Math 140
Introductory Statistics

First midterm

September 23 2010

Collection of data?

We know how to explore and look at data
Patterns, displaying, calculating relevant features
But how do we collect data?

Collection of data?

We can design surveys and polls
But since we cannot ask EVERYONE
HOW we select the people we ask will be crucial

If I am designing a poll of registered voters who they will vote for and ask 100 republicans, the answer we get is different than if we ask 100 democrats

We need to randomize

Randomization

Let chance choose the sample.
Pick pollsters at random

If this is done properly we can make predictions beyond the data we have

This is called inference

The facts are coming! The facts are coming!

Pollster:
I asked a hundred likely voters
who they planned to vote for,
and fifty-two of them said
they'd vote for you.

Politician:
Does that mean I'll win?

Pollster:
Sorry, I don't know. We haven't
studied inference in Math140 yet

Politician: What is inference?

Pollster: Drawing conclusions based on your data.
I can tell you about the hundred people I actually talked to,
but I don't yet know how to use that information
to tell you about all the likely voters.

Inference

Methods of inference
can take you beyond the data you actually have,
but only if your data is chosen properly

If you want to use 100 likely voters to tell you about
all likely voters, how you choose those 100 voters is crucial.

The quality of your inference depends on
the quality of your data

Bad data collection leads to bad conclusions.

YOU NEED TO RANDOMIZE
SIMPLE RANDOM SAMPLING

Inference

Population: Set of people or things we want to study.

Unit: Individual element of the population.

Population Size: Number of units.

Sample: set of units that you do get to study.

Census: collecting data on the entire population.

Definitions

Population: Set of people or things we want to study.
Citzens of the United States with voting rights

Unit: Individual element of the population.
A voter

Population Size: Number of units.
Number of voters in the US

Sample: set of units that you do get to study.
Number of people I will poll

Census: collecting data on the entire population.
If I poll everyone, it a census

Definitions

Parameter: some numeric value of a population.
Will they vote republican (0) or democrat (1)

Statistic: Numerical summary of the sample.
Mean, SD etc

Population Size: Number of units.
Number of voters in the US

Examples

In which of these situations do you think a census is used to collect data, and in which do you think sampling is used?

a. An automobile manufacturer inspects its new models.

b. A cookie producer checks the number of chocolate chips per cookie.

c. The U.S. president is determined by an election.

d. Weekly movie attendance figures are released each Sunday.

e. A Los Angeles study does in depth interviews with teachers to find connections between nutrition and health

Examples

D1. a. The population consists of the new models produced by the manufacturer. Census is used. Each new car is inspected. This procedure is used because every customer expects his or her new car to be perfect. Also, if a safety-related problem gets by in even one car, the cost is high.

b. The population is all chocolate chip cookies produced. Sampling is used. Counting the chocolate chips in a cookie is destructive, and counting all the chips in every cookie is time-consuming.

c. The population is the set of voters in a presidential election in the United States. Sampling is used. Theoretically, elections in the United States are a census of voters, but because not everyone who is eligible actually votes, elections are in practice a nonrandom and probably nonrepresentative sample.

d. The population consists of individual movie theater attendance figures. Sampling is used. Not every movie theater owner reports ticket sales every weekend.

e. The population is the set of all teachers in Los Angeles. Sampling is used. With thousands of teachers in the Los Angeles area, it would be too time-consuming and expensive to do in-depth interviews with all of them.

Stratification

Divide the population in subgroups that do not overlap
Take a random sample from each stratum (plural strata)

Example:
polls conducted nation by nation
or gender by gender
or by age groups

Stratified random sampling
Vs.
Simple random sampling

Why stratification?

Convenience: easier to sample countries rather than the whole world.

Coverage: each stratum is properly covered. A company may want to know how it performs in each country.

Precision: estimates may be improved

Example

Building a cement plant near people's homes
Air quality issues

Pollution Standards Index
PSI

PSI > 100 unhealthy

We will pick 10 homes for sampling

100 homes

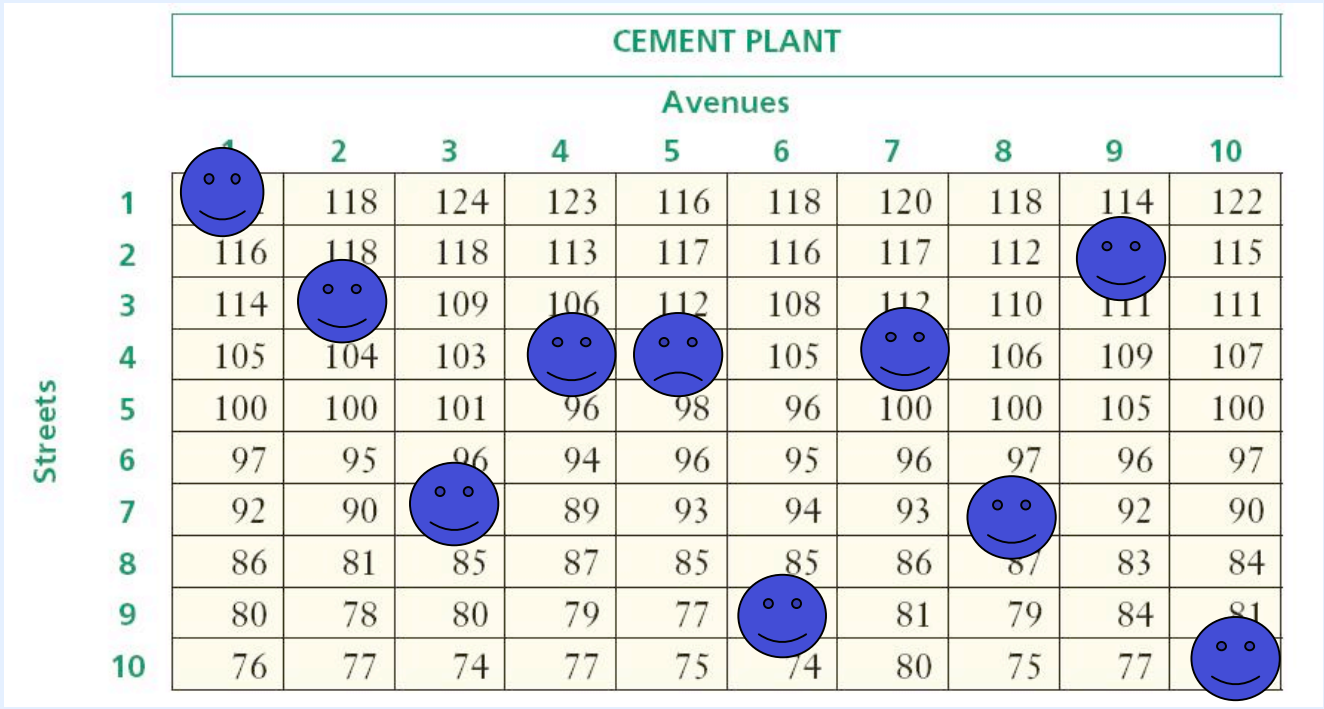
CEMENT PLANT

Streets \ Avenues	1	2	3	4	5	6	7	8	9	10
1	121	118	124	123	116	118	120	118	114	122
2	116	118	118	113	117	116	117	112	112	115
3	114	107	109	106	112	108	112	110	111	111
4	105	104	103	101	103	105	104	106	109	107
5	100	100	101	96	98	96	100	100	105	100
6	97	95	96	94	96	95	96	97	96	97
7	92	90	91	89	93	94	93	92	92	90
8	86	81	85	87	85	85	86	87	83	84
9	80	78	80	79	77	81	81	79	84	81
10	76	77	74	77	75	74	80	75	77	74

Pick 10 homes in three ways

1. Totally at random
2. One per column (may be the same row)
3. One per row (may be the same column)

Case 2.



CEMENT PLANT

Streets \ Avenues	1	2	3	4	5	6	7	8	9	10
1	[illegible]	118	124	123	116	118	120	118	114	122
2	116	118	118	113	117	116	117	112	[illegible]	115
3	114	[illegible]	109	106	112	108	112	110	111	111
4	105	104	103	[illegible]	[illegible]	105	[illegible]	106	109	107
5	100	100	101	96	98	96	100	100	105	100
6	97	95	96	94	96	95	96	97	96	97
7	92	90	[illegible]	89	93	94	93	[illegible]	92	90
8	86	81	85	87	85	85	86	87	83	84
9	80	78	80	79	77	[illegible]	81	79	84	81
10	76	77	74	77	75	74	80	75	77	[illegible]

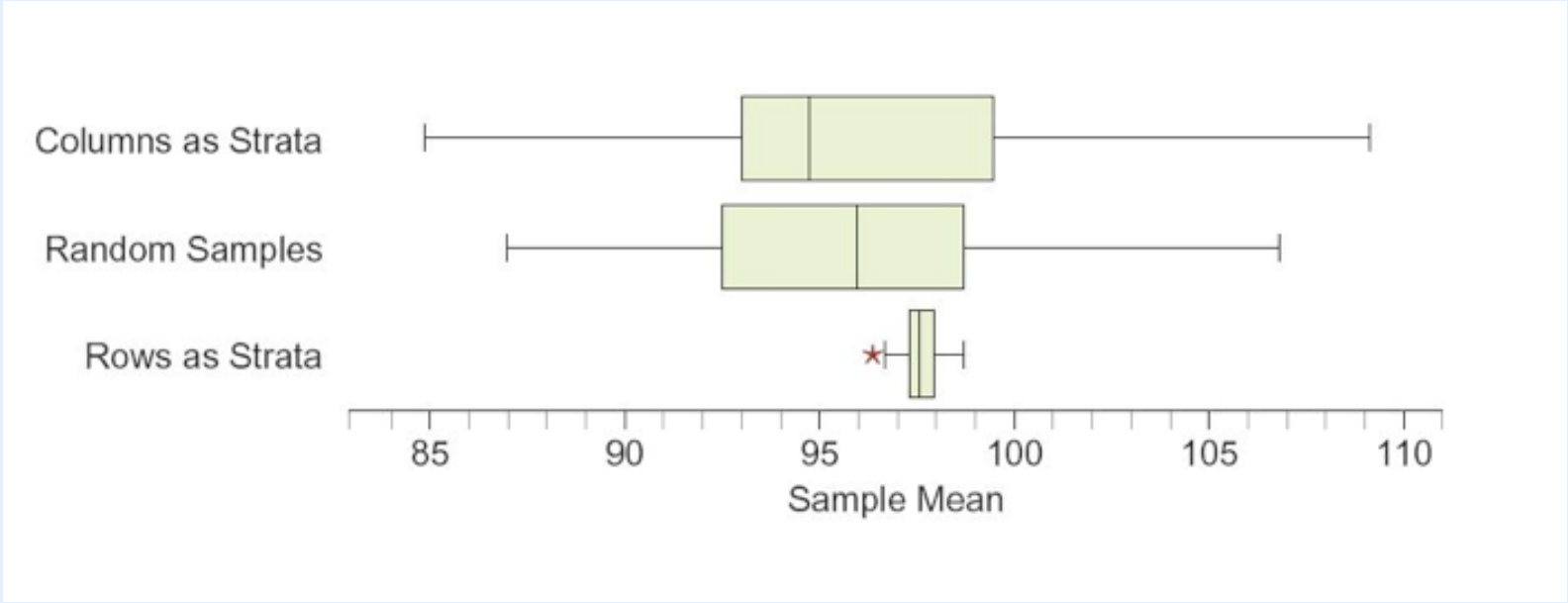
Pick 10 homes

In all three cases calculate the mean
You should end up with three estimates
For cases 1, 2, 3

What is the variability from our measures?

The average for ALL homes is 97.7

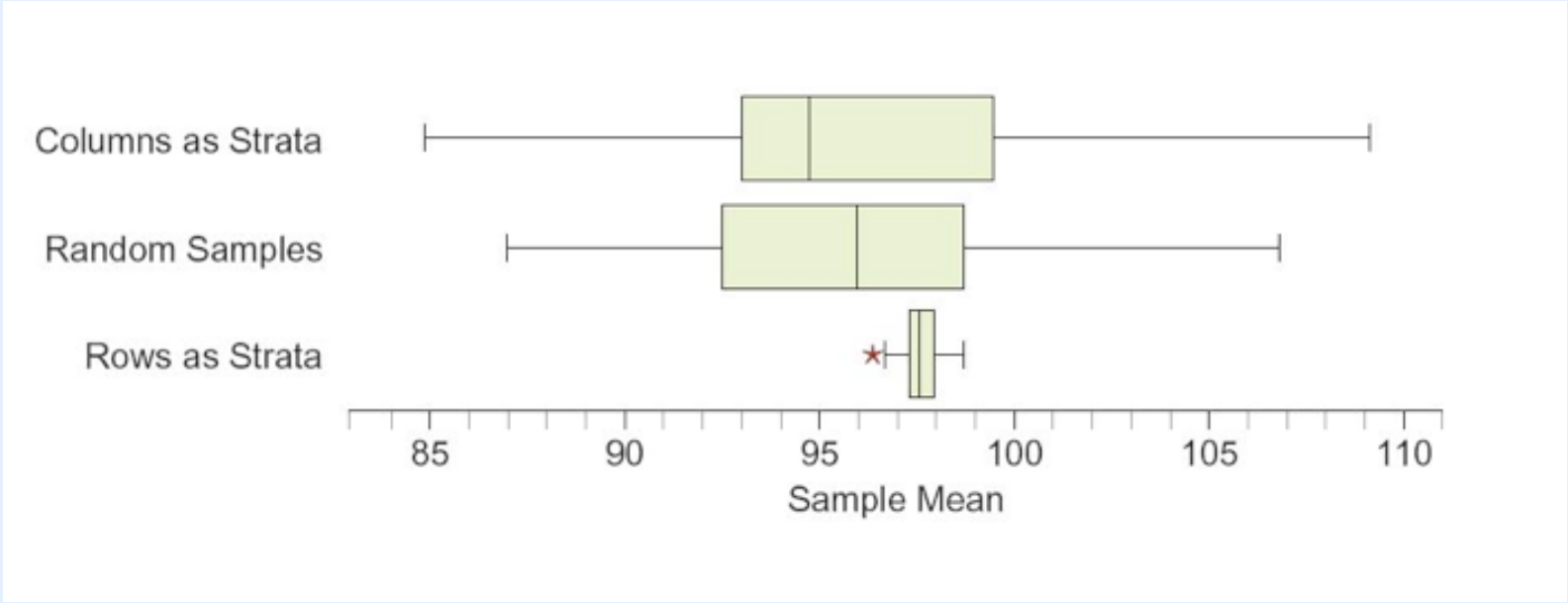
From a hypothetical class of 30



All three centers are close to the actual mean 97.7
What changes is the VARIABILITY

Stratifying by rows is much more likely
to give you values that are not too far
from the actual mean

From a hypothetical class of 30



This happened because data by row is relatively uniform within the row but quite different from row to row

How to pick strata

If strata are of same size, samples should be of same size

Example before: each stratum had 10 houses,
We pick one house from each of them

If we know that in a society 65% of the people are women and 35% are men, then a sample of 100 people stratified on gender should contain 65 women and 35 men

Cluster samples

Sometimes it is difficult to create a proper random sample

How do we go about sampling all the 4th graders in the state of California?

Cluster samples

Sometimes it is difficult to create a proper random sample

How do we go about sampling all the 4th graders in the state of California?

Maybe we can pick some schools and work there? This is an example of a cluster sample

Cluster samples

1. Make a list of all clusters (schools in CA)
2. Take a random sample of clusters
3. Obtain data from all units of each cluster

If you want to be more sophisticated

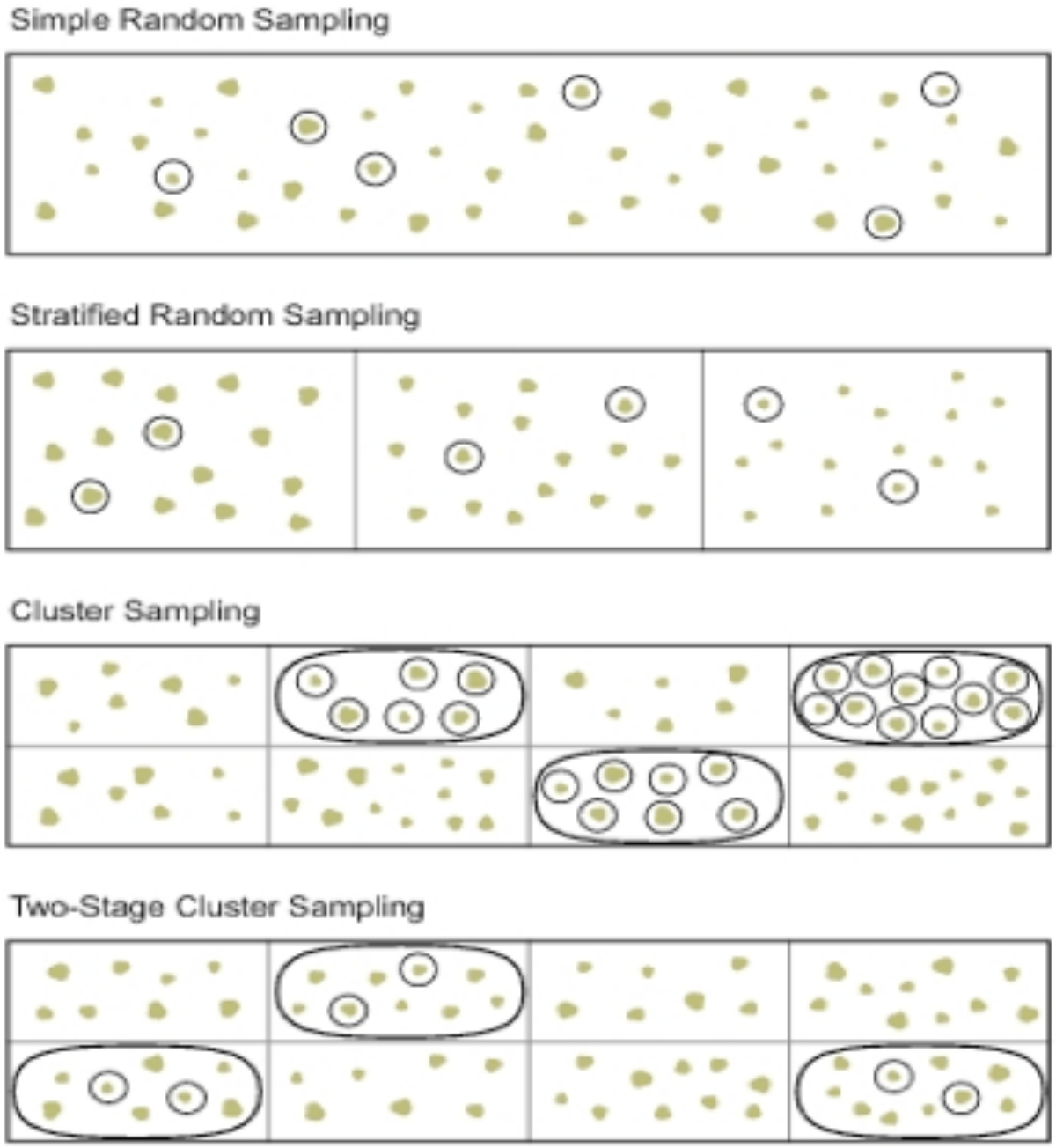
1. Make a list of all clusters (schools in CA)
2. Take a random sample of clusters
3. Obtain data from random samples from each cluster

Simple random sampling

Stratified random sampling

Cluster sampling

2- stage cluster sampling



Stratification vs. Clustering

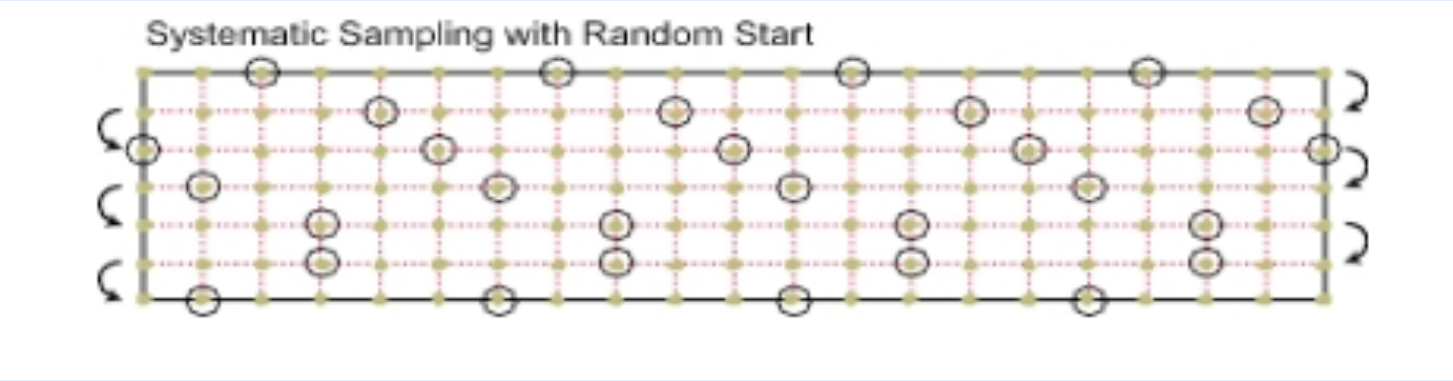
Stratified: you want the units to be relatively uniform
And we want to take samples from each statum.
We pick from all strata

Clusters: you want the clusters to contain as much variation as possible so that it reflects the variation in the total population. We pick only from some clusters

Systematic samples with random start

Let's count off. Count units by a certain value (for example eights) and then pick a number. Everyone associated with that number is your sample.

For example people in a line, or in the telephone book.



In the systematic sample case units need to be well mixed

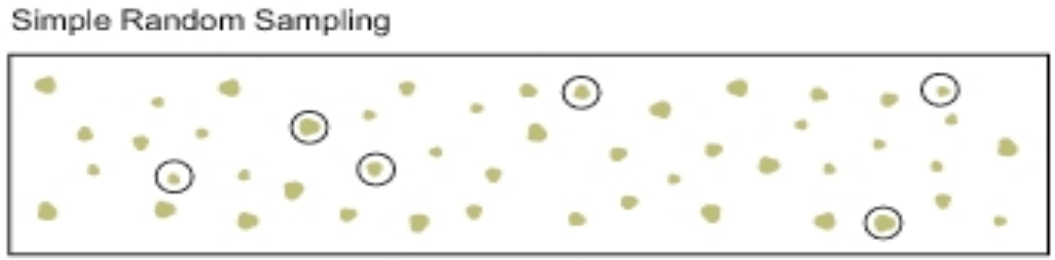
Simple random sampling

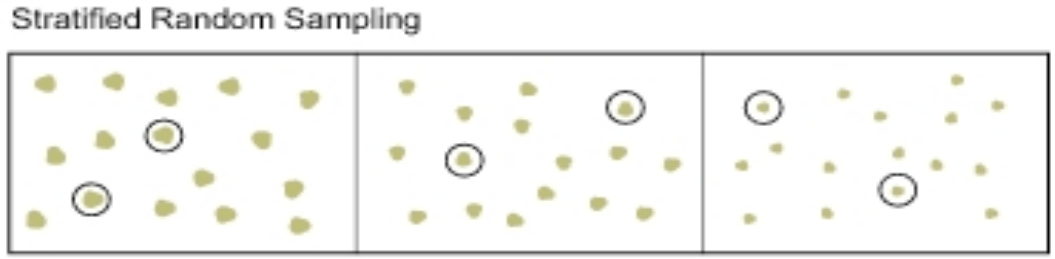
Stratified random sampling

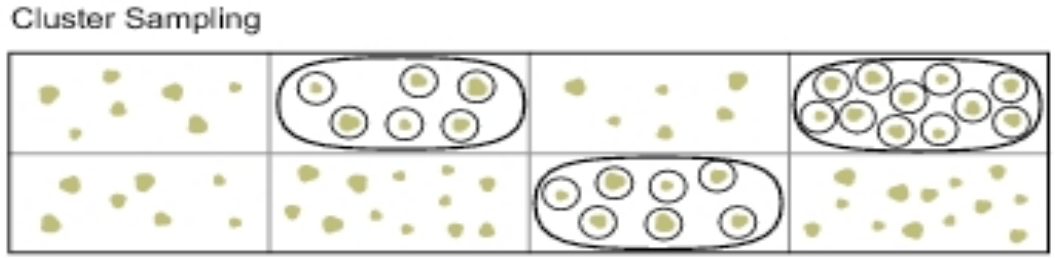
Cluster sampling

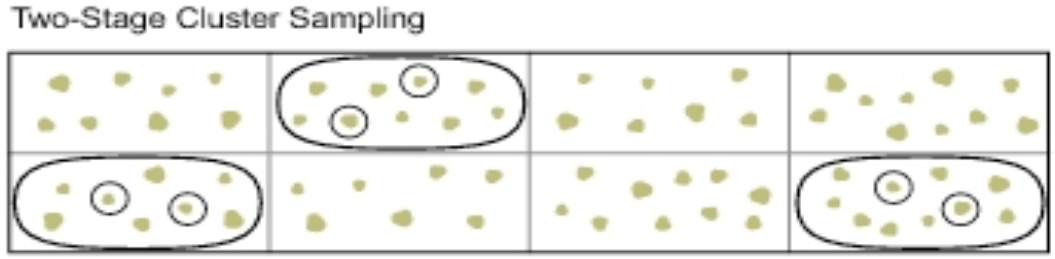
2- stage cluster sampling

Systematic sampling









You try

You are called upon a local movie theater on designing a sampling plan for survey of patrons on their attitudes about recent movies.

About 64% of the patrons are adults
About 30% are teens
About 6% are children

The theater has time and money to interview about 50 patrons.
What design should we use?
How many people from each category?

Bias: problems with survey data

A sampling method is biased
if it tends to give samples
in which some characteristic of the
population is overrepresented or under-represented.

Values are too big or too small
If we are tilting in favor of
one side from the truth

Bias: problems with survey data

A good sample is representative.

That is, it looks like a smaller version of your population. Proportions and number summaries we compute from the sample should be the corresponding ones we'd get if we used the entire population.

How do we check?

Bias: problems with survey data

There is no way of telling if a sample is representative or not. We would need to know what proportions and number summaries are from the population, which we don't know!

So what is the point?

Bias: problems with survey data

If we use the proper sampling method, we can be sure
that we are not biasing our samples
and be confident that our samples are representative

Sample bias

Size bias: Larger units are more likely to be included.

Voluntary Response Bias: Those who care about the issue respond.

Convenience Sample Bias: Units are chosen because of convenience.

Judgment Sample Bias: Units are chosen according to the judgment of someone (expert)

An Unbiased Sample Method requires that all units in the population have a chance of being in the sample.

A Sampling frame is the list of units you use to create the sample

Sample frames

A Sampling frame is the list of units you use to create the sample

Sometimes easy - people in this class
US states
Employees of Westvaco

Sometimes difficult
All the ants in Central Park
All the potato chips produced in the USA

We need a good frame

Identify the type of sampling method used in each of these surveys. Would you expect the estimate of the parameter to be too high or too low?

a. You use your statistics class to estimate the percentage of students in your school who study at least 2 hours a night.

b. You send a survey to all people who have graduated from your school in the past 10 years. You use the mean annual income of those who reply to estimate the mean annual income of all graduates of your school in the past 10 years.

c. A study was designed to estimate how long people live after being diagnosed with dementia. The researchers took a random sample of the people with dementia who were alive on a given day. The date the person had been diagnosed was recorded, and after the person died the date of death was recorded.

a) The statistics class and how much students sleep

D13. a. This is a convenience sample. The estimate could be too high if, for example, the class is an AP Statistics class where students are typically more motivated than the general population of students.

b. This is a volunteer sample, and the estimate would likely be too high. Graduates who feel they have not lived up to the expectations of old classmates might be less likely to respond. In addition, some people may exaggerate a bit in their response.

b) The income of former students

c) Dementia and life-spans

c. This is a random sample of people with dementia alive on a given day. However, size bias will affect the estimate wanted.

Case study

We want to know the percentage of voters who favor Prop X. Our population is the set of people likely to vote in the next election.

We use as frame the phone book listings for people's homes.

How well do you think the frame represents the population?

Are there important groups of individuals who belong to the population but not to the frame?

To the frame but not to the population?

If you think bias is likely, identify what kind of bias and how it might arise.

Land line polls: Obama vs. McCain

		Landline Sample (%)	Cell Phone Sample (%)
Preference	Obama	45	55
	McCain	45	36
	Other	10	9
	Sample size	1960	176

Registered voters, September 9–14, 2008.

		Landline Sample (%)	Cell Phone Sample (%)
Party	Democratic	54	62
	Republican	36	28
	Other	10	10
	Sample size	390	242

Registered voters under age 30, August and September, 2008.

Response bias

Non-Response Bias: You get no data or not enough data. e.g. 80% of people contacted refuse to answer a survey

Questionnaire Bias: Arises from the way the questions are asked.

Example: do you agree or not?

I would be disappointed if Congress cut its funding for public television.

Response bias

Non-Response Bias: You get no data or not enough data. e.g. 80% of people contacted refuse to answer a survey

Questionnaire Bias: Arises from the way the questions are asked.

Example: do you agree or not?

Cuts in funding for public television are justified as part of an overall effort to reduce federal spending.

Response bias

The two questions are basically asking the same thing
Agreeing with first statement = Disagreeing with second one

Yet:

First statement: 54% agree, 40% disagree, 6% don't know

Second statement: 52% agree, 37% disagree, 10% don't know

Hk

Page 183 E3, E5, E6, E9, E13