

AMA 2634

Introduction to Statistics

2024 Fall

Lecture 1

Basics of Statistics

Outline

- 1 Descriptive and Inferential Statistics**
- 2 Variables and Types of Data**
- 3 Data Collection and Sampling Techniques**
- 4 Observational and Experimental Studies**
- 5 Uses and Misuses of Statistics**

Statistics is the science of learning from data

Statistical knowledge helps you to

- collect data (properly)
- employ the correct analyses
- present the results (efficiently)
- draw reasonable conclusions

Statistics allow you to

- evaluate claims based on quantitative evidence
- produce trustworthy analyses and predictions

Statistics is the science of conducting studies to
collect,
organize,
summarize,
analyze, and
draw conclusions from data.

Games of chance

- Mid 18th century studies in probability, motivated by games of chance
 - Led to mathematical treatment of errors of measurement and theory that forms the foundation of statistics

Political science

- Interest in the numerical description of political units led to descriptive statistics
 - Presentation of data in tables and charts, summarization of data by numerical descriptions and graphs



Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions. (Population parameter and statistics)



A **variable** is a characteristic or attribute that can assume different values.

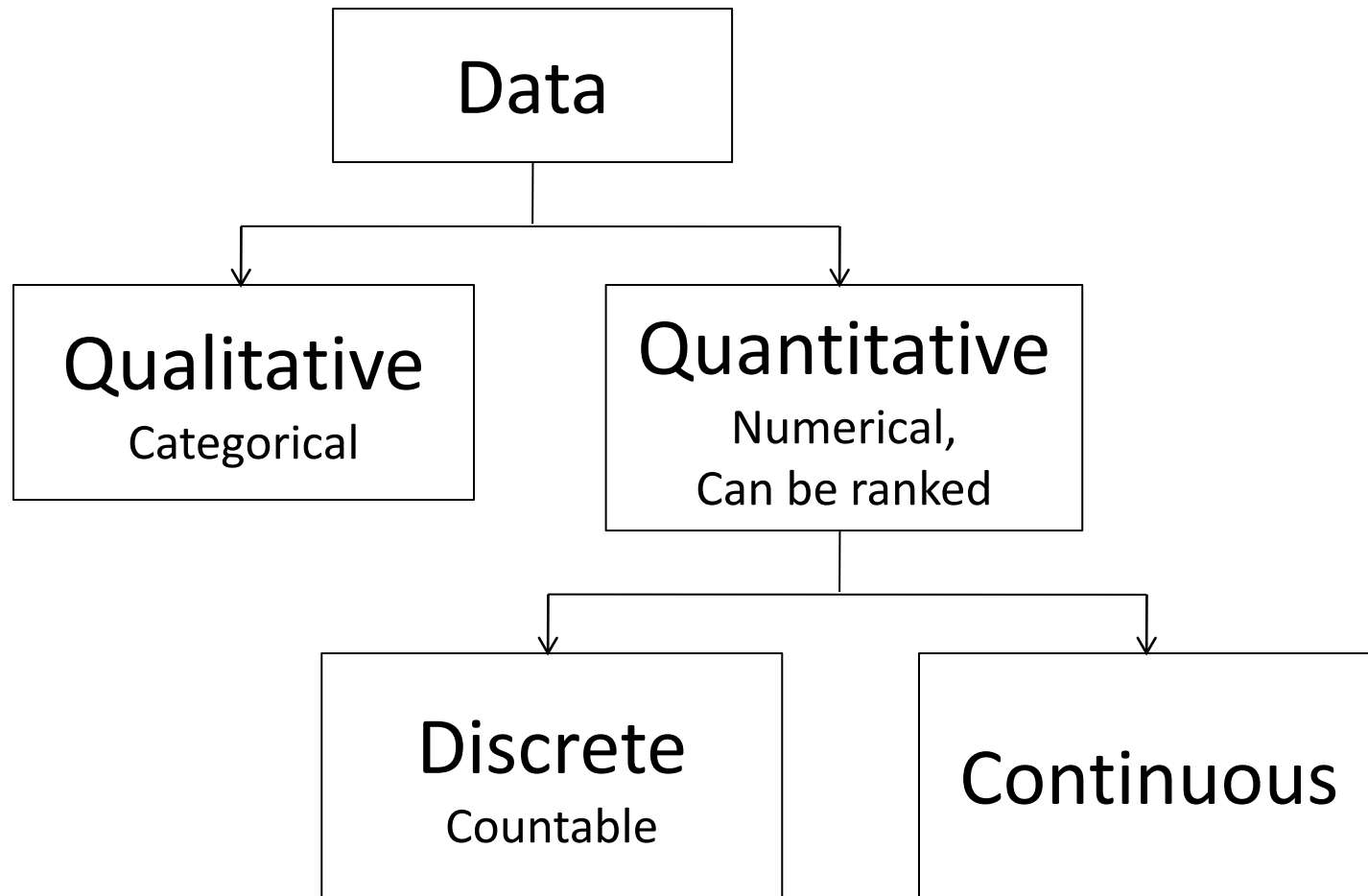
The values that a variable can assume are called **data**.

A **population** consists of all subjects (human or otherwise) that are studied.

A **sample** is a subset of the population.



Population	Unit	Variables
All students major in AMA	Student	GPA, favorite teacher, etc.
All restaurants in campus	Restaurant	Number of employees, hiring vs not hiring, etc.
All books in library	Book	Number of authors, page, etc.





Researcher Elisabeth Kvaavik and others studied factors that affect the eating habits of adults in their mid-thirties. (Source: Kvaavik E, et. al. Psychological explanatory of eating habits among adults in their mid-30's (2005)

International Journal of Behavioral Nutrition and Physical Activity (2)9.)

Classify each of the following quantitative variables considered in the study as discrete or continuous.

- a. Number of children
- b. Household income in the previous year
- c. Daily intake of whole grains (measured in grams per day)

Variable	Recorded Value	Boundaries
Length	15 centimeters (cm)	14.5-15.5 cm
Temperature	86° Fahrenheit (°F)	85.5-86.5 °F
Time	0.43 second (sec)	0.425-0.435 sec
Mass	1.6 grams (g)	1.55-1.65 g

Levels of Measurement

1. **Nominal** – categorical (names)
2. **Ordinal** – nominal, plus can be ranked (order)
3. **Interval** – ordinal, plus intervals are consistent
4. **Ratio** – interval, plus ratios are consistent, true zero

Determine the measurement level.

Variable	Nominal	Ordinal	Interval	Ratio	Level
Hair Color	Yes	No			Nominal
Zip Code	Yes	No			Nominal
Letter Grade	Yes	Yes	No		Ordinal
Temperature (F)	Yes	Yes	Yes	No	Interval
Height	Yes	Yes	Yes	Yes	Ratio
Age	Yes	Yes	Yes	Yes	Ratio

Some Sampling Techniques

Probability sampling

- Simple Random
- Systematic
- Stratified
- Cluster

Non-probability sampling

- Convenient

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an **equally likely chance of occurring**. The sample is then called a **simple random sample**.



A **systematic sample** is obtained by selecting every k^{th} individual from the population. The first individual selected is a random number between 1 and k .



Systematic random sampling

A systematic sample is obtained by selecting every k^{th} individual from the population. The first individual selected is a random number between 1 and k .

Stratified random sampling

A stratified sample is one obtained by separating the population into homogeneous, non-overlapping groups called *strata*, and then obtaining a simple random sample from each stratum.



A quality control engineer wants to obtain a systematic sample of 25 bottles coming off a filling machine to verify the machine is working properly. Design a sampling technique that can be used to obtain a sample of 25 bottles.

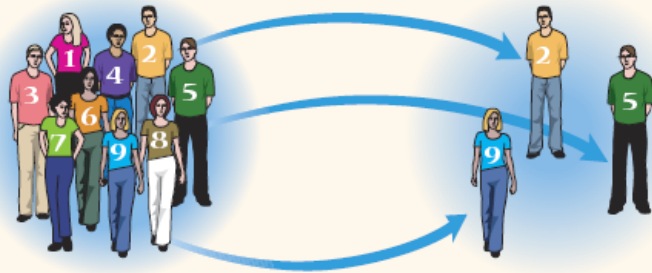
Cluster random sampling

A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.



A school administrator wants to obtain a sample of students in order to conduct a survey. She randomly selects 10 classes and administers the survey to all the students in the class.

Simple Random Sampling



Population

Sample

Stratified Sampling

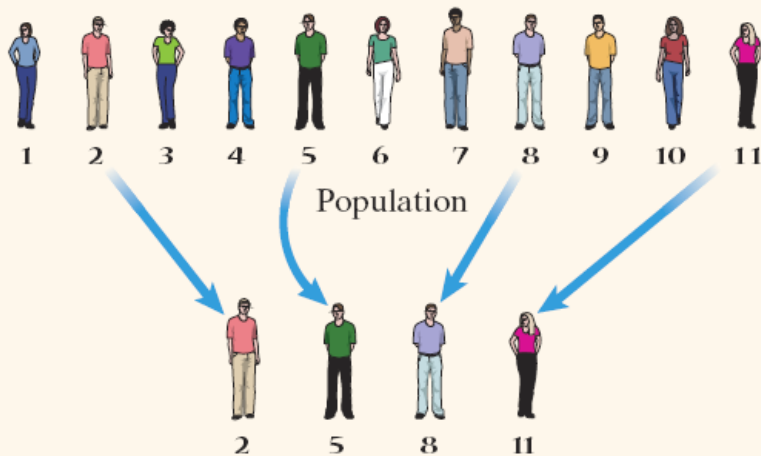


Population

Strata

Sample

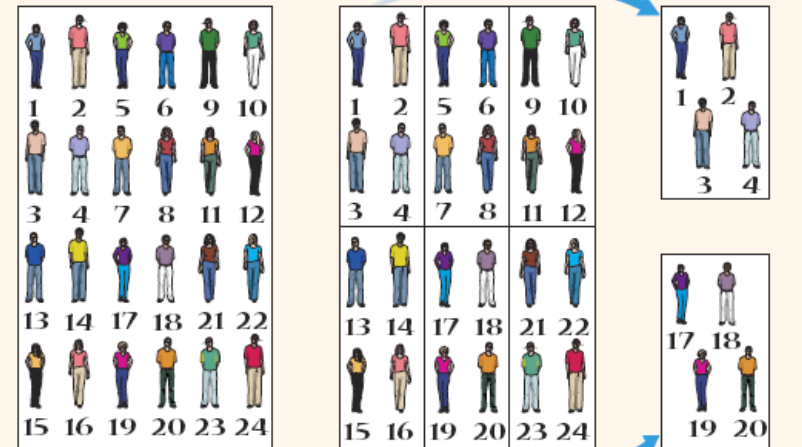
Systematic Sampling



Population

Sample (every 3rd person selected)

Cluster Sampling



Population

Cluster Population

Randomly Selected Clusters

Convenience sampling

- Uses subjects that are convenient
- Eg. Researcher interview subjects entering a local mall in a marketing research
- Non-probabilistic (Sample not representative of the general customers)

In an **observational study**, the researcher merely observes and tries to draw conclusions based on the observations.

The researcher manipulates the **independent (explanatory) variable** and tries to determine how the manipulation influences the **dependent (outcome) variable** in an **experimental study**.

A **confounding variable** influences the dependent variable but cannot be separated from the independent variable.



EXAMPLE 1

Cellular Phones and Brain Tumors

Researchers Joachim Schüz and associates wanted “to investigate cancer risk among Danish cellular telephone users who were followed for up to 21 years.” To do so, they kept track of 420,095 people whose first cellular telephone prescription was between 1982 and 1995. In 2002, they recorded the number of people out of the 420,095 people who had a brain tumor and compared the rate of brain tumors in this group to the rate of brain tumors in the general population. They found no significant difference in the rate of brain tumors between the two groups. The researchers concluded “cellular telephone use was not associated with increased risk for brain tumors.” (*Source*: Joachim Schüz et al. “Cellular Telephone Use and Cancer Risk: Update of a Nationwide Danish Cohort,” *Journal of the National Cancer Institute* 98(23): 1707–1713, 2006)

EXAMPLE 2

Cellular Phones and Brain Tumors

Researchers Joseph L. Roti Roti and associates examined “whether chronic exposure to radio frequency (RF) radiation at two common cell phone signals—835.62 megahertz, a frequency used by analogue cell phones, and 847.74 megahertz, a frequency used by digital cell phones—caused brain tumors in rats.” To do so, the researchers divided 480 rats into three groups. The rats in group 1 were exposed to the analogue cell phone frequency; the rats in group 2 were exposed to the digital frequency; the rats in group 3 served as controls and received no radiation. The exposure was done for 4 hours a day, 5 days a week for 2 years. The rats in all three groups were treated the same, except for the RF exposure.

After 505 days of exposure, the researchers reported the following after analyzing the data. “We found no statistically significant increases in any tumor type, including brain, liver, lung or kidney, compared to the control group.” (*Source*: M. La Regina, E. Moros, W. Pickard, W. Straube, J. L. Roti Roti, “The Effect of Chronic Exposure to 835.62 MHz FMCW or 847.7 MHz CDMA on the Incidence of Spontaneous Tumors in Rats,” Bioelectromagnetic Society Conference, June 25, 2002.)



- **Suspect Samples**
- **Ambiguous Averages**
- **Changing the Subject**
- **Detached Statistics**
- **Implied Connections**
- **Faulty Survey Questions**

The first thing to consider is **the sample that was used in the research study.**

Sometimes researchers use very small samples to obtain information. Several years ago, advertisements contained such statements as “Three out of four doctors surveyed recommend brand such and such.”

If only 4 doctors were surveyed, the results could have been obtained by chance alone; however, if 100 doctors were surveyed, the results might be quite different.



There are **four commonly used measures that are loosely called averages.**

They are the **mean, median, mode, and midrange.**

For the same data set, these averages can differ markedly. People who know this can, without lying, select the one measure of average that lends the most evidence to support their position.

When different values are used to represent the same data.

For example, one political candidate who is running for reelection might say, “During my administration, expenditures increased a mere 3%.” His opponent, who is trying to unseat him, might say, “During my opponent’s administration, expenditures have increased a whopping \$6,000,000.” Here both figures are correct; however, expressing a 3% increase as \$6,000,000 makes it sound like a very large increase. Here again, ask yourself, Which measure better represents the data?

A claim that uses a detached statistic is one in which no comparison is made.

For example, you may hear a claim such as “Our brand of crackers has one-third fewer calories.” Here, no comparison is made. One-third fewer calories than what? Another example is a claim that uses a detached statistic such as “Brand A aspirin works four times faster.” Four times faster than what? When you see statements such as this, always ask yourself, Compared to what?



Many claims attempt to imply connections between variables that may not actually exist.

For example, consider the following statement: “Eating fish may help to reduce your cholesterol.” Notice the words *may help*. There is no guarantee that eating fish will definitely help you reduce your cholesterol.

“Studies suggest that using our exercise machine will reduce your weight.” Here the word *suggest* is used; and again, there is no guarantee that you will lose weight by using the exercise machine advertised.

Another claim might say, “Taking calcium will lower blood pressure in some people.” Note the word *some* is used. You may not be included in the group of “some” people. Be careful when you draw conclusions from claims that use words such as *may*, *in some people*, and *might help*.

Faulty Survey Questions



When analyzing the results of a survey using questionnaires, you should be sure that the questions are properly written since the way questions are phrased can often influence the way people answer them.

For example, the responses to a question such as “Do you feel that the North Huntingdon School District should build a new football stadium?” might be answered differently than a question such as “Do you favor increasing school taxes so that the North Huntingdon School District can build a new football stadium?” Each question asks something a little different, and the responses could be radically different. When you read and interpret the results obtained from questionnaire surveys, watch out for some of these common mistakes made in the writing of the survey questions.