Example: A potato chip company produces salt and vinegar chips at one of its factories. Every night at this factor, they maintain quality control by randomly selecting 20 bags of chips from that day's production. The bags are weighed and the results are recorded in grams.

- Who: 20 bags of salt and vinegar chips
- What: weight (in grams)
- Where: a factory
- When: every night
- Why: quality control : Checking that the weight of each bag is "close" to weight printed on bag.
- How: random sample

Terminology

Data values are often referred to as **observations**.

Data Table: a table used to organize data.

- rows correspond to the individual cases about which (or whom) we have collected data. "who"
- columns correspond to the characteristics that have been recorded in each case. "what", variables

Subject: a unit (person, object, event, transaction) whose characteristics are measured/recorded in a study.

- **participant:** person involved in an experiment.
- **experimental/observational unit:** animals, plants, inanimate subject.
- **respondent:** individual who answers a survey.

Population: the entire set of subjects that we are interested in studying.

"population of interest"

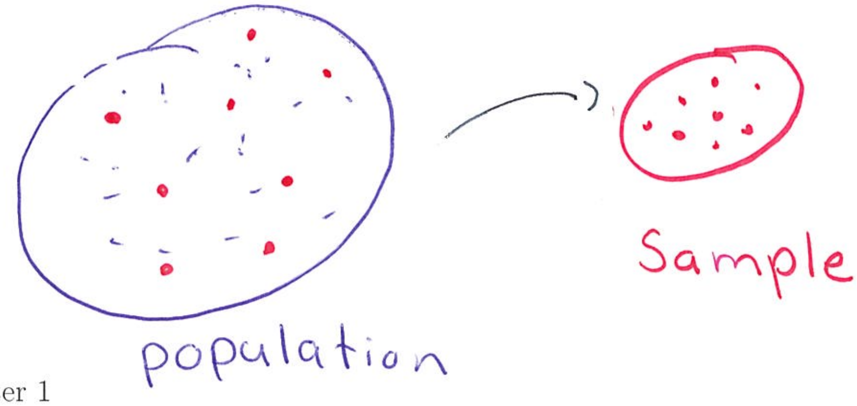
ex1: all Canadian residents

- usually too large/impractical to measure every subject.

Sample: a subset of the population selected to study. Size n

ex 1: Select $n = 2000$ Canadian residents

- used to make inferences about the population
- representative sample, need random sampling



Statistical Inference: an estimate, prediction, or generalization about a population based upon information in a sample.

- reliability

Random Sampling: selecting a sample in such a way that each subject in the population has the same chance of being included in the sample.

Parameter: a numerical summary of population data.

- usually unknown
- Greek letters

Statistic: a numerical summary of sample data.

- known (once data are collected)
- Latin letters

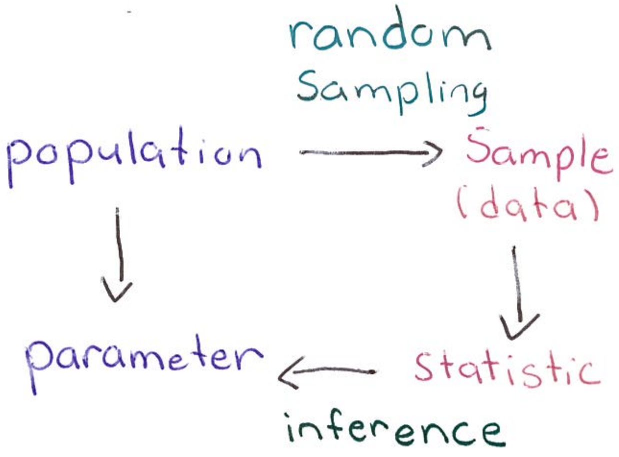
Example:

population: parameter μ = population mean

sample: statistic $\bar{y}$ = sample mean

population: parameter σ = population standard. deviation

sample: statistic s = sample s.d.



Section 1.3: Variables

Statistics is concerned with **variation**.

- change over time
- differences between people / things

Variable: a characteristic recorded about each subject in a study/survey.

"What" - has values: can be different for different subjects

Variables are classified into two main types:

1) **Qualitative or Categorical Variable:**
a variable that puts a subject into one of several groups, categories, or levels.

- Cannot be measured on a numerical scale or counted. (to obtain data)
- Numbers may appear as values, but as a label, not a quantity.

Categorical variables can be subdivided into:

- **Nominal Variable:** a categorical variable whose values are only used to name categories. It's values have no order.

- employment status: employed, not employed
- blood type: O, A, B, AB
- place of birth
- like hockey? yes, no

- **Ordinal Variable:** a categorical variable whose values have an order, but with no natural units.

- place finished in a race: 1st, 2nd, 3rd
- rating: poor, fair, good, excellent
  1, 2, 3, 4
- opinion: SA, A, N, D, SD
- age range: 0-17, 18-65, 66+
  (in years)

→ not for analysis

- **Identifier Variable:** a categorical variable which records a unique value for each subject that is used solely to identify or name it.

- social insurance number (SIN)
- licence plate number

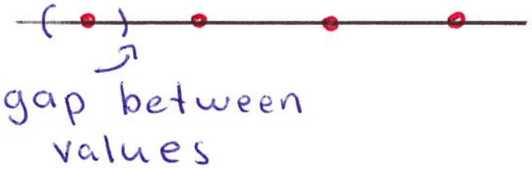
2) **Quantitative or Numerical Variable:** a variable whose values are a numerical quantity obtained from counting or measuring something. -gives the scale of the measurement

- usually has a unit - should state the units

↳ must

Quantitative variables can be subdivided into:

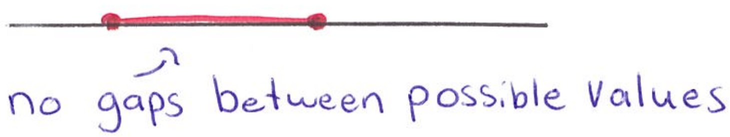
- **Discrete:** a quantitative variable that can only take on a finite (or countable) number of distinct separate values. (usually from counting)



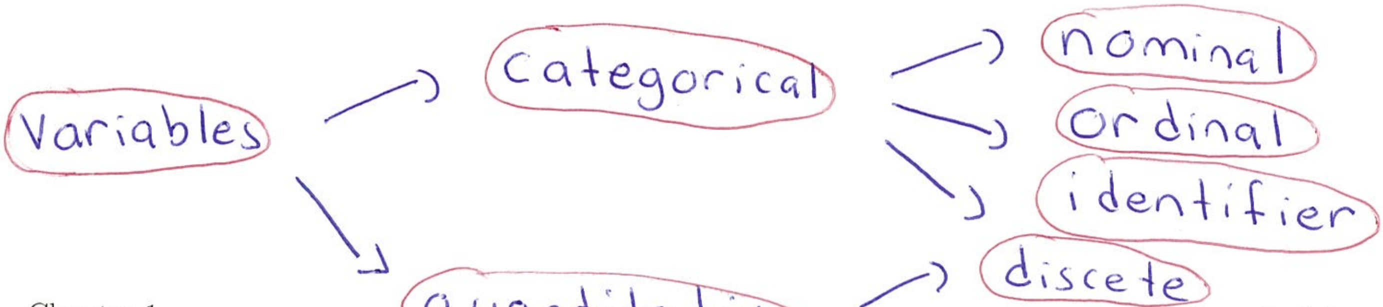
- # of Siblings
- # of cars in a parking lot
- Shoe Size: 5, 5½, 6, 6½, ..

- **Continuous:** a quantitative variable that can take on any real number as a value (or any real number in one or more intervals.) ↳ in theory

(usually from measuring)



- height of trees (meters)
- fuel efficiency (miles per gallon)
- body temperature (degrees Celcius)



Example: For each of the following, indicate the type of variable that records the given values:

1) Brand of calculator (Casio, Sharp,Texas Instrument).

   - categorical, nominal

2) Weights of packages (in grams).

   - quantitative, continuous

3) International Standard Book Number (ISBN).

   - categorical, identifier

4) Taste-test ranking (excellent, good, fair, poor).

   - categorical, ordinal

5) Number of people employed in a profession.

   - quantitative, discrete

6) Time to complete a 5 km race (in minutes).

   - quantitative, continuous

7) Prices of textbooks (in dollars).

   - quantitative, discrete

8) Political affiliation (Liberal, Conservative, NDP, Green, Independent).

   - categorical, nominal

9) Highest level of education completed (high school diploma, bachelor's, master's, Ph.D.)

   - categorical, ordinal

Example:

Data table: students registered in Stat 123 at the U of A in the Fall 2019 term.

Categorical ordinal ↓

Categorical nominal ↓

Categorical nominal ↓

Student ID	Final Grade	Course Mark	Program	Level	Status
1234567	B-	71.6	BSc General - Chem	1st Year	Enrolled
7654321	A	93.2	BA - Psych	3rd Year	Enrolled
2461357	W	-	BSc Honors - Math	2nd Year	Withdrawn

↑ Categorical identifier

↑ quantitative discrete

↑ Categorical ordinal

% of something
→ continuous

rows: "who", cases

columns: "what", variables

Chapter 2: Displaying and Describing Categorical Data

Section 2.1: Summarizing and Displaying a Single Categorical Variable

For a **categorical** variable, there are three important data summaries:

1) **Frequency:** the number of observations in each category.
   → Values, cases

2) **Relative Frequency:** the proportion of observations in each category.

3) **Percentage:** the percentage of observations in each category.

Formulas:

$$\text{Relative Frequency} = \frac{\text{Frequency}}{\text{Total Number of Observations}}$$

$$\text{Percentage} = 100 \times \text{Relative Frequency}$$

Frequency Table: a table that lists each category of a categorical variable along with its frequency.

Category	Frequency
Category labels { — — —	} counts

Relative Frequency Table: a table that lists each category of a categorical variable along with its relative frequency or percentage.

Category	Relative Frequency
Category labels {	} proportion

Example: A popular candy (Brand X) comes in four colours: green, orange, red, and yellow. A bag of this candy was opened and found to contain 24 candies. The colour of each candy has been recorded in the following table:

red	red	green	red	yellow	red
orange	orange	green	red	green	yellow
red	orange	orange	yellow	red	green
green	yellow	yellow	red	yellow	red

	Green	Orange	Red	Yellow
Tally:	𝍸	𝍷𝍷𝍷𝍷	𝍸 𝍷𝍷𝍷𝍷	𝍸 𝍷

Use this data to fill in the following table:

Frequency Table of the Candy Colours in a Bag of Brand X:

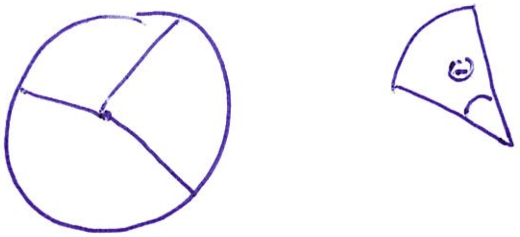
Colour	Frequency	Relative Frequency	Percentage
Green	5	0.21	21%
Orange	4	0.17	17%
Red	9	0.38	38%
Yellow	6	0.25	25%
Total	24	1	100%

Green: $\frac{5}{24} = 0.21$

$0.21 \times 100 = 21\%$

The distribution of the data obtained for a single categorical variable is often displayed graphically in a:

1) **Pie Chart:** a type of graph in which a circle has been sliced into pieces whose sizes are proportional to the relative frequencies of the categories.



<-> one category

- for each category, the angle of its slice can be computed using the formula:

$$\text{angle} \; = \text{relative frequency of category} \times 360^\circ$$

- can label pie chart using either frequency, relative frequency, or percentage. - clearly label

- useful for visualizing how each category relates to the entire data set.

  proportion of set

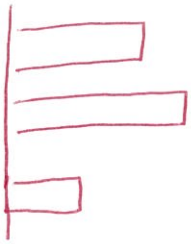
  ↳ as opposed to each other.

- to display the relative frequency distribution of a categorical variable using a pie chart:
  - each subject must fall into exactly one category.
  - the sum of the relative frequencies of all categories must equal 1.
  - the sum of the percentages of all categories must equal 100%.

2) **Bar Chart:** a type of graph which displays the distribution of a categorical variable.

leave a space between columns

or rows



- It consists of separate columns (or bars) that satisfy:
  - the columns are aligned along a common base. → x-axis
  - the widths of the bases of the columns are equal.
  - the height of each column is proportional to the (relative) frequency of the category.
  - each column is centered over its category label.

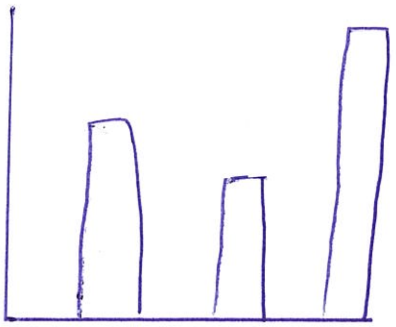
y-axis

- vertical axis can be label with either the frequency, relative frequency, or percentage.

relative frequency bar chart

- useful for comparing categories to each other.

→ not for seeing as part of whole.



Category labels