Introduction to Statistics

GLOSSARY

1. **Data** are collections of observations, such as measurements, genders, or survey responses.
2. **Population**: A population is the complete collection of all measurements or data that are being considered.
3. A **census** is the collection of data from every member of the population.
4. A **sample** is a subcollection of members selected from a population

   Ex: Population: All 38 million carbon monoxide detectors in the United States

   Sample: The 30 carbon monoxide detectors that were selected and tested
5. Critical Thinking

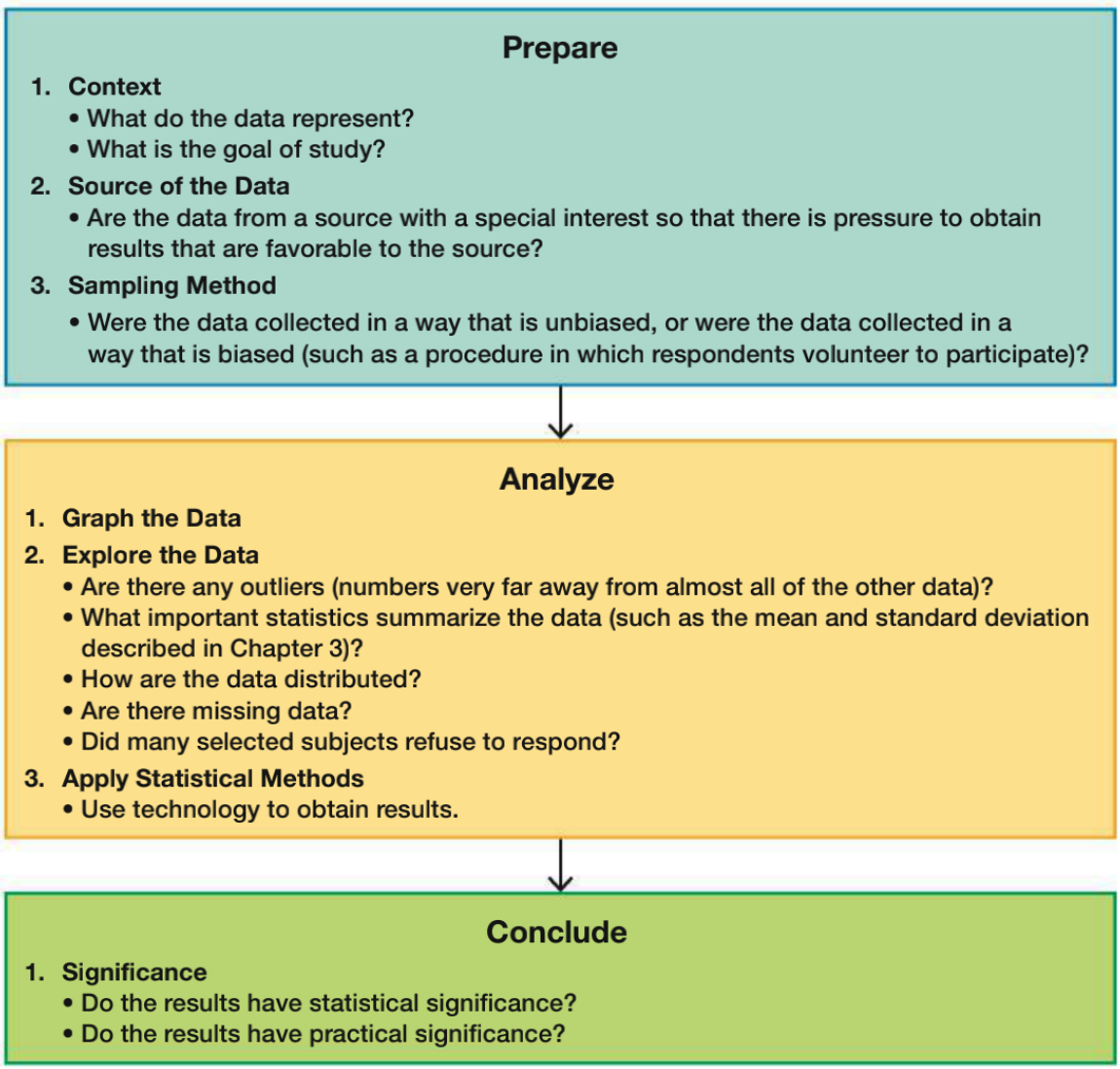


FIGURE 1-2 **Statistical and Critical Thinking**

7. A **voluntary response sample** (or self-selected sample) is one in which the respondents themselves decide whether to be included.

Ex: ■ Internet polls, in which people online can decide whether to respond

■ Mail-in polls, in which people can decide whether to reply

■ Telephone call-in polls, in which newspaper, radio, or television announcements ask that you voluntarily call a special number to register your opinion.

8. Statistical Significance vs Practical Significance

Statistical Significance *Statistical significance* is achieved in a study when we get a result that is very unlikely to occur by chance. A common criterion is that we have statistical significance if the likelihood of an event occurring by chance is 5% or less.

- Getting 98 girls in 100 random births *is* statistically significant because such an extreme outcome is not likely to result from random chance.
- Getting 52 girls in 100 births *is not* statistically significant because that event could easily occur with random chance.

Practical Significance It is possible that some treatment or finding is effective, but common sense might suggest that the treatment or finding does not make enough of a difference to justify its use or to be practical, as illustrated in Example 3.

9. Potential Fitfalls
   a. Misleading Conclusions: Correlation doesn't imply causation
   b. Sample data reported not measured: Asking weights of people instead of actually measuring it on weighing scale
   c. Loaded Questions: If the wording of the question changes the results might change too
   d. Order of questions:

"Would you say that traffic contributes more or less to air pollution than industry?" (45% blamed traffic; 27% blamed industry.)

"Would you say that industry contributes more or less to air pollution than traffic?" (24% blamed traffic; 57% blamed industry.)

   e. Non response: when someone doesn't respond
   f. Percentages:

10. A parameter is a numerical measurement describing some characteristic of a population.

    A statistic is a numerical measurement describing some characteristic of a sample.

{PP SS}

11. Quantitative (or numerical) data consist of numbers representing counts or measurements.

    Categorical (or qualitative or attribute) data consist of names or labels (not numbers that represent counts or measurements).

12. Discrete vs Continuous

DEFINITIONS

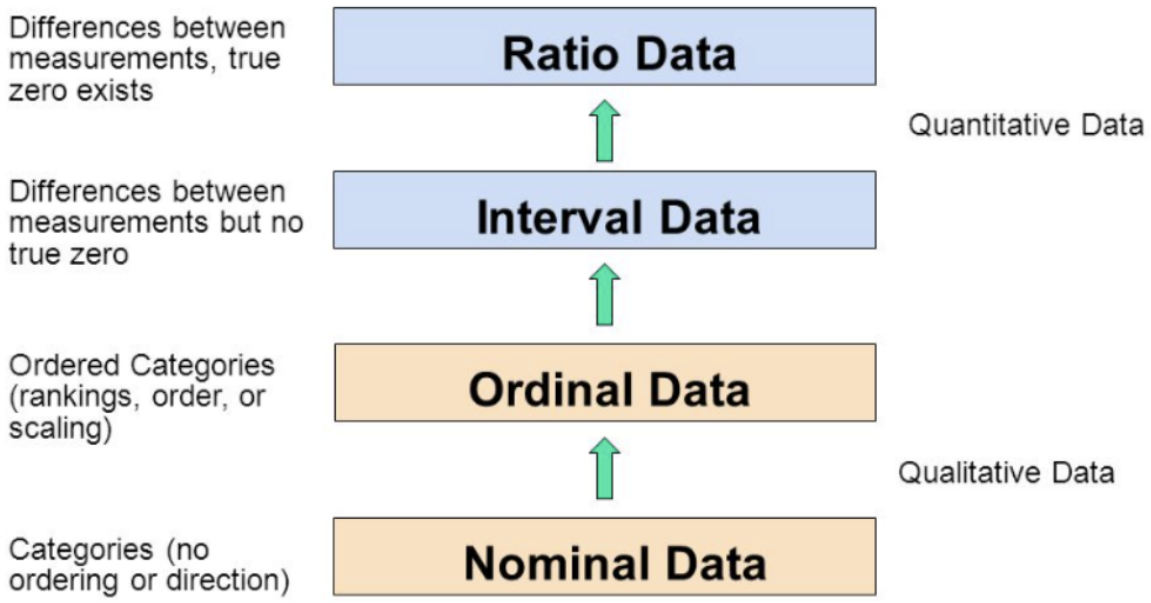
Discrete data result when the data values are quantitative and the number of values is finite, or "countable." (If there are infinitely many values, the collection of values is countable if it is possible to count them individually, such as the number of tosses of a coin before getting tails.)

Continuous (numerical) data result from infinitely many possible quantitative values, where the collection of values is not countable. (That is, it is impossible to count the individual items because at least some of them are on a continuous scale, such as the lengths of distances from 0 cm to 12 cm.)

13. LEVELS OF MEASUREMENT

TABLE 1-2 Levels of Measurement

Level of Measurement	Brief Description	Example
Ratio	There is a natural zero starting point and ratios make sense.	Heights, lengths, distances, volumes
Interval	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless.	Body temperatures in degrees Fahrenheit or Celsius
Ordinal	Data can be arranged in order, but differences either can't be found or are meaningless.	Ranks of colleges in *U.S. News & World Report*
Nominal	Categories only. Data cannot be arranged in order.	Eye colors



The difference between interval and ratio scales comes from their ability to dip below zero. Interval scales hold no true zero and can represent values below zero. For example, you can measure temperature below 0 degrees Celsius, such as -10 degrees. Ratio variables, on the other hand, never fall below zero.

14. Missing Values

    a. A data value is missing completely at random if the likelihood of its being missing is independent of its value or any of the other values in the data set. That is, any data value is just as likely to be missing as any other data value.

    b. A data value is missing not at random if the missing value is related to the reason that it is missing.

14. Observational vs Experimental Study

DEFINITIONS

In an **experiment**, we apply some *treatment* and then proceed to observe its effects on the individuals. (The individuals in experiments are called **experimental units**, and they are often called **subjects** when they are people.)

In an **observational study**, we observe and measure specific characteristics, but we don't attempt to *modify* the individuals being studied.

Experiments are often better than observational studies because well-planned experiments typically reduce the chance of having the results affected by some variable that is not part of a study. A *lurking variable* is one that affects the variables included in the study, but it is not included in the study.

EXAMPLE 2 **Ice Cream and Drownings**

Observational Study: Observe past data to conclude that ice cream causes drownings (based on data showing that increases in ice cream sales are associated with increases in drownings). The mistake is to miss the lurking variable of temperature and the failure to see that as the temperature increases, ice cream sales increase and drownings increase because more people swim.

Experiment: Conduct an *experiment* with one group treated with ice cream while another group gets no ice cream. We would see that the rate of drowning victims is about the same in both groups, so ice cream consumption has no effect on drownings.

Here, the experiment is clearly better than the observational study.

15. Designing of experiments

- **Replication** is the repetition of an experiment on more than one individual. Good use of replication requires sample sizes that are large enough so that we can see effects of treatments. In the Salk experiment in Example 1, the experiment used sufficiently large sample sizes, so the researchers could see that the Salk vaccine was effective.
- **Blinding** is used when the subject doesn't know whether he or she is receiving a treatment or a placebo. Blinding is a way to get around the **placebo effect,** which occurs when an untreated subject reports an improvement in symptoms. (The reported improvement in the placebo group may be real or imagined.) The Salk experiment in Example 1 was **double-blind,** which means that blinding occurred at two levels: (1) The children being injected didn't know whether they were getting the Salk vaccine or a placebo, and (2) the doctors who gave the injections and evaluated the results did not know either. Codes were used so that the researchers could objectively evaluate the effectiveness of the Salk vaccine.
- **Randomization** is used when individuals are assigned to different groups through a process of random selection, as in the Salk vaccine experiment in Example 1. The logic behind randomization is to use chance as a way to create two groups that are similar. The following definition refers to one common and effective way to collect sample data in a way that uses randomization.

17. A simple random sample of n subjects is selected in such a way that every possible sample of the same size n has the same chance of being chosen.

18. Common sampling methods

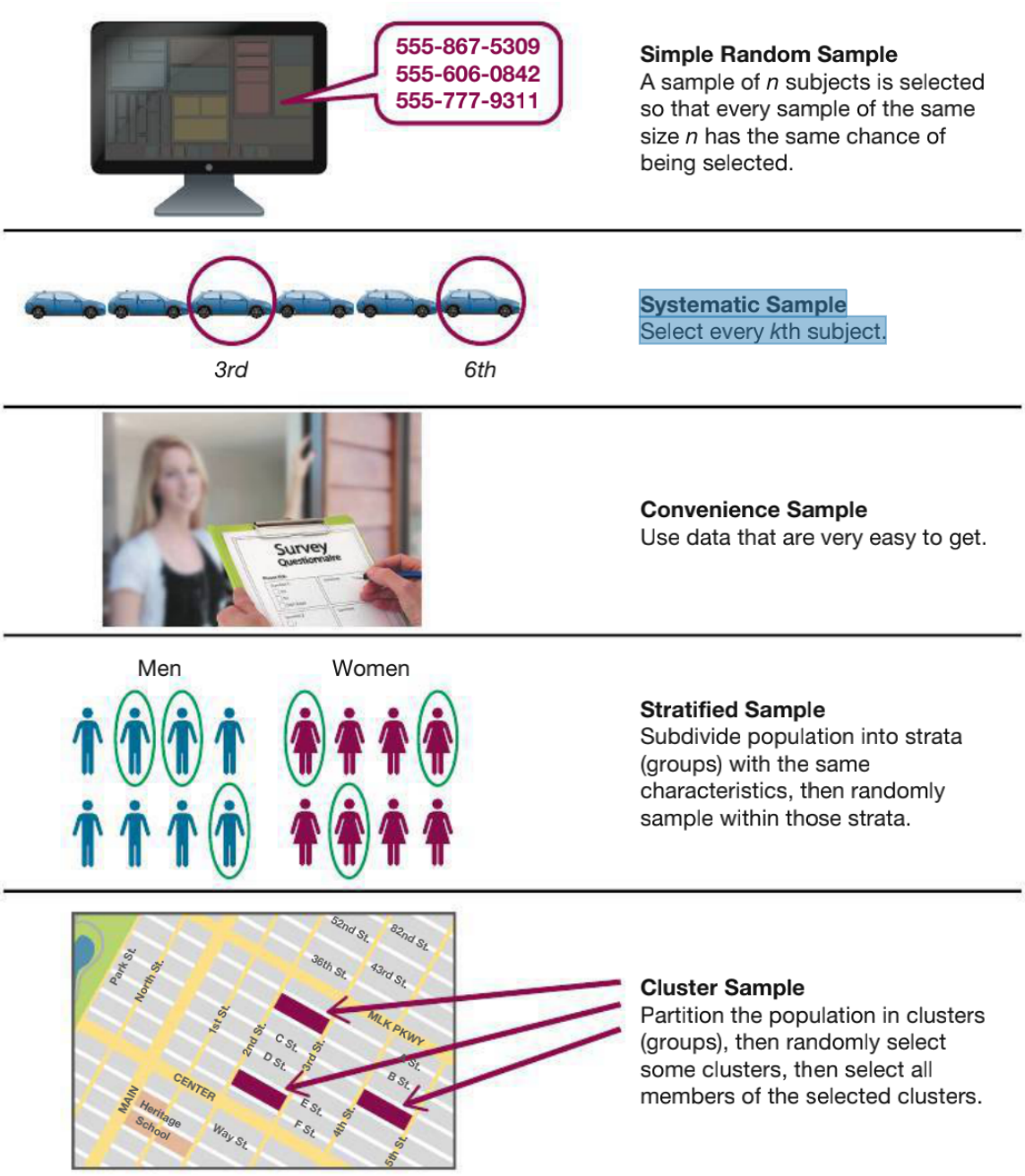
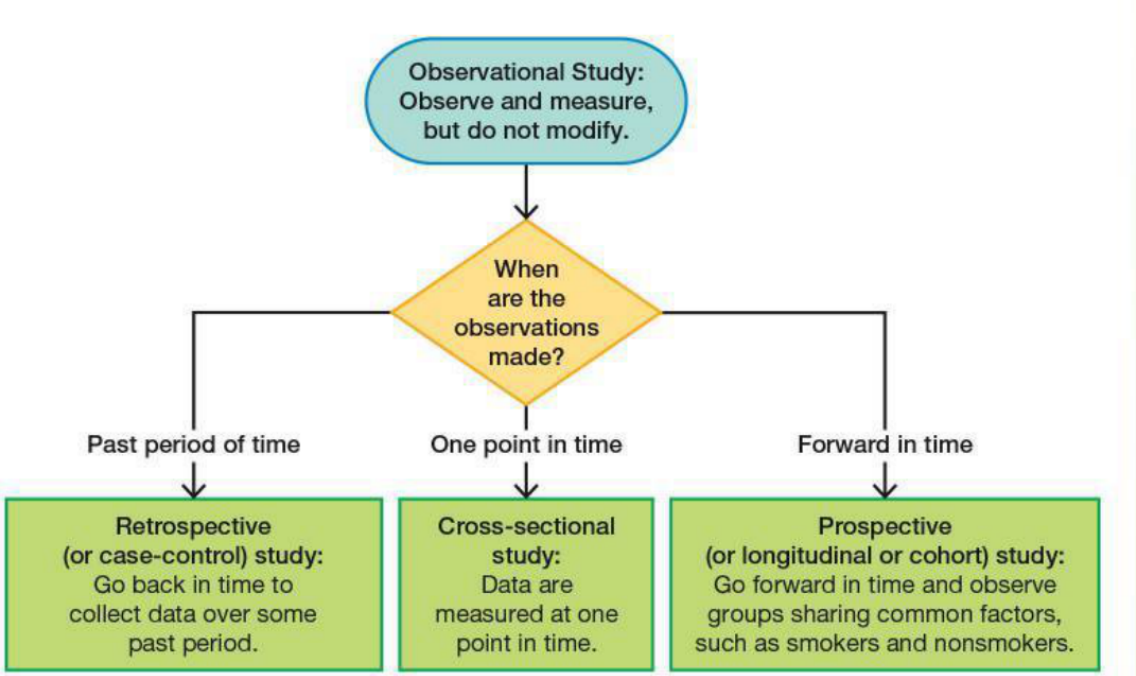


FIGURE 1-3 Common Sampling Methods

19. Types of Observational Experiments



20. Sampling Errors

DEFINITIONS

A **sampling error** (or **random sampling error**) occurs when the sample has been selected with a random method, but there is a discrepancy between a sample result and the true population result; such an error results from chance sample fluctuations.

A **nonsampling error** is the result of human error, including such factors as wrong data entries, computing errors, questions with biased wording, false data provided by respondents, forming biased conclusions, or applying statistical methods that are not appropriate for the circumstances.

A **nonrandom sampling error** is the result of using a sampling method that is not random, such as using a convenience sample or a voluntary response sample.