Statistics involves
1. Asking questions
2. Collecting data to address those questions
3. Analyzing the data
4. Drawing conclusions from the data
5. Communicating the results

- The **observational units** in a statistical study are the objects described by a set of data (people, animals, things).
- A **variable** is any characteristic of an observational unit, which can take different values (i.e., can vary) for different observational units.
  - Some variables are **quantitative**, taking numerical values on which ordinary arithmetic operations make sense.
  - Other variables are **categorical**, taking category designations.
    - A categorical variable with only two categories is called **binary**.

**Example 1-1: Variables on You**
Consider the students in this class as the cases in a statistical study.

a) For each of the following variables, indicate whether it is quantitative or categorical:

- How many *Harry Potter* books have you read?
- With which hand do you write?
- How many hours have you slept in the past 24 hours?
- Have you slept for at least 7 hours in the past 24 hours?
- On what day of the week were you born?
- Do you have a facebook account?
- How many facebook friends do you have?

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?

*Summaries are not variables.*

b) Explain why the following questions are *not* variables:

- What is the average number of *Harry Potter* books read by a student in this class?
- What percentage of students in this class are left-handed?
d) Explain why the following questions are not variables.

- Have women at Cal Poly read more *Harry Potter* books, on average, than men?

- Is the proportion of undergraduate Cal Poly students with a facebook account smaller than the proportion of Cal Poly graduate students with a facebook account?

*Research questions are not variables.*

**Example 1-2: Statistical Studies**

For each of the following statistical studies, identify the cases and variables. Also classify each variable as quantitative or categorical (also binary?).

a) How much did an average American consumer spend on Christmas presents in 2013?

b) Is the price of a house related to its size?

c) Is the residence situation of a college student (on-campus, off-campus with parents, off-campus without parents) related to how much alcohol the student consumes?

d) Can you predict how far a cat can jump based on factors such as its length?

e) Among heterosexual couples, who is more likely to say “I love you” first: the man or the woman?

f) An article in a 2006 issue of *Journal of Behavioral Decision Making* reports on a study involving 47 undergraduate students at Harvard. All of the participants were given $50, but some (chosen at random) were told that this was a “tuition rebate,” while the others were told that this was “bonus income.” After one week, the students were contacted again and asked how much of the $50 they had spent and how much they had saved. Those in the “rebate” group had
spent an average of $22.04, while those in the “bonus” group had spent an average of $9.55. Is this difference statistically significant?

g) Statistical evidence was used in the murder trial of Kristen Gilbert, a nurse who was accused of killing patients. More than one thousand 8-hour shifts were analyzed. Was the proportion of shifts with a death significantly higher for the shifts that Gilbert worked?

Much of statistics involves making inferences from a sample to a population.
- **Population**: the *entire* group of observational units (people or objects) about which information is desired
- **Sample**: a (typically small) *part* of the population from which data are gathered
  - **Sample size**: number of observational units in the sample
- **Goal**: Choose a sample that is *representative* of the population

**Example 1-3: Elvis Presley and Alf Landon**
Elvis Presley is reported to have died in his Graceland mansion on August 16, 1977. On the twelfth anniversary of this event, a Dallas record company sponsored a national call-in survey. Listeners of over 1000 radio stations were asked to call a 1-900 number (at a charge of $2.50) to voice an opinion concerning whether or not Elvis was really dead. It turned out that 56% of the callers felt that Elvis was alive. Suppose that this record company was interested in the opinions of all adult Americans on this issue.

a) Identify the population and sample in this study.

population: sample:

b) Do you think that 56% is an accurate reflection of beliefs of all Americans on this issue? If not, identify some of the flaws in the sampling method.

In 1936, *Literary Digest* magazine conducted the most extensive (to that date) public opinion poll in history. They mailed out questionnaires to over 10 million people whose names and addresses they had obtained from telephone books and vehicle registration lists. More than 2.4 million people responded, with 57% indicating that they would vote for Republican Alf Landon in the upcoming Presidential election. (Incumbent Democrat Franklin Roosevelt won the actual election, carrying 63% of the popular vote.)
c) Identify the population of interest and the sample actually used to study that population in this poll.

    population:    sample:

d) Offer an explanation as to how *Literary Digest*’s prediction could have been so much in error. In particular, comment on why its sampling method made it vulnerable to overestimating support for the Republican candidate.

- A sampling procedure is said to be **biased** if it tends *systematically* to overrepresent certain segments of the population and systematically to underrepresent others.

**Example 1-4: Self-Injuries**
An article published in the June 6, 2006 issue of the journal *Pediatrics* describes the results of a survey on the topic of college students’ intentionally injuring themselves. Researchers invited 8300 undergraduate and graduate students at Cornell University and Princeton University to participate in the survey. A total of 2875 students responded, with 17% of them saying that they have purposefully injured themselves.

a) Identify the observational units and variable in this study. Also classify the variable as categorical (also binary?) or quantitative.

b) Identify the population and sample.

    Population:    Sample:

c) What is the sample size in this study?

d) Do you think it likely that this sample is representative of the population of all college students in the world? What about of all college students in the U.S.? Explain.

**Example 1-5: Basketball Sell-Outs**
The 2008-09 Oklahoma City Thunder, a National Basketball Association team in its second year after moving from Seattle, found that their win-loss record was actually worse for home games with a sell-out crowd (3 wins and 15 losses) than for home games without have a sell-out crowd (12 wins and 11 losses). (These data were noted in the April 20, 2009 issue of *Sports Illustrated* in the Go Figure column.)
a) Identify the observational units and the two variables in this study. Also classify each variable as categorical (also binary?) or quantitative.

- **Explanatory variable**: the variable whose effect one wants to study
- **Response variable**: the variable that one suspects is affected by the explanatory variable

b) Identify which variable is explanatory and which is response in this study.

Explanatory:  
Response:

c) Calculate the proportion of wins for each group. When did the team have a higher winning percentage: in front of a sell-out crowd or not?

Sell-out crowd:  
Non-sell-out crowd:

d) Would you conclude that playing in front of a sell-out crowd causes the team to play worse, or can you think of an alternative explanation for what you found above?

- **Observational study**: one in which the researcher passively observes and records information on cases.
  - An observational study may establish an *association* between the explanatory and response variables, but not a *cause-and-effect* relationship.
- **Confounding variable**: one whose effects on the response variable are indistinguishable from the effects of the explanatory variable.
  - Explains why one cannot draw a cause/effect conclusion from an observational study: the groups could differ in more ways than just the explanatory variable.

e) Suggest a confounding variable that plausibly explains the observed relationship.

f) In fact, 22 of the 41 home games for the Thunder were against teams that won more than half of their games that season. Of these 22 games, 13 were sell-outs. Of the 19 games against opponents that won less than half of their games that season, only 5 of those games were sell-outs. Does this support the argument that the Thunder tended to have more sell-outs against “good” teams as opposed to “weaker” teams? Support your answer by calculating relevant proportions.
g) When the Thunder played a “good” team, they lost 18 of 22 games. When they played a “weaker” team, they won 11 of 19 games. Does this support the argument that the Thunder tended to lose more often when they played better teams? Again support your answer by calculating relevant proportions. *Hint:* Be sure to calculate these proportions in a consistent manner.

**Example 1-6: Housework and Divorce**

A 2012 study conducted by the Norwegian Institute for Social Research found that marriages for which the husband and wife share household chores equally are significantly more likely to end in divorce than marriages for which the wife does more housework than the husband.

a) Identify the observational units, explanatory variable, and response variable in this study. Also classify the variables.

Cases:

- Explanatory: 
- Response: 

b) Is it legitimate to conclude from this study that sharing housework equally caused the higher rate of divorce in Norwegian marriages? If so, explain. If not, identify a confounding variable and explain why its effect on the response is confounded with that of the explanatory variable.

c) Do you think that the study’s conclusion (of an association between housework and divorce) applies to marriages outside of Norway? Explain.

Two key questions that should be asked of statistical studies:

- To what population can we reasonably **generalize** the results of this study?
- Can we reasonably draw a **cause-and-effect** connection between the explanatory and response variables?