

Chapter 1 Introduction to Statistics

Types of Data (1.2)

Objectives

In this lesson, we will learn how to

- distinguish between a parameter and a statistic;
- classify data as quantitative or categorical;
- distinguish between discrete and continuous quantitative data;
- identify the four levels of measurement: nominal, ordinal, interval, and ratio;
- explain the ideas of big data, data science, and missing data.

1. Key Idea

A major use of statistics is to collect and use sample data to make conclusions about populations.

Big Picture

We usually cannot measure every member of a population, so we collect a sample and use statistics to estimate or make conclusions about population characteristics.

2. Parameter and Statistic

Parameter

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

Statistic

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

Example 1: For each statement, identify whether the number is a parameter or a statistic.

- (a) In a survey of 500 students, 62% said they use online homework regularly.
- (b) The mean age of all students currently enrolled at a college is 21.4 years.
- (c) Among 40 randomly selected cars, the mean speed was 67 mph.

Solution

- (a) Statistic, because the 62% is based on a sample of 500 students.
- (b) Parameter, because the mean describes all students currently enrolled at the college.
- (c) Statistic, because the mean speed is based on a sample of 40 cars.

3. Quantitative and Categorical Data

Quantitative Data

Quantitative data, also called numerical data, consist of numbers representing counts or measurements.

Examples:

- weights of students;
- ages of respondents;
- number of correct answers on a test.

Categorical Data

Categorical data, also called qualitative or attribute data, consist of names, labels, or categories.

Examples:

- gender categories;
- eye color;
- shirt numbers used as labels for athletes.

Example 2: Classify each variable as quantitative or categorical.

- (a) The number of text messages sent in one day.
- (b) The brand of phone a student owns.
- (c) The time needed to complete an exam.
- (d) Student ID numbers.

Solution

- (a) Quantitative; it is a count.
- (b) Categorical; it is a label or category.
- (c) Quantitative; it is a measurement of time.
- (d) Categorical; ID numbers are labels, not actual counts or measurements.

4. Discrete and Continuous Data

Quantitative data can be classified as discrete or continuous.

Discrete Data

Discrete data result when the number of possible values is finite or countable.

Examples:

- number of tosses of a coin before getting tails;
- number of students in a classroom;
- number of defective items in a shipment.

Continuous Data

Continuous data result from infinitely many possible values on a continuous scale.

Examples:

- distance from 0 cm to 12 cm;
- height;
- weight;

- time.

Example 3: Classify each quantitative variable as discrete or continuous.

- (a) Number of absences in a semester.
- (b) Amount of water in a bottle.
- (c) Length of a pencil.
- (d) Number of emails received today.

Solution

- (a) Discrete; it is a count.
- (b) Continuous; it is a measurement.
- (c) Continuous; it is a measurement.
- (d) Discrete; it is a count.

5. Levels of Measurement

Another way to classify data is to use four levels of measurement: nominal, ordinal, interval, and ratio.

Nominal Level

The nominal level consists of names, labels, or categories only. The data cannot be meaningfully arranged in order.

Example: survey responses of yes, no, and undecided.

Ordinal Level

The ordinal level consists of data that can be arranged in order, but differences between values are not meaningful.

Example: course grades A, B, C, D, and F.

Interval Level

The interval level consists of data that can be arranged in order, and differences are meaningful. However, there is no natural zero point where none of the quantity is present.

Example: years such as 1000, 1492, 1776, and 2000.

Ratio Level

The ratio level consists of data that can be arranged in order, differences are meaningful, and there is a natural zero starting point. Both differences and ratios are meaningful.

Example: class times of 50 minutes and 100 minutes.

Summary of Levels of Measurement

Level	Main Idea
<i>Nominal</i>	Categories only
<i>Ordinal</i>	Categories with order
<i>Interval</i>	Meaningful differences, but no natural zero point
<i>Ratio</i>	Meaningful differences and a natural zero point

Example 4: Identify the level of measurement.

- (a) Blood types: A, B, AB, O.
- (b) Finishing places in a race: first, second, third.
- (c) Temperatures in degrees Fahrenheit.
- (d) Weights of backpacks in pounds.

Solution

- (a) Nominal; these are categories with no natural order.
- (b) Ordinal; the places have order, but the differences are not meaningful measurements.
- (c) Interval; differences are meaningful, but 0° F does not mean "no temperature."
- (d) Ratio; weights have a natural zero point, and ratios are meaningful.

6. Big Data and Data Science

Big Data

Big data refers to data sets so large and complex that their analysis is beyond the capabilities of traditional software tools. Analyzing big data may require software running in parallel on many different computers.

Data Science

Data science involves applications of statistics, computer science, software engineering, and other relevant fields such as sociology, business, or finance.

7. Missing Data

Missing Completely at Random

A data value is missing completely at random if the likelihood of its being missing is independent of its value and independent of other values in the data set.

Missing Not at Random

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

Correcting for Missing Data

Two common approaches are:

- Delete cases: remove subjects with missing values.
- Impute missing values: substitute reasonable values for missing

data.

Example 5: Suppose a survey asks students to report their income. Many students with very high income choose not to answer. Is the missing income data likely missing completely at random or missing not at random?

Solution

The missing data are likely missing not at random, because the reason for missingness is related to the value itself. Students with very high income may be less willing to report it.

Summary

Key Ideas

- A parameter describes a population; a statistic describes a sample.
- Quantitative data are numbers that represent counts or measurements.
- Categorical data are labels or categories.
- Discrete data are countable; continuous data are measured on a continuous scale.
- The four levels of measurement are nominal, ordinal, interval, and ratio.
- Missing data should be handled carefully because it can affect conclusions.