TOPIC 1.1

An Overview of Statistics

OBJECTIVES Define statistics. Tell a population from a sample and a parameter from a statistic. Tell descriptive statistics from inferential statistics.

GETTING STARTED A school newspaper wants to know how many hours per week the 1,200 students at its high school spend on homework. A reporter asks 80 students in the cafeteria at lunch.

a.) Which group does the newspaper want to learn about? all 1,200 students

b.) Which group did the reporter actually ask? the 80 students asked

c.) Will the average for the 80 students be exactly the same as the average for all 1,200 students? Explain.

Probably not. The 80 students are only part of the school, so their average will be close to, but not exactly, the school average.

What Is Statistics?

DEFINITION **Data and Statistics**

Data are information that comes from observations, counts, measurements, or responses.

Statistics is the science of collecting , organizing , analyzing , and interpreting data in order to make decisions.

Populations and Samples

DEFINITION **Population and Sample**

A **population** is the collection of all outcomes, responses, measurements, or counts that are of interest.

A **sample** is a subset , or part, of the population. (Strictly, both are sets of responses or measurements, such as “the heights of all students.”)

A **census** collects data from every member of the population. A **sampling** collects data from part of the population.

Example 1 An inspector measures the amount of juice in 50 bottles chosen from one day's production of 20,000 bottles.

a.) Population amounts of juice in all 20,000 bottles

b.) Sample amounts of juice in the 50 bottles

Example 2 In a survey of 1,500 U.S. adults, 41% said they get most of their news from social media.

a.) Population responses of all U.S. adults

b.) Sample responses of the 1,500 adults surveyed

Parameters and Statistics

DEFINITION **Parameter and Statistic**

A **parameter** is a number that describes a population characteristic.

A **statistic** is a number that describes a sample characteristic.

Parameters use Greek letters, such as μ (mean) and σ (standard deviation). Statistics use Roman letters, such as $\bar{x}$ and s.

TIP **Match the Letters**

Population goes with **p**arameter. **S**ample goes with **s**tatistic. A parameter is a fixed number (often unknown). A statistic changes from sample to sample. **Clue words:** all, every, entire, census → population. Some, part, a poll of, randomly selected → sample.

Compare	Population	Sample
Is	all members of interest	part of the population
Described by a	parameter	statistic
Symbols	Greek: μ, σ	Roman: $\bar{x}$, s

Example 3 Decide whether each number is a parameter or a statistic. Explain.

a.) In a survey of 400 voters in a city, 54% support building a new park. statistic

54% comes from a sample of 400 voters.

b.) All 212 graduates of a nursing program had a mean starting salary of $68,400. parameter

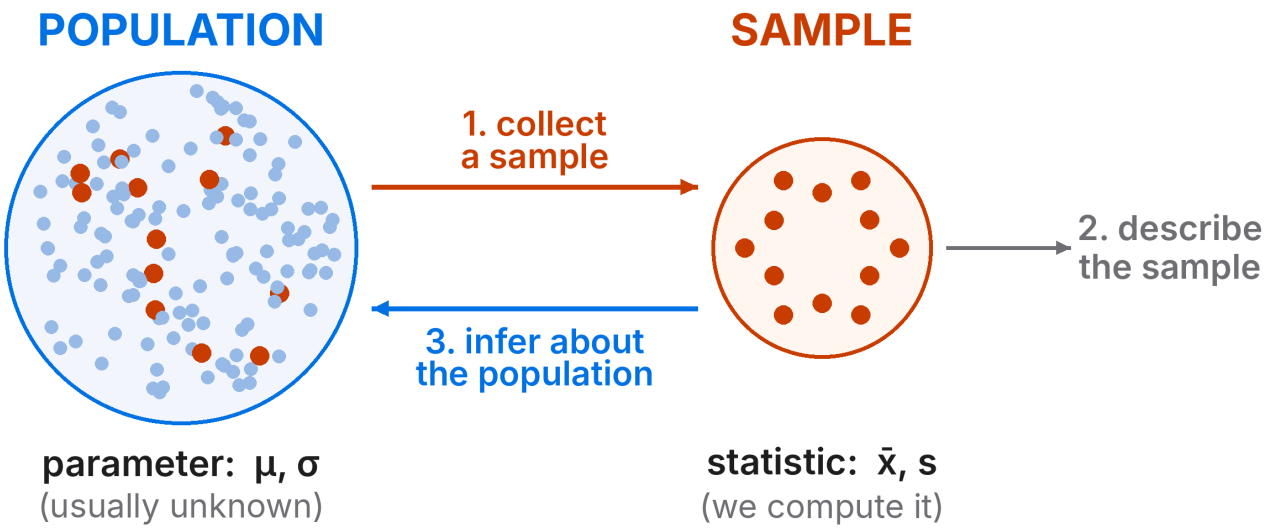
$68,400 describes all 212 graduates (population).

c.) A check of 60 randomly selected homes in a town found a mean water use of 290 gallons per day. statistic

290 gallons comes from a sample of 60 homes.

d.) Of the 32 teams in the NFL, 2 are based in Ohio (the Browns and the Bengals). parameter

2 of 32 describes ALL NFL teams (population).



Branches of Statistics

DEFINITION **Descriptive and Inferential Statistics**

Descriptive statistics is the branch of statistics that involves organizing, summarizing, and displaying data.

Inferential statistics is the branch of statistics that uses a sample to draw conclusions about a population. Probability is a basic tool of inferential statistics.

Example 4 A study of 2,000 adults ages 20 to 40 found that adults who slept less than 6 hours a night drank an average of 2.4 cups of coffee per day. Adults who slept at least 8 hours drank an average of 1.1 cups per day.

a.) Identify the population. coffee use of all adults ages 20 to 40

b.) Identify the sample. coffee use of the 2,000 adults studied

c.) Which part of the study is descriptive statistics?

The averages: 2.4 cups per day for less than 6 hours of sleep, 1.1 cups for at least 8 hours.

d.) What inference might be drawn from the study?

Adults ages 20 to 40 who sleep less tend to drink more coffee.

TIP
Push back: does coffee hurt sleep, or does poor sleep lead to coffee? A link is not a cause.

WATCH OUT Inferences Can Be Wrong

A sample result is an **estimate**. A different sample would probably give a slightly different result. Also, a link between two variables does not prove that one **causes** the other.

Example 5 Decide whether each statement uses descriptive or inferential statistics. Explain.

a.) In last year's graduating class, 72% of students took at least one AP or CCP course. descriptive

It only summarizes last year's class.
No conclusion beyond the data.

b.) Based on a survey of 600 drivers, about 1 in 4 drivers in the state text while driving. inferential

Uses a sample of 600 drivers to draw a conclusion about ALL drivers in the state.

CHECK YOUR UNDERSTANDING

Which statement describes a **parameter**?

(A) The mean height of 25 students chosen from a school of 900 students is 66 inches.
(B) In a poll of 1,000 adults, 38% own an electric car.
(C) The mean age of all 14 employees of a bakery is 31 years.
(D) In a sample of 50 phones, 3 had a defective battery.

(C) all 14 employees is the whole population

TOPIC 1.2

Data Classification

OBJECTIVES Tell qualitative data from quantitative data. Classify data by level of measurement: nominal, ordinal, interval, or ratio.

GETTING STARTED A customer survey card asks for three things: (1) your ZIP code, (2) your age in years, and (3) a rating of the service: poor, fair, good, or excellent.

a.) For which item would it make sense to find an average? (2) age

Only ages are true amounts. An "average ZIP code" or an "average rating word" means nothing.

TIP
Ask for the average ZIP code of the room. The silly answer makes the point.

b.) Which item has answers that can be put in a meaningful order but are not numbers? (3) rating

Types of Data

DEFINITION Qualitative and Quantitative Data

Qualitative data consist of attributes, labels, or nonnumerical entries.

Quantitative data consist of numerical measurements or counts.

WATCH OUT Numbers Can Be Labels

A number is quantitative only if it measures or counts something. ZIP codes, jersey numbers, and phone numbers are **qualitative**: they are labels.

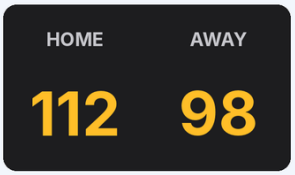
QUALITATIVE: labels



The number 23 is a name tag, not an amount.

Averaging jersey numbers means nothing.

QUANTITATIVE: counts or measures



Points are counts.

112 − 98 = 14 makes sense: a 14-point win.

Example 1 Decide whether each variable is qualitative or quantitative.

a.) Eye color qualitative

b.) Number of siblings quantitative

c.) Jersey number of an NBA player (LeBron James has worn 23 and 6) qualitative

d.) Time to run a mile, in minutes quantitative

e.) Type of phone (brand) qualitative

Levels of Measurement

DEFINITION The Four Levels of Measurement

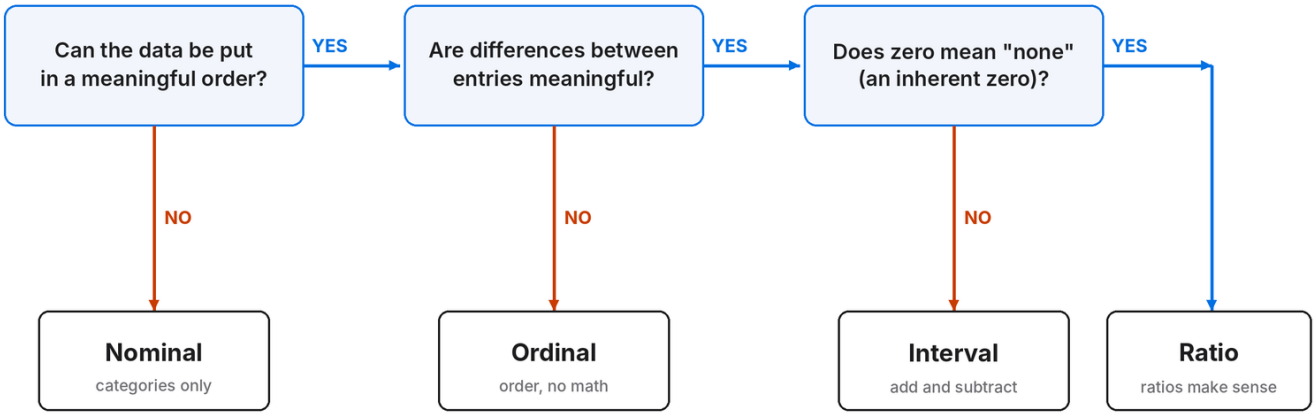
Nominal: qualitative data that can only be sorted into categories by name, label, or quality. The data cannot be put in a meaningful order.

Ordinal: data that can be put in a meaningful order, but differences between entries are not meaningful.

Interval: quantitative data that can be ordered, and differences between entries are meaningful. Zero is only a position on a scale; it does not mean "none."

Ratio: like interval, but zero is an **inherent zero** (zero means "none"). Ratios are meaningful, so "twice as much" makes sense.

Deciding the level: start at the left and stop at the first NO. Data are classified at the highest level they fit.



Example 2 Complete the table. Write Yes or No.

Level	Put in categories	Put in order	Subtract (differences)	Ratios ("twice as")
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

TIP Inherent Zero Test

Ask: "Does zero mean **none** of the quantity?" 0 pounds means no weight, so weight is ratio. 0°F does not mean no temperature, so °F is interval.

Example 3 Identify the level of measurement of each data set. Explain.

a.) The colors of the cars in a parking lot nominal

Colors are labels only, with no order.

b.) The final AP Top 25 college football rankings (1st, 2nd, 3rd, …) ordinal

Ranks have an order, but the gap between #1 and #2 is not a measured amount.

c.) Years the Cincinnati Reds won the World Series: 1919, 1940, 1975, 1976, 1990 interval

Differences make sense (1990 − 1976 = 14 years), but year 0 is not "no time," so no ratios.

d.) The heights of NBA players, in inches ratio

0 inches means no height (inherent zero). An 84-inch player is twice as tall as a 42-inch child.

e.) Daily high temperatures, in degrees Fahrenheit interval

0°F is not "no heat," so no inherent zero.

TIP
zero test: does 0 mean NONE?
If not, it is interval.

f.) The number of home runs each MLB player hit last season ratio

0 home runs means none: inherent zero.

Example 4 On Monday the high temperature was 80°F. On Tuesday it was 40°F. Is it correct to say Monday was twice as hot as Tuesday? Explain.

No. Temperature in °F is interval level. 0°F does not mean "no heat," so ratios like "twice as hot" are not meaningful.

TIP
Ask what 0°F feels like outside. Cold, but not "no temperature."

Example 5 Movie ratings G, PG, PG-13, and R are used to describe how suitable a film is for young viewers.

a.) Qualitative or quantitative? qualitative

Ratings are labels, not measurements.

b.) Level of measurement? ordinal

They can be ordered from least to most mature, but differences between them are not amounts.

CHECK YOUR UNDERSTANDING

Which data set is at the **ratio** level of measurement?

(A) The jersey numbers of a basketball team
(B) Customer ratings of 1 to 5 stars
(C) The distances, in miles, that employees drive to work
(D) Body temperatures, in degrees Celsius

(C) 0 miles means no distance (inherent zero)

TOPIC 1.3

Data Collection and Experimental Design

OBJECTIVES Design a statistical study. Tell an observational study from an experiment. Design an experiment. Choose and identify sampling techniques, and spot a biased sample.

GETTING STARTED A company wants to know whether its new energy drink improves reaction time.

Plan A: Find students who already drink it and students who do not. Compare their reaction times.

Plan B: Take 40 volunteers. Randomly assign 20 to drink the energy drink and 20 to drink a look-alike drink with no caffeine. Compare their reaction times.

Which plan gives better evidence that the drink **causes** faster reactions? Explain. Plan B

In Plan A the groups may differ in other ways (sleep, sports, age). Random assignment in Plan B makes the groups alike except for the drink.

TIP
Have students vote first, then defend their pick. Only random assignment can show cause.

1. **PROCEDURE** **Designing a Statistical Study**

1. Identify the **variable(s)** of interest and the **population** of the study.
2. Plan how to collect the data. If you use a sample, make sure it **represents** the population.
3. Collect the data.
4. Describe the data with **descriptive** statistics.
5. Interpret the data and make decisions about the population with **inferential** statistics.
6. Identify any possible **errors**.

Ways to Collect Data

DEFINITION Four Ways to Collect Data

Observational study: researchers observe and measure characteristics of interest but do not change existing conditions.

Experiment: a treatment is applied to part of a population, called the **treatment group**, and responses are observed. Another part, the **control group**, may get no treatment or a fake treatment.

Simulation: uses a mathematical or physical model to reproduce the conditions of a situation or process. Often done with computers.

Survey: an investigation of one or more characteristics of a population, usually by asking people questions .

Example 1 Decide which method of data collection fits best: observational study, experiment, simulation, or survey.

a.) Testing whether a new fertilizer increases the number of tomatoes per plant experiment

b.) Since 1948, the Framingham Heart Study has followed thousands of residents of Framingham, Massachusetts, recording their health and habits observational study

c.) Using a computer model to predict damage to a car in a crash simulation

d.) Finding the percent of students who want a later school start time survey

Experimental Design

DEFINITION Key Elements of an Experiment

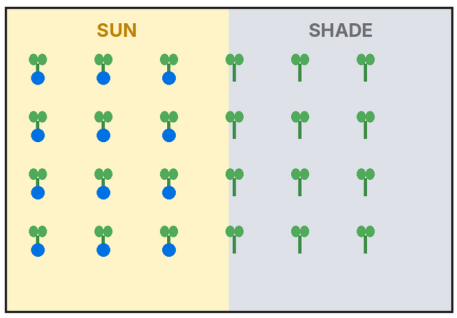
Control: limit the influence of other factors. A **confounding variable** occurs when you cannot tell the difference between the effects of different factors on a variable.

Placebo: a harmless, fake treatment that looks like the real one. The **placebo effect** occurs when a subject reacts favorably to a placebo.

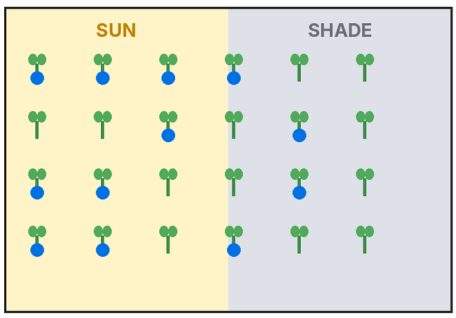
Blinding: subjects do not know whether they get the treatment or the placebo. **Double-blind:** neither the subjects nor the experimenter knows.

Randomization: subjects are assigned to groups at random .

Replication: the experiment is repeated on a large group of subjects. The **sample size** is the number of subjects.



Confounded: fertilizer only in the sun



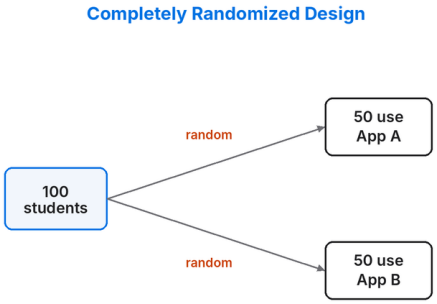
Better: fertilizer assigned at random

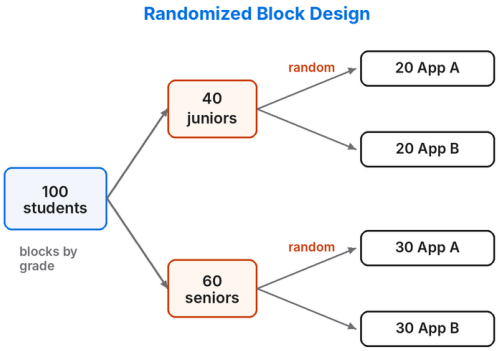
● = plant given the new fertilizer

DEFINITION Two Common Designs

Randomized block design: first divide subjects with similar characteristics into blocks. Then randomly assign treatments within each block.

Matched-pairs design: subjects are paired up by similar characteristics. One subject in each pair gets one treatment and the other gets a different treatment.





Example 2 A real experiment. In 1954, about 400,000 children took part in a test of the new Salk polio vaccine. About half were randomly assigned to get the vaccine. The other half got a placebo shot that looked the same. Neither the children, their parents, nor the doctors knew who got which. (Source: Francis Report, 1955)

a.) Describe the treatment group. the children who got the vaccine

b.) Describe the control group. the children who got the placebo

c.) Why give the control group a placebo shot instead of nothing?

To account for the placebo effect, and so doctors judge both groups the same way. Everyone gets a shot.

d.) Was the experiment blind or double-blind? Explain. double-blind

Neither the subjects (children) nor the doctors who diagnosed polio knew who got the vaccine.

TIP
Tell the story: the children were called Polio Pioneers. In April 1955 the vaccine was declared safe and effective.

e.) Polio struck young children more often than older children. Describe a randomized block design.

Divide the children into age blocks, such as ages 6–7 and 8–9. Within each block, randomly assign half to the vaccine, half to the placebo.

Example 3 A gym tests a new protein shake. Members who lift weights 5 days a week drink the shake. Members who lift 2 days a week do not. After 8 weeks, the shake group gained more strength.

a.) Identify a problem with this experiment.

Confounding variable: days of lifting per week. The shake group also lifted more often, so we cannot tell whether the shake or the extra workouts caused the gain.

b.) Suggest a way to fix it.

Use a randomized block design: block by days of lifting per week, then randomly assign half of each block to the shake.

Example 4 A company tests a new energy bar. Four runners eat the bar before a race and four do not. The bar group ran faster.

a.) Identify a problem with this experiment.

The sample size is too small. With only 4 runners per group, the difference could easily be due to chance.

b.) Suggest a way to fix it.

Replication: repeat the experiment with many more runners, randomly assigned to the two groups.

Sampling Techniques

DEFINITION Sampling Techniques

Random sample: every member of the population has an equal chance of being selected.

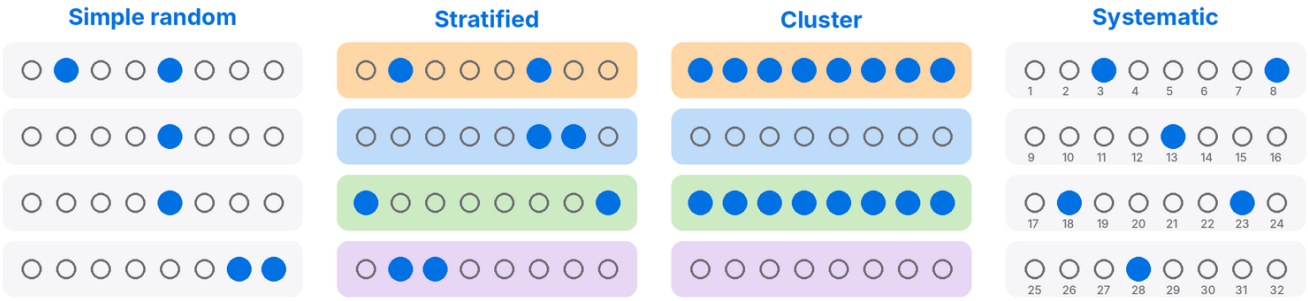
Simple random sample: every possible sample of the same size has the same chance of being selected.

Stratified sample: divide the population into groups (**strata**) that share a characteristic, then randomly select members from each stratum.

Cluster sample: divide the population into groups (**clusters**), randomly select some clusters, and use all members of the chosen clusters.

Systematic sample: number the members, pick a random starting number, then select every *k*th member.

Convenience sample: uses only members who are easy to reach. It is usually biased.



TIP Why Stratify?

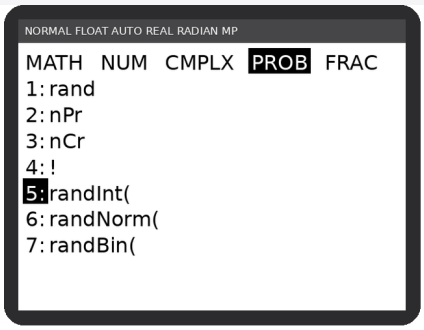
A simple random sample of 20 students could, just by chance, be mostly freshmen. Randomly choosing 5 students from **each** grade guarantees that every grade is represented.

WATCH OUT Biased Samples and Sampling Error

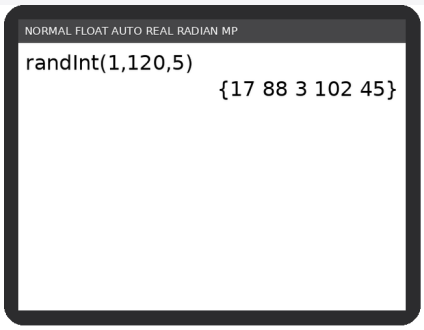
A sample is **biased** if some part of the population is favored or left out. Even a good random sample will not match the population exactly. The difference between a sample result and the population result is **sampling error**.

TIP TI-84: Random Integers

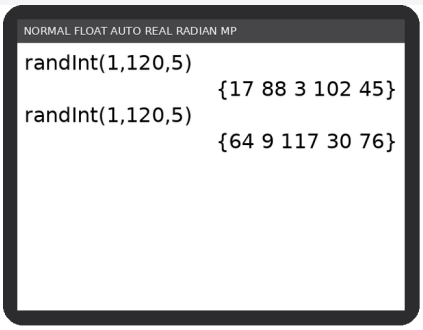
To choose 5 random numbers from 1 to 120: press **MATH**, arrow right to **PROB**, choose **5: randInt(**, and enter **randInt(1, 120, 5)**. If a number repeats, skip it and press **ENTER** again for more.



1 MATH, arrow right to PROB, choose 5:randInt(



2 Type 1, 120, 5 and press ENTER



3 ENTER again gives a new set

Example 5 Identify the sampling technique. Then tell whether the sample could be biased.

a.) A school picks a random name among the first 15 names on its alphabetical roster, then selects every 15th name after it. Systematic

every 15th name → systematic. Not likely biased.

b.) College athletes are divided by NCAA division (I, II, III), and 100 athletes are randomly selected from each division. Stratified

random from EACH group → stratified.

c.) A researcher randomly selects 5 of the 30 MLB teams and surveys every player on those 5 teams. cluster

ALL members of some groups → cluster. Could be biased if the 5 teams differ from the rest.

TIP
Stratified: SOME from EACH group. Cluster: ALL from SOME groups.

d.) A reporter interviews the first 30 people who walk out of a coffee shop. convenience

Easy to reach → convenience. Biased: only coffee shop customers at that time are heard.

e.) A company puts every employee ID number into a random number generator and chooses 50 numbers. Simple random

Every group of 50 has the same chance.

Example 6 A famous biased sample. In 1936, *Literary Digest* magazine mailed poll cards to about 10 million people, chosen mostly from telephone books and car registration lists. About 2.4 million replied, and the poll predicted that Alf Landon would easily beat President Franklin Roosevelt. Roosevelt won 46 of the 48 states.

a.) Why was the sample biased?

During the Great Depression, people with phones and cars were wealthier than most voters, and wealthier voters favored Landon. Also, only people who chose to reply were counted.

b.) The sample had 2.4 million people. Why didn't the large size fix the problem?

A biased method stays biased no matter how big the sample is. More of the wrong people is still the wrong people.

TIP
Repeat all year: a bigger sample does not fix a biased method.

Example 7 A teacher numbers 28 students from 1 to 28 and wants to choose 4 of them at random. What should the teacher enter on the TI-84? randInt(1, 28, 4)

randInt(lowest, highest, how many)
randInt(1, 28, 4)

Example 8 Decide whether you would take a census or use a sample. Explain.

a.) The mean height of the 15 players on an NBA team's roster census

The population is small and easy to measure.

b.) The mean number of hours a brand of light bulb lasts sample

Testing burns out the bulb. Testing every bulb would leave none to sell.

CHECK YOUR UNDERSTANDING

A researcher divides a town into 200 city blocks, randomly selects 10 blocks, and surveys every household on those blocks. Which sampling technique is this?

(A) stratified
(B) cluster whole blocks are chosen, every household used
(C) systematic
(D) convenience