Data and Variables

Sources of Data

Definitions:

- A ***population*** is the entire group of objects or individuals under study, about which information is wanted.

- ***Data*** are observations (such as measurements or survey responses) taken from members of a population.
  - A ***census*** is the collection of data from every element in a population.
  - A ***sample*** is a subset of a population that is actually observed and used to get information.

Caveat: Almost all data we see come from samples!

Reason:

- It would take too much time and money to study the entire population.

- May even not be feasible.

Definitions

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

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

Examples

Determine whether the given value is a parameter or a statistic:

- All current Dickinson students' records are examined and 17% of them are found to be international students.

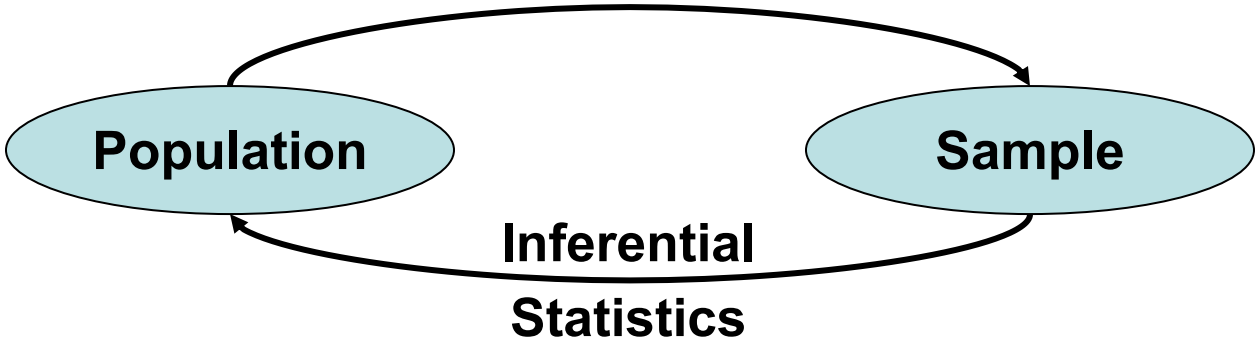
- Several Domino sugar packs are randomly selected, and the average weight of the contents is 3.65 g.

Representative Samples

Statistical inference is when we draw conclusions based on our data.

Fundamental Rule for Using Sample Data for Inference

- Samples can be used to make inferences about a population only if the data can be considered to be *representative* with regard to the question(s) of interest.
- Random sampling



Definitions

- Definition: A ***unit*** or ***object*** is an item we observe. When the unit is a person, we refer to the unit as a ***subject***.

- Definition: An ***observation*** is a piece of information or characteristic recorded for each unit.

- Definition: A characteristic that can vary from unit to unit is called a ***variable***.

- Definition: A set of observations on one or more variables from a collection of units is called a ***data set***.

Variable Types

- Definition: ***Categorical (qualitative)*** variables are those which classify the units into categories. The categories may or may not have a natural ordering to them.
    - Example: Education level (high-school, college,…)

- Definition: A ***numeric*** (***quantitative)*** variable is one that measures a numerical characteristic.
    - Example: Years of education since first grade

  – A ***discrete*** variable can only consist of integer values, e.g., $\mathbb{N}$ = {1, 2, 3, …}
  – A ***continuous*** variable can take on any value within a given interval. e.g., 1.5088039, 5, 693.84975 …

Variable Types

Categorical variables come in two types:

- An **ordinal variable** is a qualitative variable whose categories have a natural ordering to them.
  - Example: Titanic's ticket classes
- A **nominal variable** is a qualitative variable whose categories have no natural ordering to them.
  - Example: Nationality, hometown, gender

Examples for Types of Variables

- Categorical
  - Ordinal
  - Nominal
- Numeric
  - Discrete
  - Continuous

Movie Rating
Yes/No (binary i.e. 0-1 variable)
Eye color
Time to react to a stimulus
Candidate chosen in last election
Melting point
Number of elevators in a building

Examples for Types of Variables

- Categorical
  - Ordinal
  - Nominal
- Numeric
  - Discrete
  - Continuous

Time to run the 100-meter dash
Blood type
Drink size (S M L) at a restaurant
Number of women in the Senate
An individual's ZIP code

Multivariate Data Sets

It is often the case that more than one variable is measured for each unit. That is, each unit has a vector of observations associated with it. These variables can be of any of the types discussed previously.

Examples:

a) Patients have height, weight, and waist size recorded.

Units:

Variables:

b) Students have SAT score, High school average, and First-Year GPA measured.

Units:

Variables:

c) Anthropologists make five measurements on skulls unearthed at a dig site.

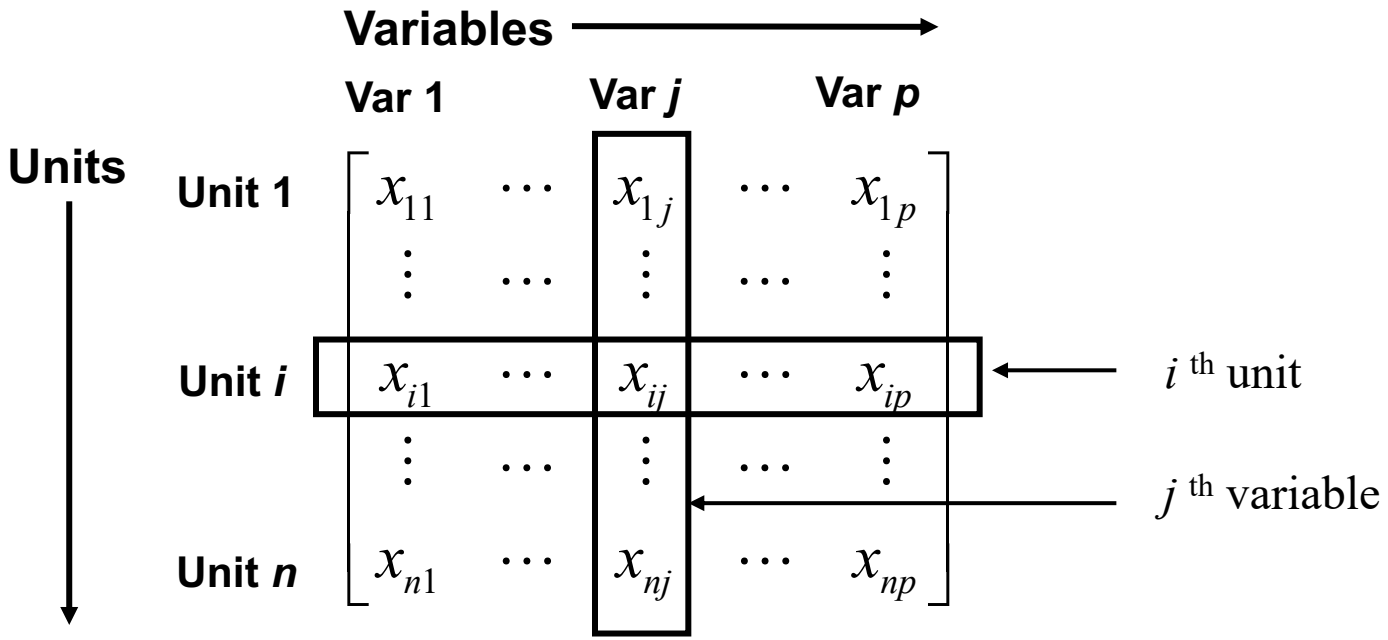
Units:

Variables:

Data Matrices

If we are given data on n units, each possessing p variables of interest. Then the $n \times p$ **data matrix** for this collection is the rectangular array given by

Data Matrix:



Matrices are usually denoted with bold capital letters **A**, **B**, **X**.

Data Matrix

Sample of Crime data for 2007, scaled to rate per 100,000 population.

		Pop	Violent crime	Murder	Rob-bery	Aggra-vated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft	Arson
Buffalo	2007	273,832	1,275	20	560	635	5,893	1,603	3,461	829	43
New York	2007	8,220,196	614	6	265	332	1,819	254	1,403	161	68
Rochester	2007	206,686	1,133	24	497	552	5,419	1,238	3,388	794	98
Syracuse	2007	139,880	1,026	14	319	646	4,264	1,276	2,587	401	34
Yonkers	2007	198,071	443	5	214	202	1,521	324	1,007	190	18

This information can be condensed into the following data matrix.

$$\mathbf{X} = \begin{bmatrix} & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \end{bmatrix}$$

Data Frames in R

The basic structure for storing multivariate data in R is the **data frame**. This is a tightly coupled collection of variables which share many of the properties of matrices and of lists and is used as the fundamental data structure by most of R's modeling software.

A data frame is a matrix-like structure whose columns may be of differing variable types (numeric, factor, logical, and so on).

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
2	John Holland	Boston Celtics	30.0	SG	27.0	6-5	205.0	Boston University	NaN
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
4	Jonas Jerebko	Boston Celtics	8.0	PF	29.0	6-10	231.0	NaN	5000000.0
5	Amir Johnson	Boston Celtics	90.0	PF	29.0	6-9	240.0	NaN	12000000.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
9	Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0

Note: The function `class` can be used to see what the class of a particular object stored in R is (very useful!)