Lecture 3

Sampling

Thought Question 1:

What do you think is the **major difference between** a ***survey*** (such as a public opinion poll) and an ***experiment*** (such as the heartbeat experiment in Case Study 1.1)?

Thought Question 2:

Suppose a properly chosen **sample of 1600 people** across the United States was asked if they **regularly watch** a certain television program, and **24% said yes**.

How close do you think that is to the percentage of the entire country who watch the show? **Within 30%? 10%? 5%? 1%? Exactly the same?**

Thought Question 3:

Many television stations conduct polls by asking viewers to call one phone number if they feel one way about an issue and a different phone number if they feel the opposite.

Do you think the results of such a poll represent the feelings of the community?

Do you think they represent the feelings of all those watching the TV station at the time or the feelings of some other group? Explain.

Thought Question 4:

Suppose you had a telephone directory listing all the businesses in a city, alphabetized by type of business.

If you wanted to phone 100 of them to get a representative sampling of opinion on some issue, **how would you select which 100 to phone**?

Why would it **not be a good idea to simply use the first 100 businesses listed**?

4.1 Common Research Strategies

Sample Surveys

- A subgroup of a large population is questioned on a set of topics.
- No intervention or manipulation of the respondents, simply asked to answer some questions.

Randomized Experiments

- Measures the effect of manipulating the environment in some way.
- **Randomized experiment** = manipulation is assigned to participants on a random basis.
- **Explanatory variable** = the feature being manipulated.
- **Response variable** = outcome of interest.
- **Randomization** helps to make the groups *approximately equal* in all respects except for the explanatory variable.

Observational Studies

- Manipulation occurs **naturally**, not imposed.
- Can’t assume the **explanatory variable** is the only one responsible for any observed differences in the **response variable**.
- **Case-control study** attempts to include an appropriate control group.
- Sometimes results more readily extend to the real world than in an experiment – no artificial manipulation.

Meta-Analyses

- Quantitative review of a **collection of studies** all done on a **similar topic**.
- Combining information can lead to emergence of patterns or effects not readily seen in the individual studies.
- More on meta-analyses in Chapter 25.

Case Studies

- In-depth examination of one or a small number of individuals.
- Descriptive and do not require statistical methods.
- Generally can’t be extended to any person or situation other than the one studied.

Example 1: Two Studies That Compared Diets

News Story 20: "Eating organic foods reduces pesticide concentration in children."

- **Observational Study**: parents kept a food diary for children for 3 days. Result: there was a difference (for children eating organic versus conventional produce).
- **Plus**: subjects followed their natural diet
- **Minus**: can't determine if difference due to food choice or other factors.

News Story 3: "Rigorous veggie diet found to slash cholesterol."

- **Experiment**: volunteers randomly assigned to 1 of 3 diets. Reduction in cholesterol measured and compared.
- **Plus**: other variables that affect cholesterol should be similar across the groups.
- **Minus**: would anyone really follow the diet in real world?

4.2 Defining a Common Language

- A **unit** is a single individual or object to be measured.
- The **population** (or **universe**) is the entire collection of units about which we would like information or the entire collection of measurements we would have if we could measure the whole population.

- The **sample** is the collection of units we actually measure or the collection of measurements we actually obtain.
- The **sampling frame** is a list of units from which the sample is chosen. Ideally, it includes the whole population.
- In a **sample survey**, measurements are taken on a subset, or sample, of units from the population.
- A **census** is a survey in which the entire population is measured.

Example 2: Determining Monthly Unemployment in the U.S.

Background:

Bureau of Labor Statistics visits approx 60,000 households, chosen from list of all known households in country. Each adult classified: employed, unemployed, “not in labor force.” Unemployment rate: number of unemployed persons divided by the sum of the employed and unemployed.

- **Units**: adults in the labor force
- **Population** of *units*: all adults in the labor force. **Population** of *measurements*: employment status (working or not working) of everyone in the labor force.
- **Sampling frame**: list of all known households in country.
- **Sample** of *units*: people who were asked about their employment status.
- **Sample** of *measurements*: employment status of sample.

4.3 The Beauty of Sampling

With proper sampling methods, based on a sample of 1500 adults we can almost certainly estimate, to within 3%, the percentage of the entire population who have a certain trait or opinion.

This result does *not* depend on
how large the (large) population is.

Why is Sampling used

- Resources are needed to conduct a Census
- CSO Spends about £20million to conduct the 5 year Census of Population
- Sometimes the measuring process destroys the thing being measured, e.g. if we were to test the strength of a weld or in testing an individuals blood - who among us would be willing to donate all of our blood in a test?
- Because of the work involved in a Census it is much faster to conduct a survey, sometimes it is important to have results fast.

Accuracy of a Sample Survey: Margin of Error

For a properly conducted sample survey:

The sample proportion differs from the population proportion by more than the **margin of error** less than 5% of the time, or in fewer than 1 in 20 surveys.

$$\boxed{\textbf{Margin of error} \cong {}^{1}\!/\!_{\sqrt{n}}}$$

Margin of Error $\cong \frac{1}{\sqrt{n}}$

With a sample of n = 1600, we usually get an estimate that is accurate to within 2.5% of the truth:

$$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1600}} = \frac{1}{40} = 0.025$$

"55% of respondents support the president's economic plan. The margin of error for this survey is plus or minus 2.5 percentage points."

⇨ Almost certain that between 52.5% and 57.5% of the entire population support the plan.

Such intervals miss covering truth about 1 in 20 times.

Example 3: Measuring Teen Drug Use

News Story 13: "3 factors key for drug use in kids."
QEV Analytics surveyed ***1,987 children*** *ages 12 to 17 and* ***504 parents*** *…*
The margin of error was plus or minus ***two*** *percentage points for* ***children*** *and plus or minus* ***four*** *percentage points for* ***parents****.*

For children: $1/\sqrt{n} = 1/\sqrt{1987} = 0.0224$ or about 2.2%.

For parents: $1/\sqrt{n} = 1/\sqrt{504} = 0.0445$ or about 4.45%.

Study result: 20% of the children said they could buy marijuana in an hour or less.

We can be fairly confident that somewhere between **18% and 22% of *all* teens in the *population*** represented by those in this survey would respond that way if asked.

Example: A poll was conducted to see what proportion of Americans think terrorists in this country are prepared to launch a major attack.

1009 people adults were included in the survey.

904 people believe that terrorists are prepared to launch a major attack.

What is $\hat{p}$, the estimate of ALL Americans that believe terrorists are prepared?

$$\hat{p} = 904/1009 = 0.8958 \approx 0.90 \text{ or } 90\%$$

What is the margin of error for this study?

$$\text{ME} = 1/\sqrt{n} = 1/\sqrt{1009} = 0.031 \approx 0.03 \text{ or } 3\%$$

This means that it is almost certain that between 87% and 93% of American population believe that terrorists are prepared to launch a major attack.

To obtain a confidence interval: Add and subtract the ME from the sample value, and the resulting interval almost surely covers the true population value.

Other Advantages of Sample Surveys

- When measurements destroy the units being tested, a **census is not feasible**.
- **Faster to collect a sample** than a census if the population is large.
- Can devote resources to getting the **most accurate information** possible from the sample.

4.4 Simple Random Sampling

Probability sampling plans:
everyone in the population has a specified chance of making it into the sample.

Simple Random Sample: every conceivable group of people of the required size has the *same chance* of being the selected sample.

Need: (1) **List of units** in the population.

(2) Source of **random numbers**.

Example 4: How to Sample From Your Class

Background:

Population = 200 students in your class.
Plan to take a simple random sample of size *n* = 25 students.

Margin of error: $1 / \sqrt{25} = 1/5 = 0.20$ or about 20%.

Step 1: Obtain a **list of students** in the class, numbered 1 to 200.

Step 2: Obtain **25 random numbers** between 1 and 200.
Write 1 to 200 on same-sized slips of paper,
put in bag, mix well, draw out 25; or
Use computer or calculator program (e.g. Minitab).

Step 3: **Locate and interview the people** on your list whose numbers were **selected**.

4.5 Other Sampling Methods

- Stratified Random Sampling
- Cluster Sampling
- Systematic Sampling
- Random Digit Dialing
- Multistage Sampling

Stratified Random Sampling

Divide population into groups (strata) and take a simple random sample from each.

Advantages:

1. Have **individual estimates** for each stratum.
2. If variable measured gives more consistent values within each strata than within population, can get **more accurate estimates** of population values.
3. If strata are geographically separated, may be **cheaper** to sample them separately.
4. May use different interviewers within each strata.

Cluster Sampling

Divide population into groups (clusters), take a random sample of clusters and measure only the selected clusters.

Advantage: need only a list of clusters,
not a list of all individual units.

Example: Sample students living in a dorm at a college

College has 30 dorms, each dorm has 6 floors
⇨ 180 floors form the clusters.

Take a random sample of floors
and measure everyone on those floors.

Systematic Sampling

Divide population list into as many consecutive segments as you need, randomly choose a starting point in the first segment, then sample at that same point in each segment.

Example:

- List of 5000 names, want a sample of 100.
- Divide the list into 100 consecutive segments of 50.
- Randomly choose a starting point in the first 50 names.
- Then sample every 50th name after that.

Random Digit Dialing

Results in a sample that approximates a simple random sample of all households in the U.S. that have telephones.

Steps:

- List all possible ***exchanges*** (area code + next 3 digits).
- Use white pages to approximate proportion of all households in country that have each exchange.
- Use computer to generate a sample with approximately the same proportions.
- Repeat above to sample ***banks*** (next 2 digits)
- Computer randomly generates last two digits to complete the phone number.

Example 5: Finding Teens and Parents Willing to Talk

News Story 13: "3 factors key for drug use in kids."
Researchers started with an 'initial pool of random telephone numbers' consisting of 94,184 numbers, which 'represented all 48 continental states in proportion to their population, and were prescreened by computer to eliminate as many unassigned or nonresidential telephone numbers as possible.'

More Details about the 94,184 numbers:

Resulted in only 1,987 completed interviews …

- 12,985 were not in service
- 25,471 ineligible since no resident in required age group
- 27,931 refused to provide information
- 8,597 abandoned because of no answer (after 4+ call backs)

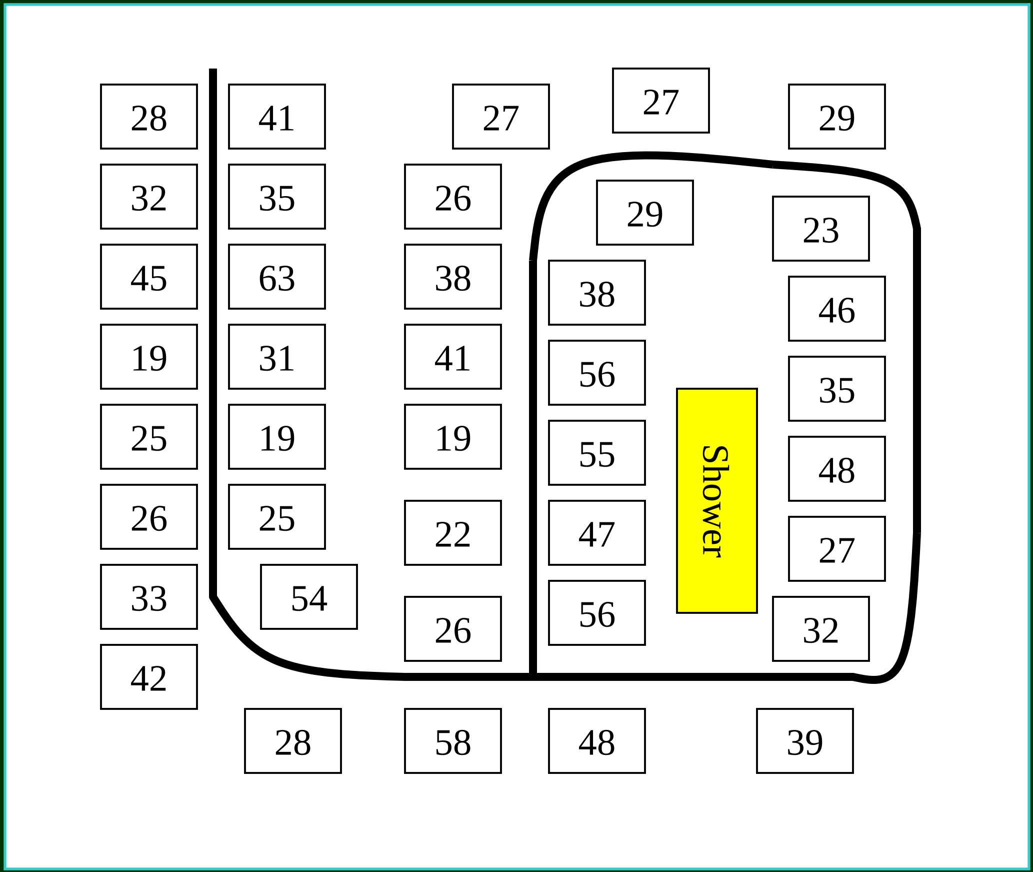
Multistage Sampling

Sampling plan that uses a combination of sampling methods in various stages.

Example:

- Stratify by region of the country; then
- stratify by urban, suburban, and rural; then
- choose a random sample of communities within those strata.
- Divide those communities into city blocks or fixed areas, as clusters, and sample some of those.
- Everyone on the block or within fixed area may then be sampled.

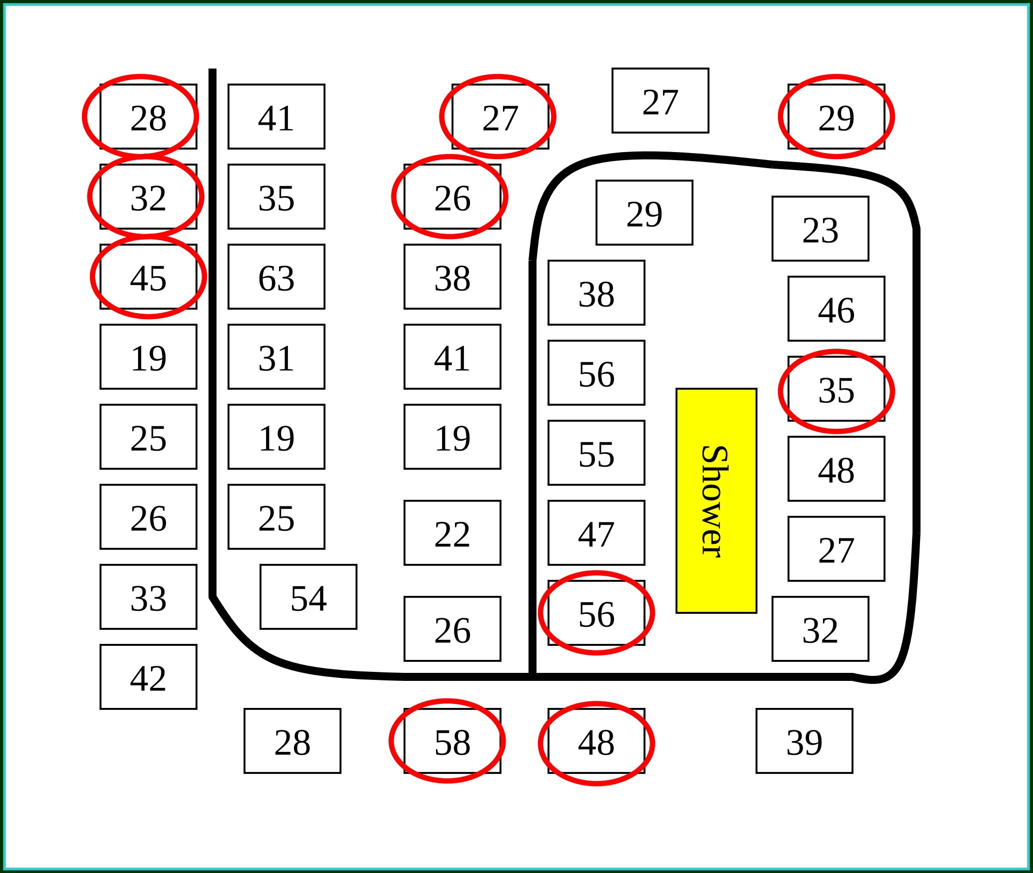
Class Example: Estimate Average Age of Campers:



Complete enumeration:

Average age = 35.95 years

Class Example: Estimate Average Age of Campers:

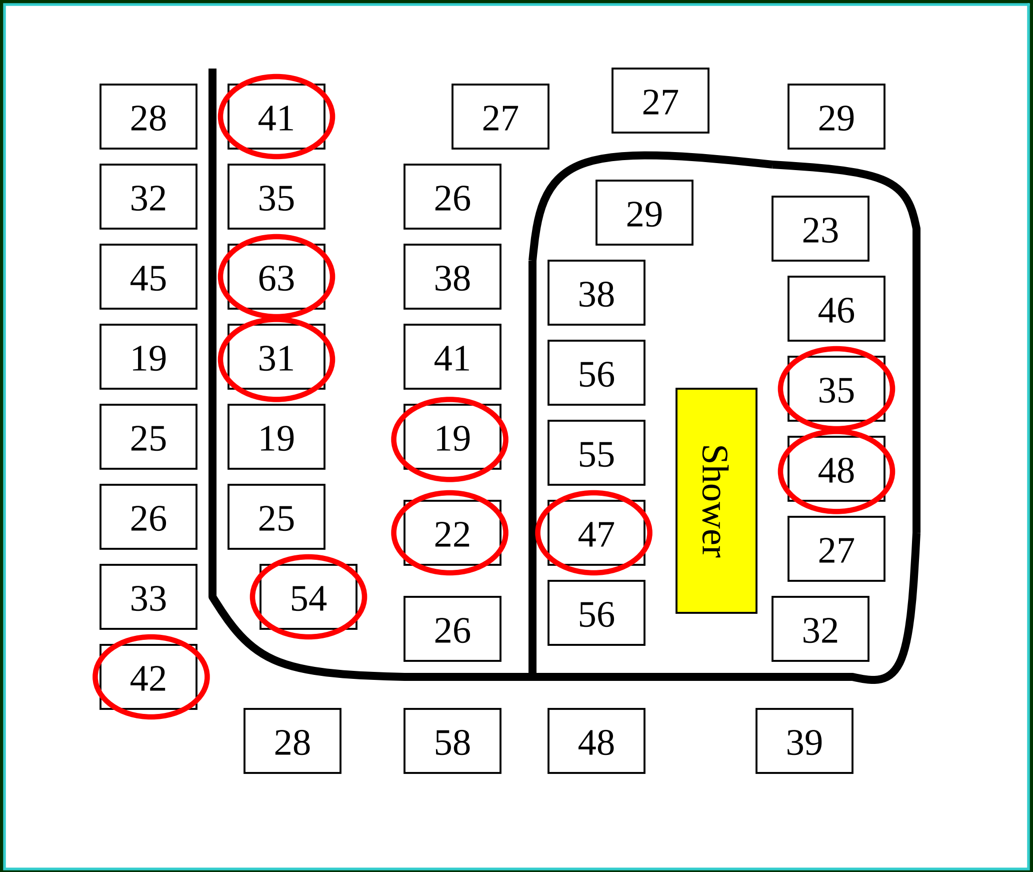


Simple Random Sampling:
n = 10

Average age = 38.4 ± 7.6 years

True average = 35.95

Class Example: Estimate Average Age of Campers:

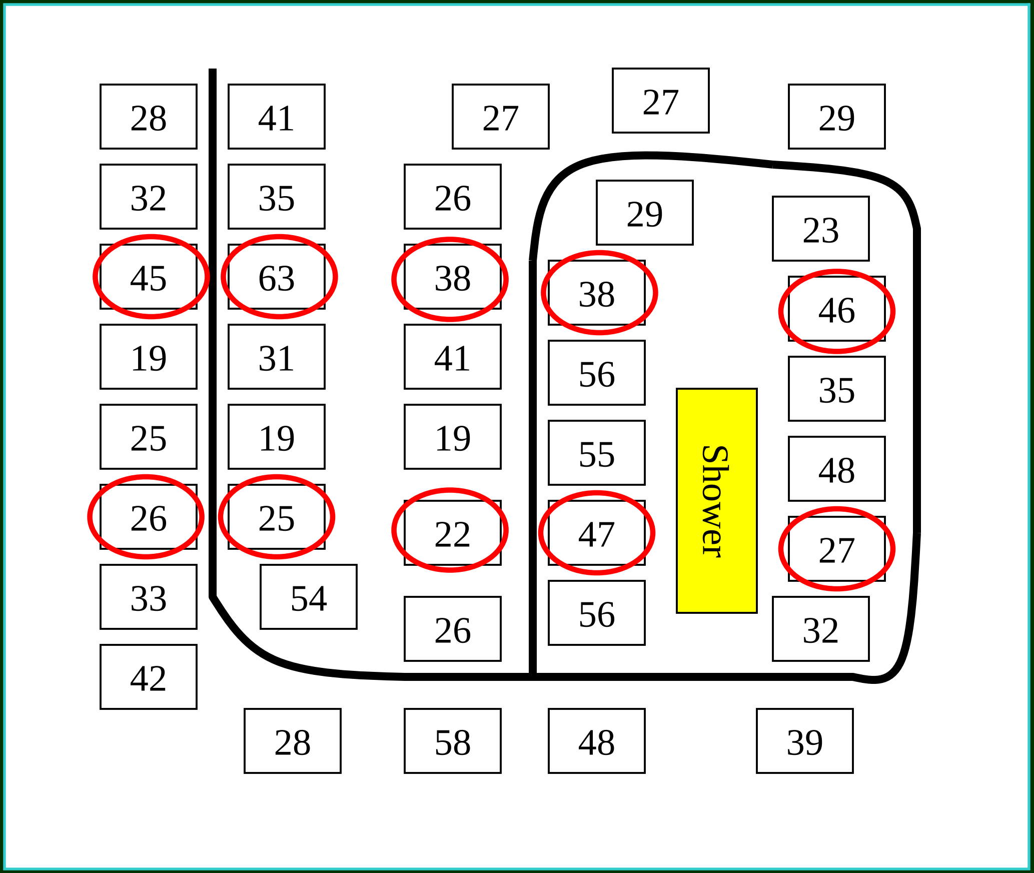


Simple Random Sampling:
n = 10

Average age =
40.2 ± 8.5 years

True average = 35.95

Class Example: Estimate Average Age of Campers:

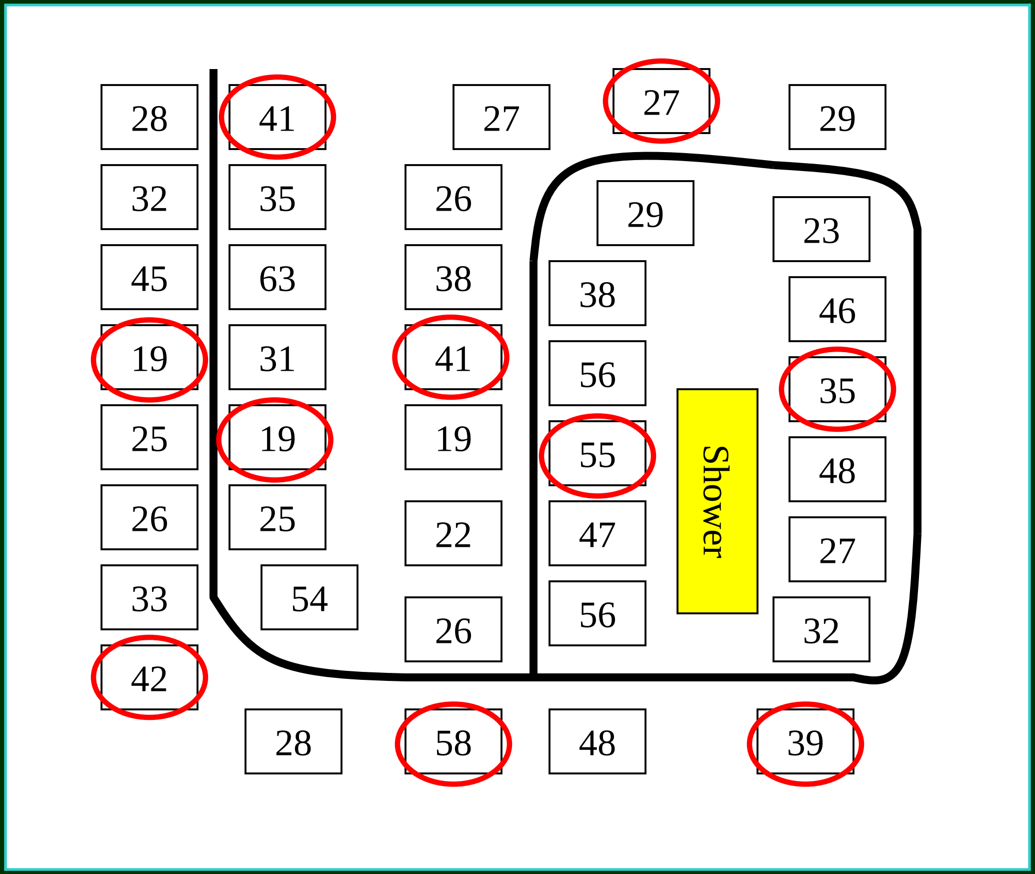


Systematic Sampling:
n = 10

Average age =
37.7 ± 8.0 years

True average = 35.95

Class Example: Estimate Average Age of Campers:

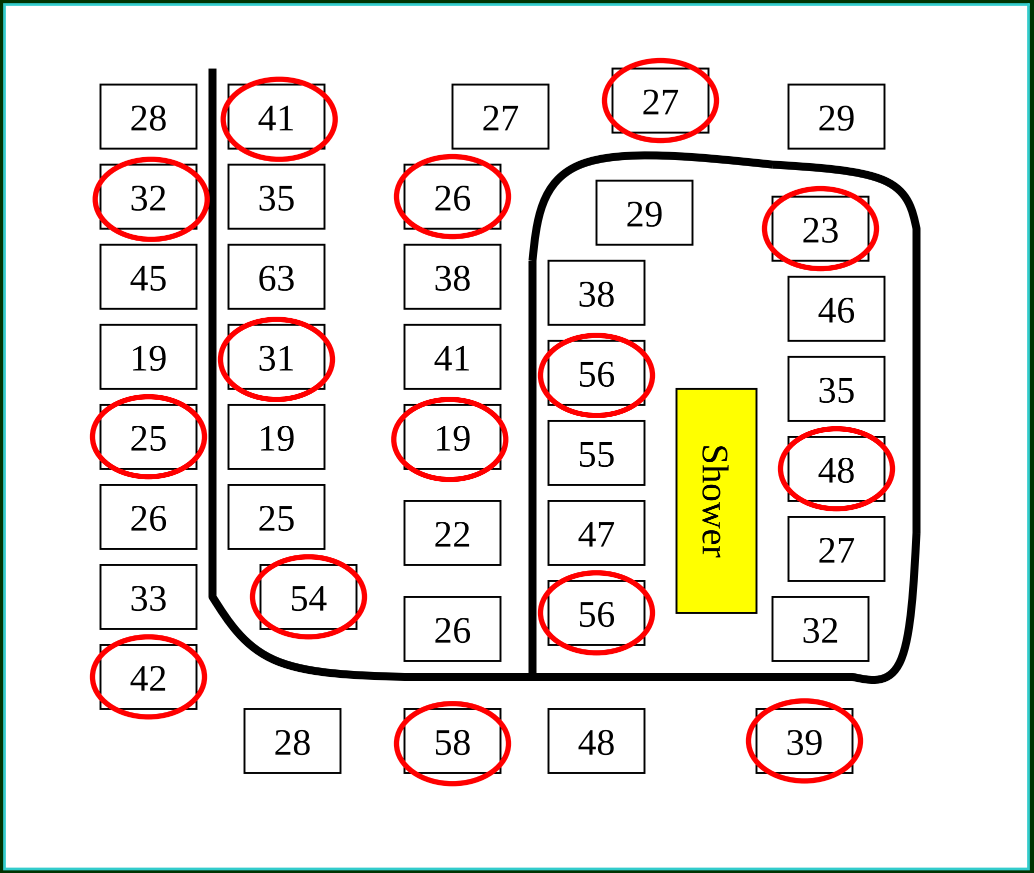


Systematic Sampling:
n = 10

Average age = 37.6 ± 8.2 years

True average = 35.95

Class Example: Estimate Average Age of Campers:



Systematic Sampling:
n = 15

Average age =
38.5 ± 5.9 years

True average = 35.95

4.6 Difficulties and Disasters in Sampling

Difficulties

1. Using the wrong sampling frame
2. Not reaching the individuals selected
3. Having a low response rate

Disasters

4. Getting a volunteer or self-selected sample
5. Using a convenience or haphazard sample

Difficulties in Sampling

Using the Wrong Sampling Frame

Sometimes sampling frame either will include unwanted units or exclude desired units.

Examples

- Using list of registered voters to predict election results.
  Problem: includes those who are not likely to vote.
  Solution: first ask questions about voting history.
- Using telephone directory to survey general population.
  Problem: excludes those who move often, those with unlisted home numbers, those without a phone.
  Solution: use random digit dialing.

Difficulties in Sampling

Not Reaching the Individuals Selected

Even if a proper sample of units is selected, the units may not be reached.

Examples

- Telephone surveys reach a disproportionate number of women because they are more likely to answer phone. *Solution*: ask to speak with oldest adult male at home.
- Beware of ‘quickie polls’ – “most likely to be wrong because questions are hastily drawn and poorly pretested, and it is almost impossible to get a random sample in one night” (Crossen, 1994, p. 102)

Difficulties in Sampling

Having a Low Response Rate

Even the best surveys are not able to contact everyone on their list, and not everyone contacted will respond.

- Response rates should be reported in research summaries.
- The lower response rate => the less results can be generalized to the population.
- Survey response is voluntary, those who respond likely to have stronger opinions than those who do not.
- Use or reminders, follow up calls can decrease nonresponse rate.

Disasters in Sampling

Getting a Volunteer or Self-Selected Sample

Relying on volunteer *responses* presents difficulties, but relying on a volunteer *sample* is a waste of time.

Example 7: A Meaningless Poll

Television poll vs. properly conducted study

Q: Do you support the president's economic plan?	**TV Poll (volunteer sample)**	**Survey (random sample)**
Yes	42%	75%
No	58%	18%
Not sure	0%	7%

Such ‘volunteer polls’ are just a count of who bothered to go to the telephone and call.

Voluntary response example

- Advice columnists Ann Landers asked her readers, "If you had it to do over again, would you have children?“ Almost 10,000 readers wrote in, and 70% of them said No.
- A statistically design opinion poll a few months later found that 91% of parents studied said Yes!
- This is a self-selected sample, worthless as indicators of opinion of the population.
- Biased because people with strong opinions (especially negative) tend to respond.
- Similar example: TV shows call-in polls, internet polls, etc.

Disasters in Sampling

Using a Convenience or Haphazard Sample

Using the most convenient group available or deciding on the spot who to sample can produce misleading results.

Example:

Students in Introductory Statistics Classes
may be representative of all students at a university
on extent of drug use in the HS they attended,
but not on how many hours they study each week.

Case Study 4.1: *The Infamous Literary Digest Poll of 1936*

Election of 1936: Democratic incumbent Franklin D. Roosevelt and Republican Alf Landon

Literary Digest Poll:

- Sent questionnaires to 10 million people from magazine subscriber lists, phone directories, car owners, who were more likely wealthy and unhappy with Roosevelt.
- Only 2.3 million responses for 23% response rate. Those with strong feelings, the Landon supporters wanting a change, were more likely to respond.
- (Incorrectly) Predicted a 3-to-2 victory for Landon.

Case Study 4.1: *The Infamous Literary Digest Poll of 1936*

Election of 1936: Democratic incumbent Franklin D. Roosevelt and Republican Alf Landon

Gallup Poll:

- George Gallup just founded the American Institute of Public Opinion in 1935.
- Surveyed a random sample of 50,000 people from list of registered voters. Also took a random sample of 3000 people from the *Digest* lists.
- (Correctly) Predicted Roosevelt the winner. Also predicted the (wrong) results of the *Literary Digest* poll within 1%.