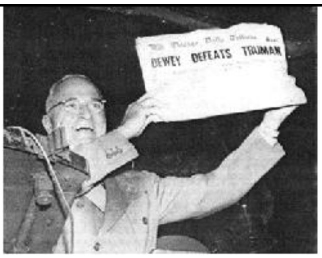
# Ch. 12: Sample Surveys

AN AVERAGE TEACHER MIGHT ASSIGN THREE PAGES OF HOMEWORK.
A MEAN TEACHER WOULD ASSIGN TEN PAGES.
OK
YOU SAID "MEAN" AND "AVERAGE" WERE THE SAME THING.
MATHEMATICALLY.
I JUST DON'T WANT TO BELIEVE THE AVERAGE PERSON IS MEAN.

Copyright © 2003 United Feature Syndicate, Inc.

***"If you don't believe in random sampling, the next time you have a blood test tell the doctor to take it all."***



**The election of 1948**
**The Predictions**

| The Candidates | Crossley | Gallup | Roper | **The Results** |
|---|---|---|---|---|
| Truman | 45 | 44 | 38 | **50** |
| Dewey | 50 | 50 | 53 | **45** |

## Objectives

### Producing data: sampling

- Observation versus experiment
- Population versus sample
- Sampling methods
- Simple random samples
- Stratified samples
- Caution about sampling surveys
- Learning about populations from samples

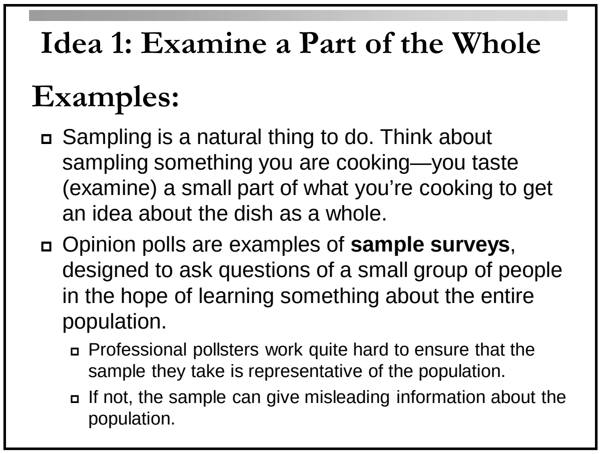
## Background

- We have learned ways to display, describe, and summarize data, but have been limited to examining the particular batch of data we have.
- To make decisions, we need to go beyond the data at hand and to the world at large.
- Let's investigate three major ideas that will allow us to make this stretch…

## 3 Key Ideas That Enable Us to Make the Stretch

## Idea 1: Examine a Part of the Whole

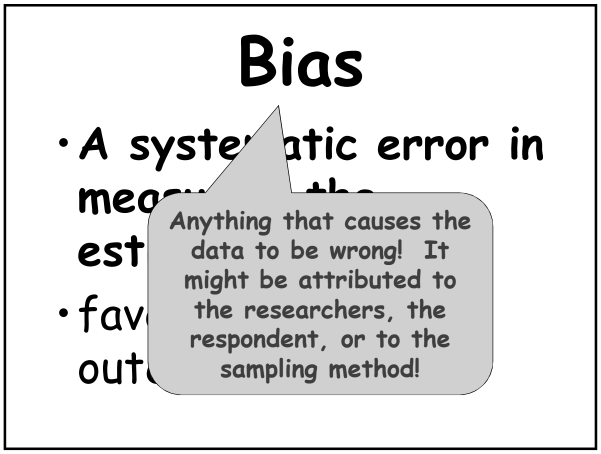
- The first idea is to draw a **sample**.
  - We'd like to know about an entire **population** of individuals, but examining all of them is usually impractical, if not impossible.
  - We settle for examining a smaller group of individuals—a **sample**—selected from the population.



## Idea 1: Examine a Part of the Whole

### Examples:

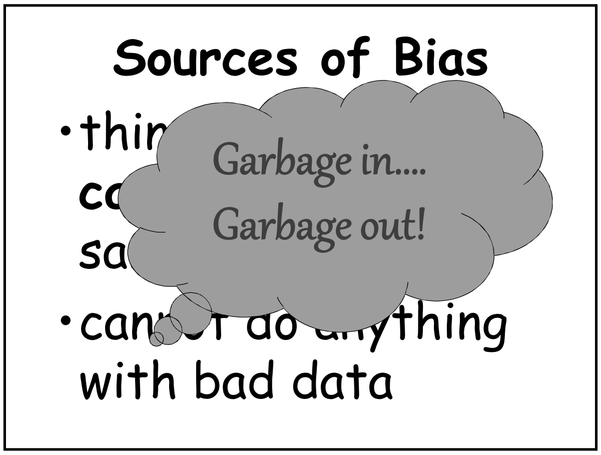
- Sampling is a natural thing to do. Think about sampling something you are cooking—you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- Opinion polls are examples of **sample surveys**, designed to ask questions of a small group of people in the hope of learning something about the entire population.
  - Professional pollsters work quite hard to ensure that the sample they take is representative of the population.
  - If not, the sample can give misleading information about the population.



## Bias

- A syste atic error in mea the est
- fav out

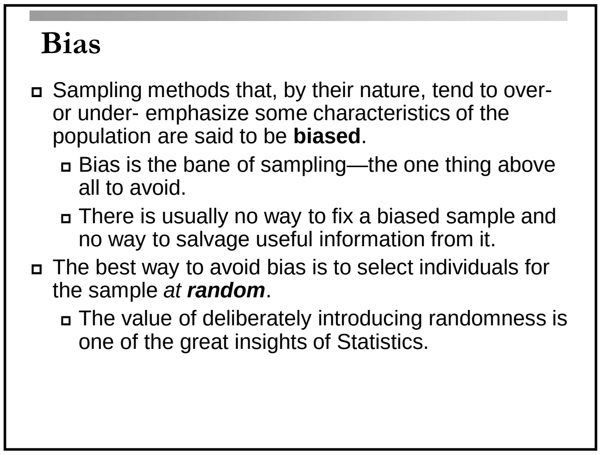
Anything that causes the data to be wrong! It might be attributed to the researchers, the respondent, or to the sampling method!



## Sources of Bias

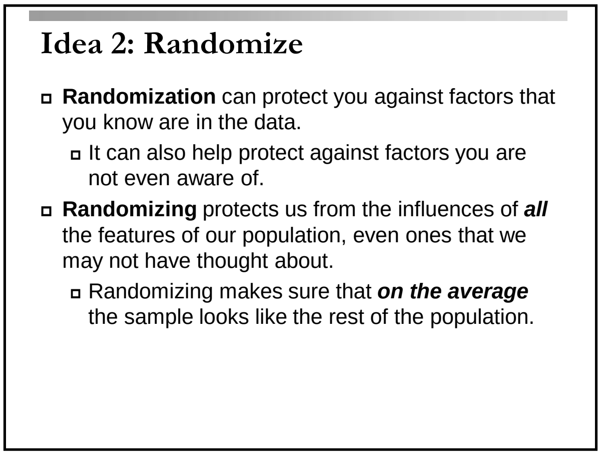
- thin co sa
- can ot do anything with bad data

Garbage in....
Garbage out!



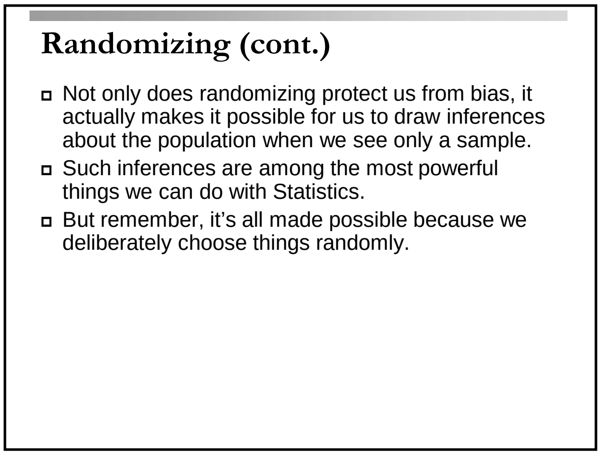
## Bias

- Sampling methods that, by their nature, tend to over- or under- emphasize some characteristics of the population are said to be **biased**.
  - Bias is the bane of sampling—the one thing above all to avoid.
  - There is usually no way to fix a biased sample and no way to salvage useful information from it.
- The best way to avoid bias is to select individuals for the sample *at* ***random***.
  - The value of deliberately introducing randomness is one of the great insights of Statistics.



## Idea 2: Randomize

- **Randomization** can protect you against factors that you know are in the data.
  - It can also help protect against factors you are not even aware of.
- **Randomizing** protects us from the influences of ***all*** the features of our population, even ones that we may not have thought about.
  - Randomizing makes sure that ***on the average*** the sample looks like the rest of the population.



## Randomizing (cont.)

- Not only does randomizing protect us from bias, it actually makes it possible for us to draw inferences about the population when we see only a sample.
- Such inferences are among the most powerful things we can do with Statistics.
- But remember, it's all made possible because we deliberately choose things randomly.

### *In contrast:*

**Probability** or **random sampling:**

Individuals are randomly selected (chosen by chance). No one group should be over-represented.

**Sampling randomly gets rid of bias.**

Random samples rely on the absolute objectivity of random numbers. There are books and tables of **random digits** available for random sampling.

(See **TABLE B**)

Statistical software can generate random digits (e.g., Excel "=random()").

## Idea 3: It's the Sample Size

- How large a random sample do we need for the sample to be reasonably representative of the population?
- **It's the size of the sample**, not the size of the population, that makes the difference in sampling.
  - Exception: If the population is small enough and the sample is more than 10% of the whole population, the population size *can* matter.
- The *fraction* of the population that you've sampled doesn't matter. It's the *sample size* itself that's important.

## Example:

i) In the city of Chicago, Illinois, 1,000 likely voters are randomly selected and asked who they are going to vote for in the Chicago mayoral race.

ii) In the state of Illinois, 1,000 likely voters are randomly selected and asked who they are going to vote for in the Illinois governor's race.

iii) In the United States, 1,000 likely voters are randomly selected and asked who they are going to vote for in the presidential election.

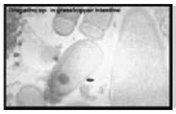
- Which survey has more accuracy?
  - **All the surveys have the same accuracy**

## Observation *vs.* experiment

***Observational study*:** Record data on individuals without attempting to influence the responses. We typically cannot prove anything this way.

***Example:*** Based on observations you make in nature, you suspect that female crickets choose their mates on the basis of their health. → Observe health of male crickets that mated.

***Experimental study*:** Deliberately impose a treatment on individuals and record their responses. Influential factors can be controlled.



***Example:*** Deliberately infect some males with intestinal parasites and see whether females tend to choose healthy rather than ill males.

## Does a Census Make Sense?

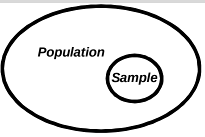
- Why bother worrying the sample size?
- Wouldn't it be better to just include everyone and "sample" the entire population?
  - Such a special sample is called a **census**.

## Does a Census Make Sense?

- There are problems with taking a census:
  - **<u>Practicality</u>**: It can be difficult to complete a census—there always seem to be some individuals who are hard to locate or hard to measure.
  - **<u>Timeliness</u>**: populations rarely stand still. Even if you could take a census, the population changes while you work, so it's never possible to get a perfect measure.
  - **<u>Expense</u>**: taking a census may be more complex than sampling.
  - **<u>Accuracy</u>**: a census may not be as accurate as a good sample due to data entry error, inaccurate (made-up?) data, tedium.

## Population *vs.* sample

- **Population:** The entire group of individuals in which we are interested but can't usually assess directly
  - Note that "***individual***" does not have to mean people
- **Sample:** The part of the population we actually examine and for which we do have data
  - How well the sample represents the population depends on the sample design.
- ***Example***: All humans, all working-age people in California, all crickets



- A ***parameter*** is a number describing a characteristic of the ***population***.
- A ***statistic*** is a number describing a characteristic of a ***sample.***

Are the **BOLD** numbers parameters or statistics?

- A telem uses a device that dials res mbers in that city at rando found to be **$243.27**. Th ause the average electri es that month was **$241.73**.
- The Bureau of nth interviewe abor force and **$49, 056**.

Statistic
(from a

Parameter
(from a
Population)

Statistic
(from a Sample)

## Census *vs.* Survey

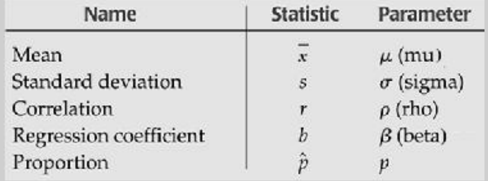
A ***census*** is a study in which every member of a population provides information of interest.

A ***survey*** is a study in which a sample of a population provides information of interest.

## Sample Statistics Estimate Parameters

- Values of **population** parameters are unknown; in addition, they are *unknowable.*
- ***Example:*** The distribution of heights of adult females (at least 18 yrs of age) in the United States is approximately symmetric and mound-shaped with mean **μ**. **μ** is a **population** parameter whose value is unknown and *unknowable*
- The heights of 1,500 females are obtained from a sample of government records. The sample mean $\bar{x}$ of the 1,500 heights is calculated to be 64.5 inches.
- The sample mean $\bar{x}$ is a sample statistic that we use to estimate the unknown population parameter **μ**

**We typically use Greek letters to denote parameters and Latin letters to denote statistics.**



| Name | Statistic | Parameter |
|---|---|---|
| Mean | $\bar{x}$ | $\mu$ (mu) |
| Standard deviation | $s$ | $\sigma$ (sigma) |
| Correlation | $r$ | $\rho$ (rho) |
| Regression coefficient | $b$ | $\beta$ (beta) |
| Proportion | $\hat{p}$ | $p$ |

## Sampling methods

**Voluntary Response Sampling:**

- Individuals selves to be involved
- in
- in all-

***Bias*** outcome.

Remember – the way to determine voluntary response is:

Self-selection!!

- Ann Landers: "If you had it to do over again, would you have children?"
- 10,000 parents responded…
  **"70% of parents say kids not worth it."**

***Bias***: Most letters to newspapers are written by disgruntled people. A random sample showed that 91% of parents WOULD have kids again.

"Hey, Pops, what was that letter you sent off to Ann Landers yesterday?"

## Sampling methods

**Convenience sampling:** Just ask whoever is around.

Q

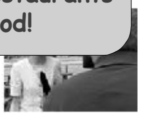
•

*Whic*

- As
  an
  dif
- Ev
  su

The data obtained by a convenience sample will be biased – however this method is often used for surveys & results reported in newspapers and magazines!

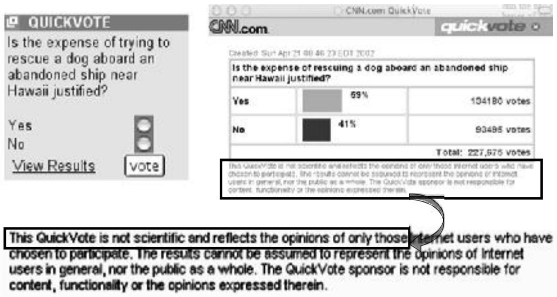
survey. Another example is the surveys left on tables at restaurants - a convenient method!



***Bias:*** Opinions limited to individuals present

**CNN on-line surveys:**

***Bias:*** People have to care enough about an issue to bother replying. This sample is probably a combination of people who hate "wasting the taxpayers' money" and "animal lovers."


QUICKVOTE

Is the expense of trying to rescue a dog aboard an abandoned ship near Hawaii justified?

Yes ○
No ○
View Results vote

CNN.com QuickVote

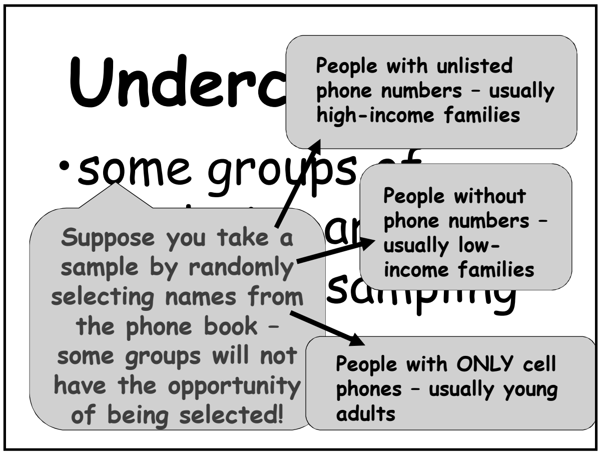
Created Sun Apr 21 00:46:23 EDT 2002

| Is the expense of rescuing a dog aboard an abandoned ship near Hawaii justified? | | |
|---|---|---|
| Yes | 59% | 134180 votes |
| No | 41% | 93495 votes |
| | | Total: 227,675 votes |

This QuickVote is not scientific and reflects the opinions of only those Internet users who have chosen to participate. The results cannot be assumed to represent the opinions of Internet users in general, nor the public as a whole. The QuickVote sponsor is not responsible for content, functionality or the opinions expressed therein.

## Bias in Sampling

- The design of a study is ***biased*** if it systematically favors certain outcomes.
  - A ***voluntary response sample*** is biased in that it favors negative outcomes regardless of the question.
  - A ***convenience sample*** is usually biased in that it favors the opinions of people in a certain location at a certain time.
    - There is no guarantee that such opinion is representative of the population as a whole
  - In both cases, a conscious *choice* is made to include/exclude a respondent
    - We want a method in which the choice is random and does not depend on any individual



# Underc

- some groups of
  ar
  sampling

People with unlisted phone numbers – usually high-income families

Suppose you take a sample by randomly selecting names from the phone book – some groups will not have the opportunity of being selected!

People without phone numbers – usually low-income families

People with ONLY cell phones – usually young adults

# Nonresponse

•

Becaus of huge telemarketing effo the past few years, tele urveys have a MAJOR

One way to help with the problem of nonresponse is to make follow up contact with the people who are not home when you first contact them.

response!

# Response Bias

refers to anything in the survey design that influences the responses.

- Interviewer bias
- Untruthful responses
- The *wording* of a question

Work hard to avoid influencing responses!

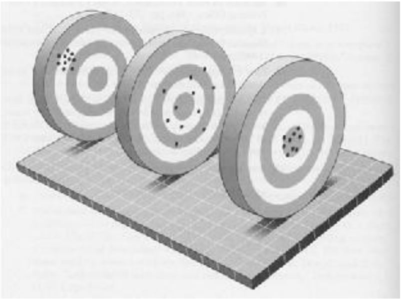
THE WIZARD OF ID
parker and hart
HOW WOULD YOU RATE THE KING'S JOB PERFORMANCE?
OPINION POLL
WOULD YOU SAY HE'S DOING A GOOD JOB, AN OUTSTANDING JOB OR A TERRIFIC JOB?
I THINK I KNOW WHO IS PAYING FOR THIS POLL

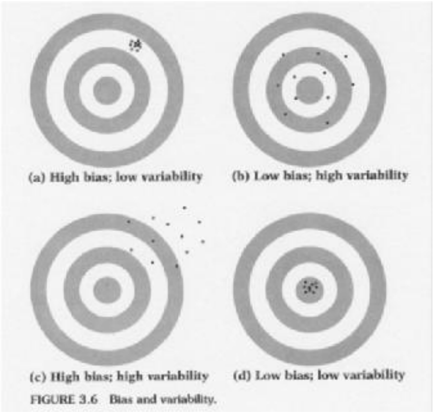
## bias ~~equal~~ error

- ***Sampling error*** is just ***sampling variation.***
- Sampling error simply describes the natural variability in results that will be observed from one sample to the next, none of them exactly capturing the truth in the population.

***Bias*** (ugh!) found in the sampling method…..

Something about the design *systematically distorts the results* so that they are unlikely to reflect reality.





(a) High bias; low variability

(b) Low bias; high variability

(c) High bias; high variability

(d) Low bias; low variability

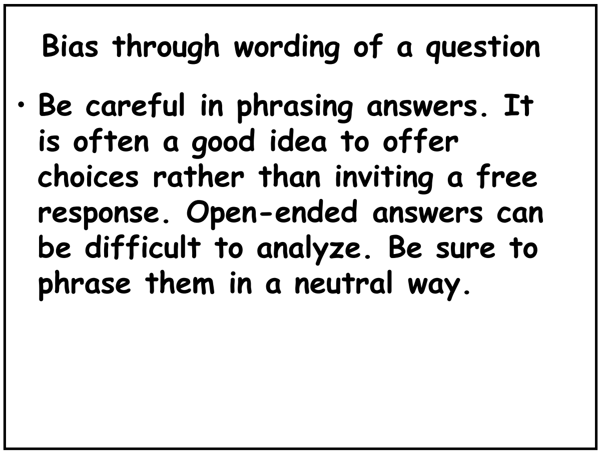
FIGURE 3.6 Bias and variability.

more response bias…

Other examples:

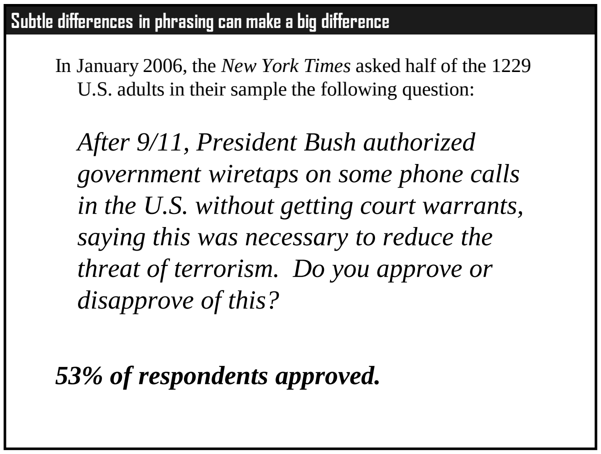
•A uniformed campus police office visits your class and asks every student about their drug use in the last 30 days…

•Your boss at work announces that they need to trim the workforce (read: they need to fire some people), then interviews and asks every employee:
*"Are you satisfied with your current job at this company?"*



## Bias through wording of a question

- **Be careful in phrasing answers. It is often a good idea to offer choices rather than inviting a free response. Open-ended answers can be difficult to analyze. Be sure to phrase them in a neutral way.**

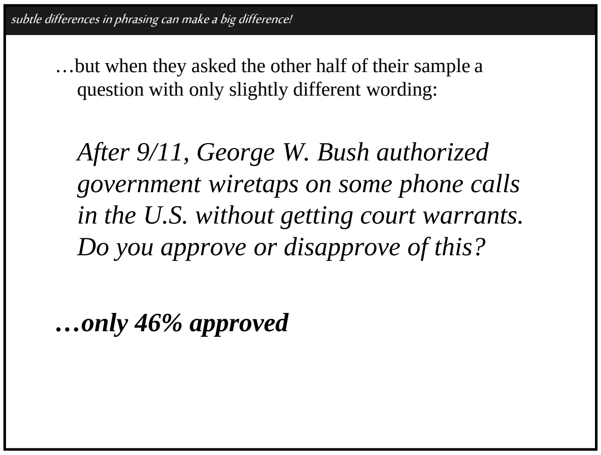


### Subtle differences in phrasing can make a big difference

In January 2006, the *New York Times* asked half of the 1229 U.S. adults in their sample the following question:

*After 9/11, President Bush authorized government wiretaps on some phone calls in the U.S. without getting court warrants, saying this was necessary to reduce the threat of terrorism. Do you approve or disapprove of this?*

***53% of respondents approved.***

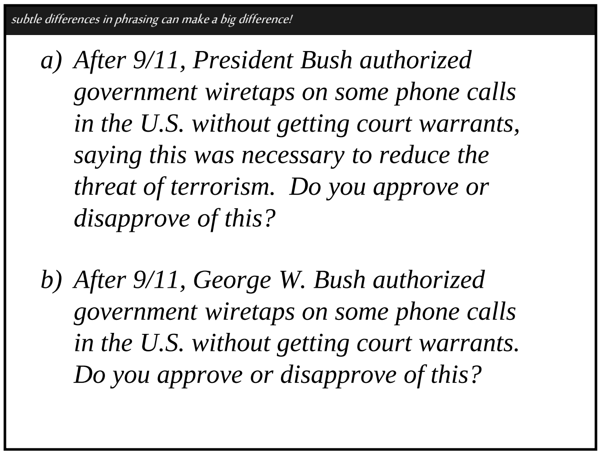


### *subtle differences in phrasing can make a big difference!*

…but when they asked the other half of their sample a question with only slightly different wording:

*After 9/11, George W. Bush authorized government wiretaps on some phone calls in the U.S. without getting court warrants. Do you approve or disapprove of this?*

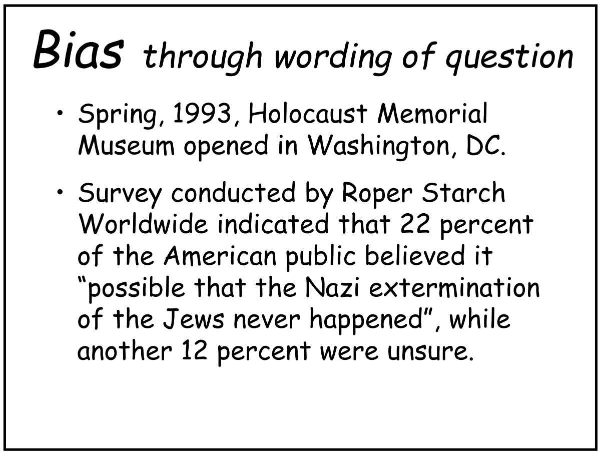
***…only 46% approved***



### *subtle differences in phrasing can make a big difference!*

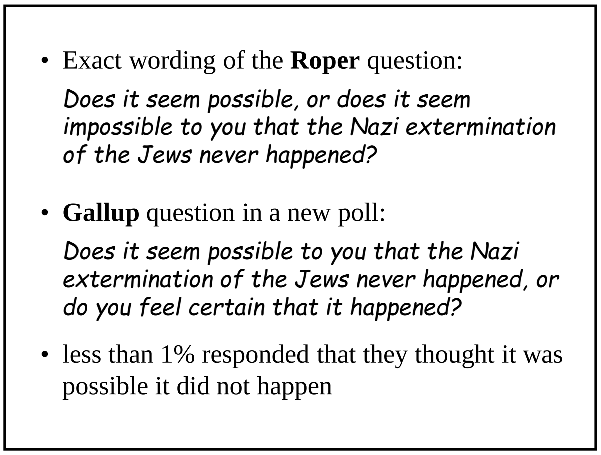
*a) After 9/11, President Bush authorized government wiretaps on some phone calls in the U.S. without getting court warrants, saying this was necessary to reduce the threat of terrorism. Do you approve or disapprove of this?*

*b) After 9/11, George W. Bush authorized government wiretaps on some phone calls in the U.S. without getting court warrants. Do you approve or disapprove of this?*



## *Bias through wording of question*

- Spring, 1993, Holocaust Memorial Museum opened in Washington, DC.
- Survey conducted by Roper Starch Worldwide indicated that 22 percent of the American public believed it "possible that the Nazi extermination of the Jews never happened", while another 12 percent were unsure.



- Exact wording of the **Roper** question:

  *Does it seem possible, or does it seem impossible to you that the Nazi extermination of the Jews never happened?*

- **Gallup** question in a new poll:

  *Does it seem possible to you that the Nazi extermination of the Jews never happened, or do you feel certain that it happened?*

- less than 1% responded that they thought it was possible it did not happen

## Source of Bias?

1) Before the presidential election of 1936, FDR against Republican ALF Landon, the magazine *Literary Digest* predicting Landon winning the election in a 3-to-2 victory. A survey of 2.8 million people. George Gallup surveyed only

**Undercoverage – since the Digest's survey comes from car owners, etc., the people selected were mostly from high-income families and thus mostly Republican! (other answers are possible)**

2) Suppose that you want to estimate the total amount of money spent by students on textbooks each semester at

**Convenience sampling – easy way to collect data**

**or**

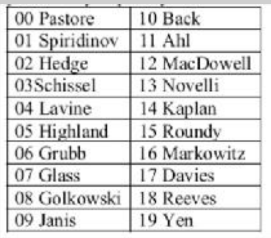
**Undercoverage – students who buy books from on-line bookstores are not included.**

## Example: Hospital employee drug use

- Name the kind of bias that might be present if the administration decides that instead of subjecting people to random testing they'll just…
  - a) interview employees about possible drug abuse.
  - **Response bias:** people will feel threatened, won't answer truthfully
  - b) ask people to volunteer to be tested.
  - **Voluntary response bias:** only those who are "clean" would volunteer

## Example: Hospital employee drug use

- Listed in the table are the names of the 20 pharmacists on the hospital staff. Use the random numbers listed below to select three of them to be in the sample.
- 04905 83852
  29350 91397
  19994 65142
  05087 11232



| | |
|---|---|
| 00 Pastore | 10 Back |
| 01 Spiridinov | 11 Ahl |
| 02 Hedge | 12 MacDowell |
| 03Schissel | 13 Novelli |
| 04 Lavine | 14 Kaplan |
| 05 Highland | 15 Roundy |
| 06 Grubb | 16 Markowitz |
| 07 Glass | 17 Davies |
| 08 Golkowski | 18 Reeves |
| 09 Janis | 19 Yen |

## Simple random samples (SRS)

The **simple random sample (SRS)** is made of randomly selected individuals. Each individual in the population has the same probability of being in the sample and no individual chooses to include/exclude a member of the population. All possible samples of size ***n*** have the same chance of being drawn.

- To select a sample at random, we first need to define where the sample will come from.
  - The **sampling frame** is a list of individuals from which the sample is drawn.
- Once we have our sampling frame, the easiest way to choose an SRS is to assign a random number to each individual in the sampling frame.

## Simple random samples (SRS)

- ***Technically speaking:*** Choose a set of *n* individuals from a population in a manner such that *all sets of size n* had an equal chance of being chosen.

- Samples drawn at random generally differ from one another.
  - Each draw of random numbers selects *different* people for our sample.
  - These differences lead to different values for the variables we measure.
  - We call these sample-to-sample differences **sampling variability.**

# Simple Random Sample

- Advantages
  - Unbiased
  - Easy
- Disadvantages
  - Large variance
  - May not be representative
  - Must have sampling frame (list of population)

## Simple random samples (SRS)

**How to choose an SRS of size *n* from a population of size *N:***

- **LABEL:** We first label each individual in the population with a number (typically from 1 to *N,* or 0 to *N* – 1).
- **TABLE:** A list of random digits is parsed into digits the same length as *N* (if *N* = 233, then its length is 3; if *N* = 18, its length is 2).
- Choose digits in groups sized according to the numbered population
  - ❑ 10 or less individuals, use 1 digit: 0 – 9
  - ❑ 11 – 100 individuals, use 2 digits: 00 – 99
  - ❑ 101 – 1000 individuals, use 3 digits: 000 - 999 etc.
- The parsed list is read in sequence, and the first *n* digits corresponding to a label in our population are selected.
- The individuals with these selected labels thus constitute our SRS.
- Ignore duplicate numbers or numbers beyond the population range.

## Choosing a Simple Random Sample

❑ **From a population of 25 individuals, choose an SRS of size 5 using this table:**

**19223 95034 05752 28713 06409 12531**

19: choose
22: choose
39: ignore (there is not a person number 39)
50: ignore
34: ignore
05: choose
75: ignore
22: ignore (person number 22 is already in the SRS)
87: ignore
13: choose
06: choose

## Choosing a Simple Random Sample

We need to select a random sample of 5 from a class of 20 students.

1) List and number all members of the population, which is the class of 20.
2) The number 20 is two-digits long.
3) Parse the list of random digits into numbers that are two digits long. Here we chose to start with line 103, for no particular reason.

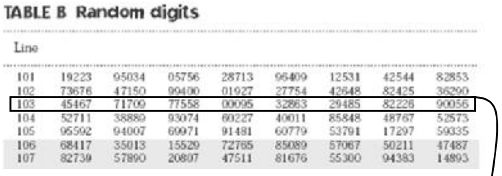


TABLE B Random digits

| Line | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 101 | 19223 | 95034 | 05756 | 28713 | 96409 | 12531 | 42544 | 82853 |
| 102 | 73676 | 47150 | 99400 | 01927 | 27754 | 42648 | 82425 | 36290 |
| 103 | 45467 | 71709 | 77558 | 00095 | 32863 | 29485 | 82226 | 90056 |
| 104 | 52711 | 38889 | 93074 | 60227 | 40011 | 85848 | 48767 | 52573 |
| 105 | 95592 | 94007 | 69971 | 91481 | 60779 | 53791 | 17297 | 59335 |
| 106 | 68417 | 35013 | 15529 | 72765 | 85089 | 57067 | 50211 | 47487 |
| 107 | 82739 | 57890 | 20807 | 47511 | 81676 | 55300 | 94383 | 14893 |

45 46 71 17 09 77 55 80 00 95 32 86 32 94 85 82 22 69 00 56

45 46 71 **17** **09** 77 55 80 00 95 32 86 32 94 85 82 22 69 00 56

52 71 **13** 88 89 93 **07** 46 **02** …

| | |
|---|---|
| 1 | Allison |
| 2 | Amy |
| 3 | Brigitte |
| 4 | Darwin |
| 5 | Emily |
| 6 | Fernando |
| 7 | George |
| 8 | Harry |
| 9 | Henry |
| 10 | John |
| 11 | Kate |
| 12 | Max |
| 13 | Moe |
| 14 | Nancy |
| 15 | Ned |
| 16 | Paul |
| 17 | Ramon |
| 18 | Rupert |
| 19 | Tom |
| 20 | Victoria |

4) Choose a random sample of size 5 by reading through the list of two-digit random numbers, starting with line 103 and on.
5) The first five random numbers matching numbers assigned to people make the SRS.

The first individual selected is Ramon, number 17. Then Henry (9 or "09"). That's all we can get from line 103.

We then move on to line 104. The next three to be selected are Moe, George, and Amy (13, 7, and 2).

***• Remember that 1 is 01, 2 is 02, etc.***
***• If you were to hit 17 again before getting five people, don't sample Ramon twice—you just keep going.***

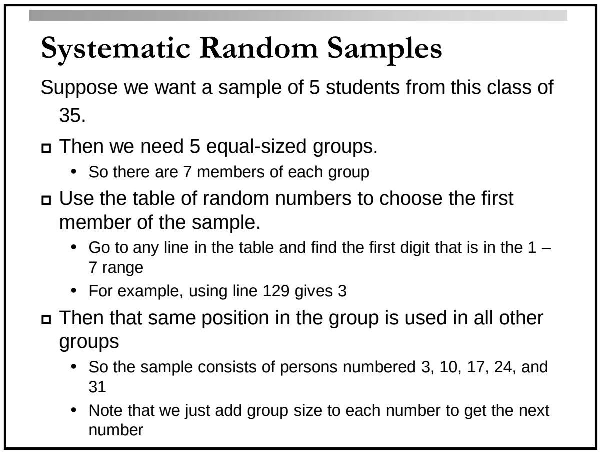
## Systematic Random Samples

A **Systematic Random Sample** is an alternative to an SRS that needs only one random number.

The population is numbered and divided into equal sized groups so that there are as many groups as the desired sample size.

One member of the first group is randomly chosen to be in the sample.

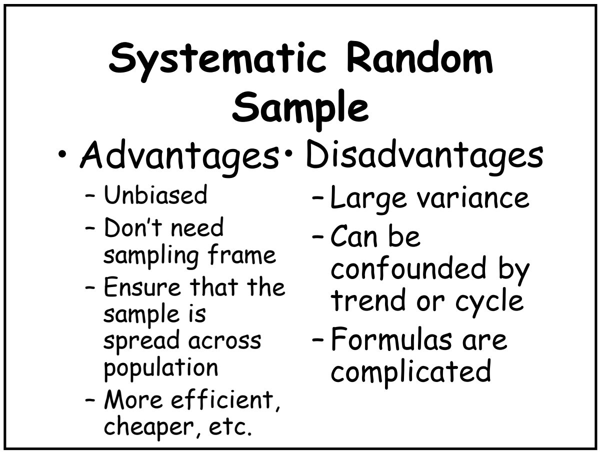
The same-positioned member of all the other groups is then automatically included in the sample.



## Systematic Random Samples

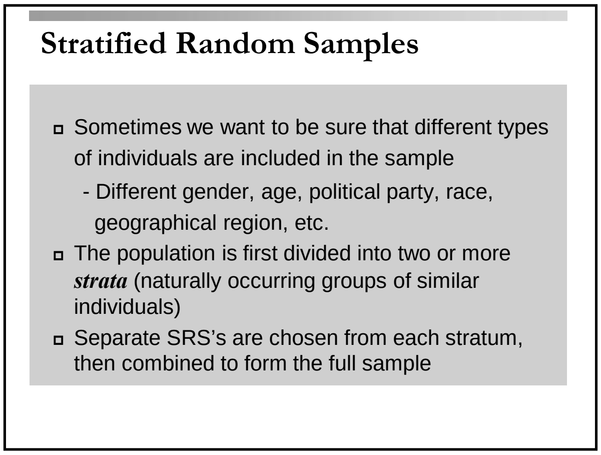
Suppose we want a sample of 5 students from this class of 35.

- Then we need 5 equal-sized groups.
  - So there are 7 members of each group
- Use the table of random numbers to choose the first member of the sample.
  - Go to any line in the table and find the first digit that is in the 1 – 7 range
  - For example, using line 129 gives 3
- Then that same position in the group is used in all other groups
  - So the sample consists of persons numbered 3, 10, 17, 24, and 31
  - Note that we just add group size to each number to get the next number



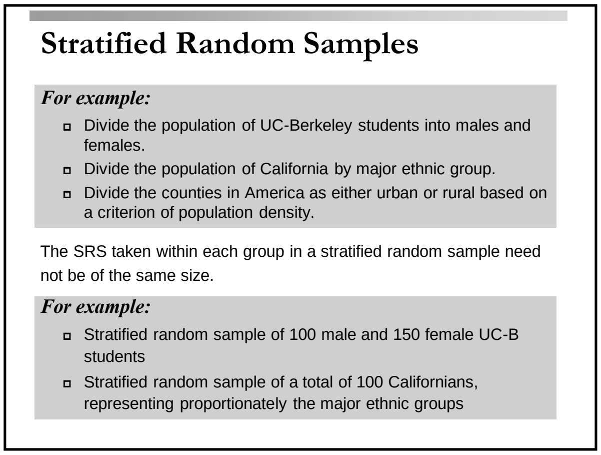
## Systematic Random Sample

- Advantages
  - Unbiased
  - Don't need sampling frame
  - Ensure that the sample is spread across population
  - More efficient, cheaper, etc.
- Disadvantages
  - Large variance
  - Can be confounded by trend or cycle
  - Formulas are complicated



## Stratified Random Samples

- Sometimes we want to be sure that different types of individuals are included in the sample
  - Different gender, age, political party, race, geographical region, etc.
- The population is first divided into two or more ***strata*** (naturally occurring groups of similar individuals)
- Separate SRS's are chosen from each stratum, then combined to form the full sample



## Stratified Random Samples

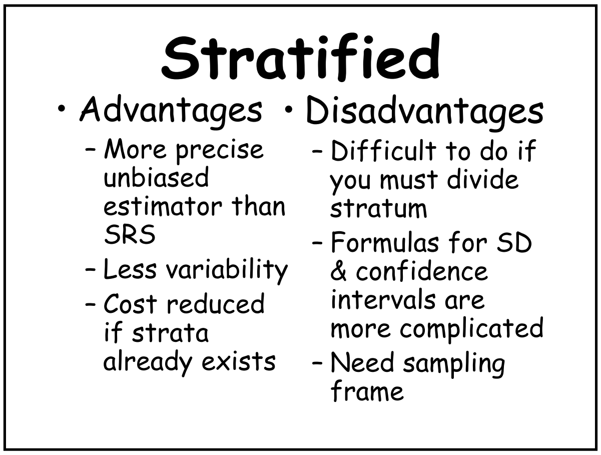
***For example:***

- Divide the population of UC-Berkeley students into males and females.
- Divide the population of California by major ethnic group.
- Divide the counties in America as either urban or rural based on a criterion of population density.

The SRS taken within each group in a stratified random sample need not be of the same size.

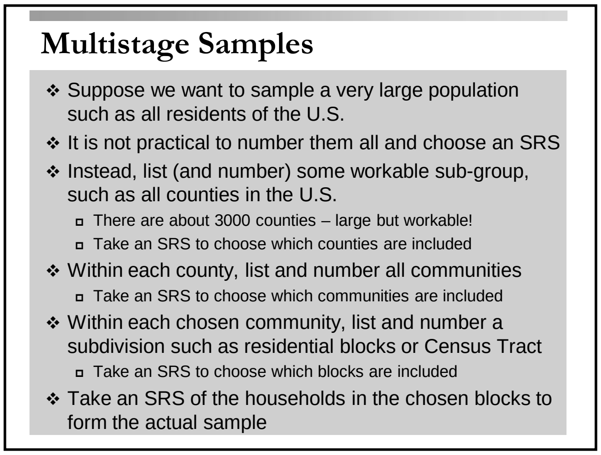
***For example:***

- Stratified random sample of 100 male and 150 female UC-B students
- Stratified random sample of a total of 100 Californians, representing proportionately the major ethnic groups



## Stratified

- Advantages
  - More precise unbiased estimator than SRS
  - Less variability
  - Cost reduced if strata already exists
- Disadvantages
  - Difficult to do if you must divide stratum
  - Formulas for SD & confidence intervals are more complicated
  - Need sampling frame



## Multistage Samples

- ❖ Suppose we want to sample a very large population such as all residents of the U.S.
- ❖ It is not practical to number them all and choose an SRS
- ❖ Instead, list (and number) some workable sub-group, such as all counties in the U.S.
  - There are about 3000 counties – large but workable!
  - Take an SRS to choose which counties are included
- ❖ Within each county, list and number all communities
  - Take an SRS to choose which communities are included
- ❖ Within each chosen community, list and number a subdivision such as residential blocks or Census Tract
  - Take an SRS to choose which blocks are included
- ❖ Take an SRS of the households in the chosen blocks to form the actual sample

***Multistage samples*** use multiple stages of stratification. They are often used by the government to obtain information about the U.S. population.

***Example:*** Sampling both urban and rural areas, people in different ethnic and income groups within the urban and rural areas, and then individuals of different ethnicities within those strata.

Data are obtained by taking an SRS for each substrata.

Statistical analysis for multistage samples is more complex than for an SRS.

Welcome to the American Community Survey web site!

Source: U.S. Census Bureau, Demographic Surveys Division,
Created: August 7, 2001

## Caution about sampling surveys

- ❒ ***Nonresponse:*** People who feel they have something to hide or who don't like their privacy being invaded probably won't answer. Yet they are part of the population.
- ❒ ***Response bias:*** Fancy term for lying when you think you should not tell the truth. Like if your family doctor asks: ***"How much do you drink?"*** Or a survey of female students asking: ***"How many men do you date per week?"*** People also simply forget and often give erroneous answers to questions about the past.
- ❒ ***Wording effects:*** Questions worded like ***"Do you agree that it is awful that…"*** are prompting you to give a particular response. Confusing or leading questions can push toward a certain result.

❑ ***Undercoverage:***

***Undercoverage*** occurs when parts of the population are left out in the process of choosing the sample.

Because the U.S. Census goes "house to house," homeless people are not represented. Illegal immigrants also avoid being counted. Geographical districts with a lot of undercoverage tend to be poor ones. Representatives from richer areas typically strongly oppose statistical adjustment of the census.

Historically, clinical trials have avoided including women in their studies because of their periods and the chance of pregnancy. This means that medical treatments were not appropriately tested for women. This problem is slowly being recognized and addressed.

1) To assess the opinions of students at the Ohio State University regarding campus safety, a reporter interviews 15 students he meets walking on the campus late at night who are willing to give their opinions.

→ ***What is the sample here? What is the population? Why?***

- All those students walking on campus late at night
- All students at universities with safety issues
- The 15 students interviewed
- All students approached by the reporter

2) An SRS of 1200 adult Americans is selected and asked: ***"In light of the huge national deficit, should the government at this time spend additional money to establish a national system of health insurance?"*** Thirty-nine percent of those responding answered yes.

→ ***What can you say about this survey?***

- The sampling process is sound, but the **wording is biased**. The results probably understate the percentage of people who do favor a system of national health insurance.

***Should you trust the results of the first survey? Of the second? Why?***

## Cluster Sampling

- Sometimes stratifying isn't practical and simple random sampling is difficult.
- Splitting the population into <u>similar parts</u> or clusters can make sampling more practical.
- Then we could select one or a few clusters at random and perform a census within each cluster.
- This sampling design is called cluster sampling.
- If each cluster fairly represents the full population, cluster sampling will give us an unbiased sample.

## Cluster Sampling Useful When…

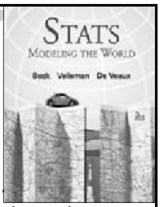
- it is difficult and costly to develop a complete list of the population members (making it difficult to develop a simple random sampling procedure.)

- e.g., all items sold in a grocery store

- the population members are widely dispersed geographically.

- e.g., all Toyota dealerships in North Carolina

### Mean length of sentences in our course text



- We would like to assess the reading level of our course text based on the length of the sentences.
- Simple random sampling would be awkward:
  - number each sentence in the book?
- Better way:
  - choose a few pages at random (the pages are the clusters, and it's reasonable to assume that each page is representative of the entire text).
  - count the length of the sentences on those pages

### Cluster sampling - not the same as stratified sampling!!

- We stratify to ensure that our sample represents different groups in the population, and sample randomly within each stratum.

Strata are homogenous (e.g., male, female) but differ from one another

- Clusters are more or less alike, each heterogeneous and resembling the overall population.
  - We select clusters to make sampling more practical or affordable.
  - We conduct a census on or select a SRS from each selected cluster.

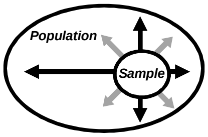
# Cluster Samples

- Advantages
  - Unbiased
  - Cost is reduced
  - Sampling frame may not be available (not needed)
- Disadvantages
  - Clusters may not be representative of population
  - Formulas are complicated

### Learning about populations from samples

The techniques of inferential statistics allow us to draw inferences or conclusions about a population from a sample.

- Your estimate of the population is only as good as your sampling design → Work hard to eliminate biases.
- Your sample is only an estimate—and if you randomly sampled again, you would probably get a somewhat different result.
- The bigger the sample the better. *We'll get back to it in later chapters.*



# Wording of the

Questions must be worded as neutral as possible to avoid influencing the response.

he
answ that are given
- connotation of words
- use of "big" words or technical words

### Identify the sampling design

1) The Educational Testing Service (ETS) needed a sample of colleges. ETS first divided all colleges into groups of similar types (small public, small private, etc). Then they randomly selected 3 colleges from each group.

**Stratified random sample**

## Identify the sampling design

2) A county commissioner wants to survey people in her district to determine their opinions on a particular law up for adoption. She decides to randomly select blocks in her district and then survey all who live on those blocks.

**Cluster sampling**

## Identify the sampling design

3) A local restaurant manager wants to survey customers about the service they receive. Each night the manager randomly chooses a number between 1 & 10. He then gives a survey to that customer, and to every 10th customer after them, to fill it out before they leave.

**Systematic random sampling**

A research group wishes to know the mean GPA of all 2544 students at XYZ High School. To estimate this, they take a random sample of 189 students that have zone classes in the C-wing, and pull those records. The mean GPA of the students in the sample is 2.98. According to the school registrar, the GPA of all 2544 students at XYZ is 3.09.

**Identify the following**

a)Population (of interest): all XYZ HS students

b)Parameter of interest: mean GPA of all students

c)Sampling frame: just students with zone in C-wing

d)Sample: the 189 students selected

A neighborhood interest group wants to know what proportion of households in Austin watch the TV show "So You Think You Can Dance." They select a random sample of 59 houses from Northwest Austin, and find that 35.6% of those families watch the program regularly. Local ratings indicate that about 22% of all households watch SYTYCD on a regular basis.

**Identify the following**

a) Population (of interest): households in Austin
b) Parameter of interest: proportion of households that watch SYTYCD
c) Sampling frame: households in NW Austin
d) Sample: the 59 houses selected

## Just Checking

- Why is each of the following claims not correct?

**It is always better to take a census than to draw a sample**

**It can be hard to reach all members of a population, and it can take so long that circumstances change, affecting the responses. A well-designed sample is often a better choice.**

## Just Checking

- **Stopping students on their way out of the cafeteria is a good way to sample if we want to know about the quality of the food there.**
- **The sample is probably biased – students who didn't like the food at the cafeteria might choose not to eat there.**

## Just Checking

- We drew a sample of 100 from the 3000 students in a school. To get the same level of precision for a town of 30,000 residents we will need a sample of 1000
- Only the sample size matters, not the fraction of the overall population.
- A poll taken at a statistic support website garnered 12,357 responses. The majority said they enjoy doing statistics homework. With a sample size that large we can be pretty sure that most statistics students feel this way, too.
- Students who frequent this website might be more enthusiastic about stats than the overall population of stat students. A large sample cannot compensate for this bias.

## Just Checking

- The true percentage of all Stats students who enjoy the homework is called the "population statistic"
- It's the population "parameter." "Statistic" describe the samples.

## Just Checking

- We need to survey a random sample of 300 of the passengers on a flight from San Francisco to Tokyo. Name each sampling method described:
- 1) Pick every 10th passenger as people board the plane
- **Systematic**
- 2) From the boarding list randomly sample 5 people flying first class and 25 of the other passengers
- **Stratified**

## Just Checking

- We need to survey a random sample of 300 of the passengers on a flight from San Francisco to Tokyo. Name each sampling method described:
- 3) Randomly generate 30 seat numbers and survey the passengers who sit there
- **Simple**
- 4) Randomly select a seat position (right window, right center, right aisle, etc.)
- **Cluster**

# The Valid Survey

- A valid survey yields the information we are seeking about the population we are interested in:
- Before setting out to survey, ask yourself:
- What do I want to know?
- Am I asking the right respondents?
- Am I asking the right questions?
- What would I do with the answers if I had them; would they address the things I want to know?

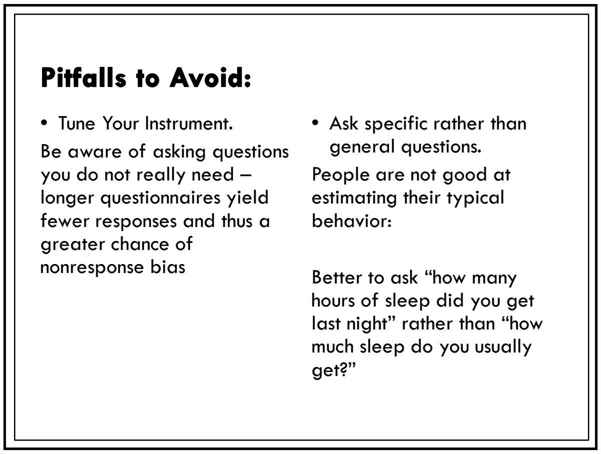
## Pitfalls to Avoid:

- Know what you want to know!

Have a clear idea of what you hope to learn and about whom you hope to learn it.

- Use the right frame.

Be sure you have an appropriate *sampling frame*: have you identified the population of interest and sampled from it appropriately?



## Pitfalls to Avoid:

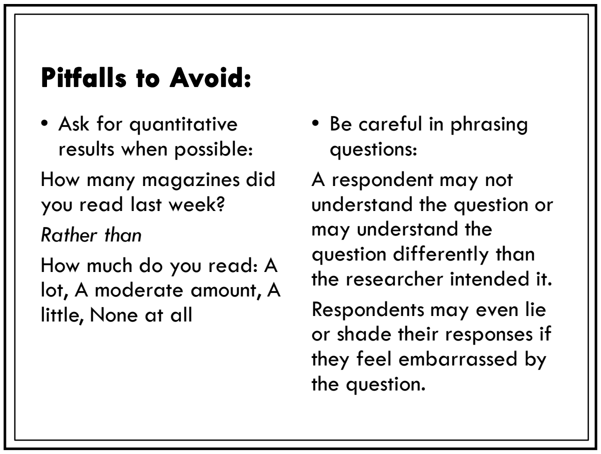
- Tune Your Instrument.

Be aware of asking questions you do not really need – longer questionnaires yield fewer responses and thus a greater chance of nonresponse bias

- Ask specific rather than general questions.

People are not good at estimating their typical behavior:

Better to ask "how many hours of sleep did you get last night" rather than "how much sleep do you usually get?"



## Pitfalls to Avoid:

- Ask for quantitative results when possible:

How many magazines did you read last week?

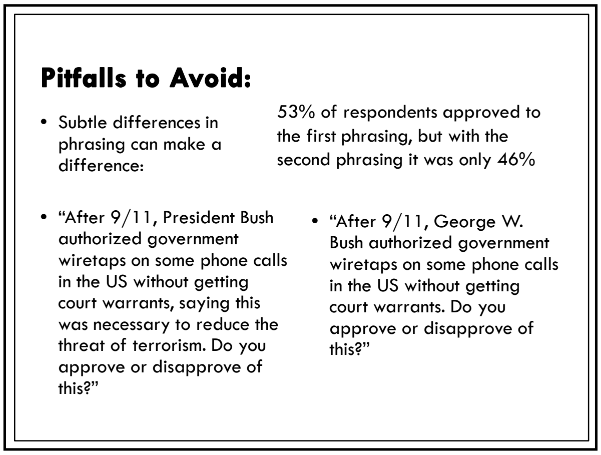
*Rather than*

How much do you read: A lot, A moderate amount, A little, None at all

- Be careful in phrasing questions:

A respondent may not understand the question or may understand the question differently than the researcher intended it.

Respondents may even lie or shade their responses if they feel embarrassed by the question.



## Pitfalls to Avoid:

- Subtle differences in phrasing can make a difference:

53% of respondents approved to the first phrasing, but with the second phrasing it was only 46%

- "After 9/11, President Bush authorized government wiretaps on some phone calls in the US without getting court warrants, saying this was necessary to reduce the threat of terrorism. Do you approve or disapprove of this?"

- "After 9/11, George W. Bush authorized government wiretaps on some phone calls in the US without getting court warrants. Do you approve or disapprove of this?"