individual classes, especially the education classes, are not heterogeneous, so cluster sampling should not be used.