

Chapter 1 Data and Statistics

What's statistics?

Statistics is the study or science that deals with collecting, analyzing, presenting and interpreting data. It's part of mathematics.

Data and data set

Data are the raw material of statistics.

Data are the facts and figures that are collected, summarized, analyzed and interpreted.

A data set is the data collected in a particular study.

9. For example: we want to conduct a study of students' study hours and their academic performance. We have the following data set.

A data set of students' study hours and their academic performance

Student	Identification Number	Grade Point Average	Age(in year)	Gender	Rank in Class	Study hours/day
Adam	1234	2.89	17.1	Male	15	3
Bob	8978	2.01	17.6	Male	25	1.5
Jason	6578	3.97	18.4	Male	2	4.5
Mary	2345	3.98	17.9	Female	1	4
Michelle	8901	2.67	18.3	Female	18	2
Paula	7789	2.94	17.5	Female	12	2.2
Webster	6780	3.77	18.2	Male	3	4.5

In a data set, there are Elements, Variables and Observations.

Elements, Variables and Observations

The **elements** are the entities/subjects on which data are collected.

Examples:

- In the table, we collect data on students, so a student is an element
a. How many elements are in the data set? 7
- If we want to study companies R&D expenditure, we collect data on companies, so a company is an element.
- If we study the relationship between class size and student's average GPA. We collect data on different classes, so a class is an element.

A **variable** is a characteristic of interest (necessary for a particular study) for an element. In our example, we are interested in student ID GPA, daily study hour, age and gender, so they are all variables. They are characteristics of interest for a student.

- b. How many variables are in this data set? 6

An **observation** is the set of measurements collected for a particular element

For each student, we collect ID GPA etc. so all the data for one student is one observation

- c. How many observations are in this data set? 7

Qualitative and quantitative data

Qualitative data are labels or names used to identify an **attribute** of each element. Qualitative data are often referred to as categorical data

Ex: Grade level: freshman, sophomore, junior, senior are **qualitative data**

d. Which variables are qualitative in the above dataset?

ID, Gender, Rank

How exactly can we tell a variable is qualitative? Here are some properties of qualitative data.

- Qualitative data are discrete; they can not be broken down into a smaller unit and add additional meaning.
- They may have ordinal meaning, ex., Rank indicates orders
- Arithmetic operations do not make sense for qualitative data
Ex: **ID: 001+002?** The sum of two ID's bears no meaning.
Junior × Senior? Multiplying one grade level by the other does not mean anything:
- Qualitative data can be either numeric or nonnumeric
Student ID, SSN are numeric qualitative variables
Denote grade levels from freshman to senior by 1, 2, 3, 4, you get numeric qualitative data.
In a post office, the mailboxes are numbered from 1 to 1,000. These numbers represent qualitative data.

Quantitative data indicate how many or how much; they can be discrete or continuous.

(1) **discrete** ex. number of children in a family; number of books a person has

- Discrete data is based on counts, and the values can not be subdivided meaningfully.

(2) **continuous** ex. weight, height, distance, earnings

- Continuous data can be meaningfully subdivided into finer and finer increment.
e. which are quantitative in the dataset?
Grade point average, Age, study hours are quantitative
- Quantitative data are always numeric
- Ordinary arithmetic operations, such as addition, subtraction, division, multiplication are meaningful with quantitative data

Ex: We can do difference on age to get **age difference**

We can divide one income by the other to get **income ratio**

Or we can add two incomes together to get **total income...**

Cross-sectional data vs. time series data

Cross-sectional data are data collected at the same or almost the same point in time.

Ex. data on annual medical expenditure of US retirees in 1999

Time series data are data collected over several usually equal spaced time points.

Ex. Data on daily opening prices of stock ABC for the last 360 days

Data on exchange rates from march 1999-march 2006

Population: the set of all the elements of interest in a particular study

Assume you want to research on the SAT scores of 500 college applicants, 500 applicants are a population.

Research the employees' opinion on the new manager in a small company, the overall 20 employees are the population

Sample: a subset of the population

In the SAT score research, if you sample 30 applicants to do your research, then the 30 selected applicants is a sample.

Census: a survey to collect data on the entire population

Sample survey: a survey to collect data on a sample

Descriptive statistics: tabular, graphical, and numerical summaries of data

In a sample of 800 students in a university, 30% are Business majors. The 30% is an example of descriptive statistics.

The average age of a freshman class is 17.5 years, the 17.5 years is also an example of descriptive statistics. It indicates the mean age.

Statistical inference: the process of using data obtained from a sample to make estimates or test hypothesis about the characteristics of a population

In a sample of 100 students in a university, 20, or 20%, are Business majors. Based on the above information, the school's paper reported that "20% of all the students at the university are Business majors." This report is an example of Statistical inference

To learn the average working hours of econ professors in US, a sample of 1000 is drawn and estimate based on the sample is statistical inference.

Exercise: The following shows the temperatures (high, low) and weather conditions in a given Sunday for some selected world cities. For the weather conditions, the following notations are used: c = clear; cl = cloudy; sh = showers; pc = partly cloudy.

City	Hi	Lo	Condition
Acapulco	99	77	pc
Bangkok	92	78	pc
Mexico City	77	57	sh
Montreal	72	56	pc
Paris	77	58	c
Rome	88	68	cl
Toronto	78	61	c

- How many elements are in this data set?
- How many variables are in this data set?
- How many observations are in this data set?
- Name the variables and indicate whether they are qualitative or quantitative.
- For which variables are arithmetic operations appropriate and for which are they not appropriate?