MATH 201: Lecture 1b Handout

Section 1.2 Data Basics

Name:_____ Date:_____

Let's take a look at data on our solar system. It is organized into a **data matrix**, where:

- each row is called a _____.
- each column is called a _____.

name	mass	length_of_day	mean_temp	n_moons	ring_system	surface_pressure
Mercury	0.330	4222.6	positive	0	FALSE	0.00
Venus	4.870	2802.0	positive	0	FALSE	92.00
Earth	5.970	24.0	positive	1	FALSE	1.00
Mars	0.642	24.7	negative	2	FALSE	0.01
Jupiter	1898.000	9.9	negative	95	TRUE	NA
Saturn	568.000	10.7	negative	146	TRUE	NA
Uranus	86.800	17.2	negative	28	TRUE	NA
Neptune	102.000	16.1	negative	16	TRUE	NA

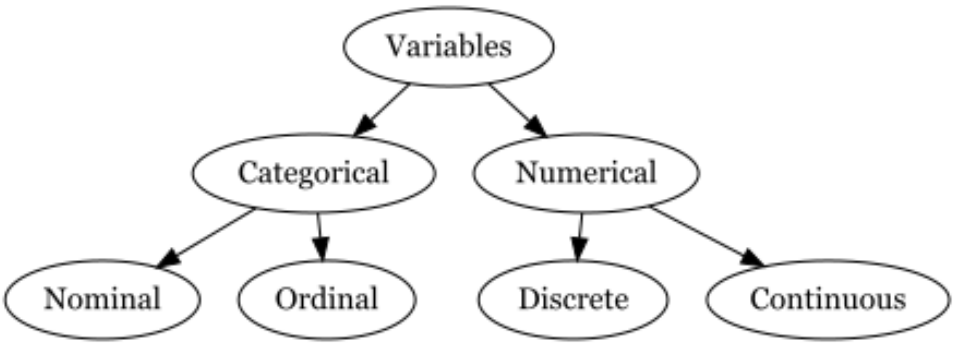
It is always good to look for a description of your data to better understand how the data was collected and on whom/what.

variable	description
name	Name of planet
mass	Mass in 10^24 kg
length_of_day	Length of day in hours
mean_temp	Whether mean temperature in Celsius is positive or not
n_moons	Number of moons
ring_system	Whether the planet has a set of rings around it
surface_pressure	Surface pressure in bars

Describe some patterns or features you notice about this data.

What do you think the `NA` in the `surface_pressure` column means?

Variable types



A _____ variable (sometimes called quantitative) has numeric meaning and we can do mathematical operations with it.

- A _____ variable is a type of _____ variable with a finite or countable number of possibilities.
- A _____ variable is a type of _____ variable with an infinite number of possibilities.

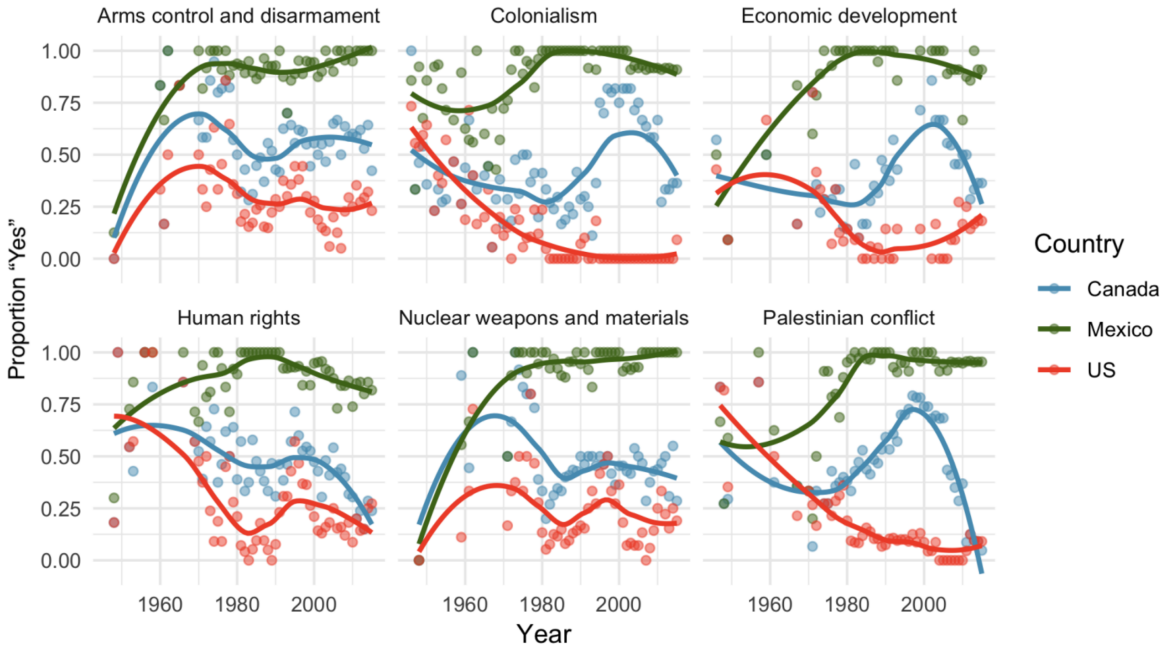
A _____ variable (sometimes called qualitative) does not have numeric meaning, even if it is recorded as a number, and we cannot do mathematical operations with it.

- A _____ variable is a type of _____ variable with a natural ordering to its possible values.
- A _____ variable is a type of _____ variable without a natural ordering to its possible values.

For each variable in the `planets` dataset determine its type.

OpenIntro Statistics Exercise 1.12: UN Votes

The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue.



Try to draft the data matrix used to produce this visualization, specifically write out the column names and fill in at least four rows.

Indicate each variable type.

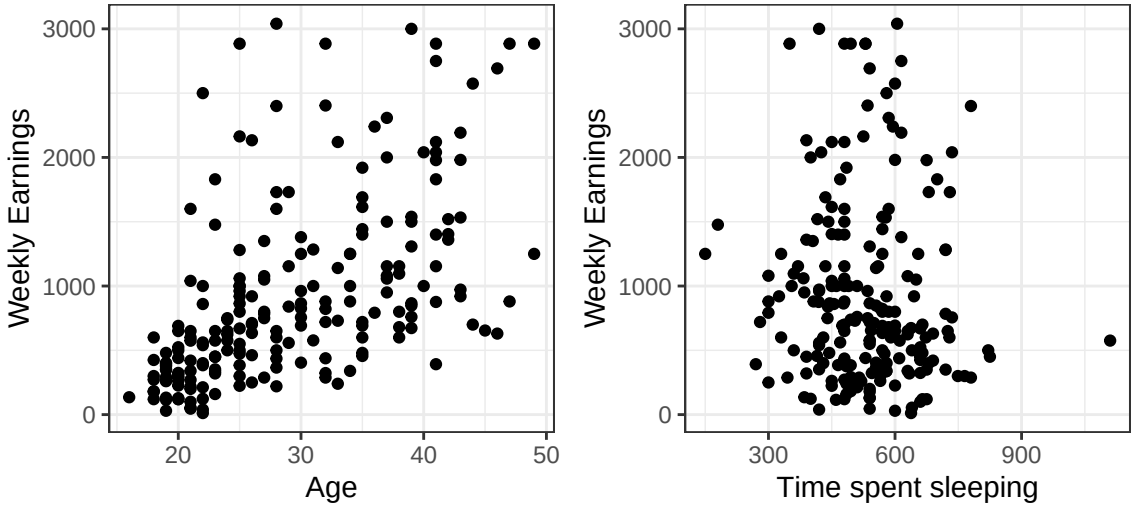
Relationships between variables

If two variables

- have a correspondence, we say they are **associated**.
- do not have a correspondence, we say they are **not associated**.

Can you think of an unrecorded variable that is likely associated with the mass of a planet or having a ring system?

The U.S. Bureau of Labor Statistics publishes the American Time Use Survey which has been subset to include only respondents who are enrolled in college or university.



Does there appear to be an association between weekly earnings and age? If so, would you say it is a **positive association**, meaning that an increase in one variable is associated with an increase in the other, or a **negative association**, meaning increase in one is associated with a decrease in the other?

Does there appear to be an association between weekly earnings and time spent sleeping? If so, is it positive or negative?

If we want to investigate if a change in a variable causes another we call the first variable an **explanatory variable** and the second one the **response variable**.

explanatory variable $\rightarrow$ response variable

Which would you label response and explanatory between weekly earnings and age?

Study types: Experiment versus observational study

In an _____ the researchers do not cause any intervention, they collect data in a way that does not directly interfere with how the data arise.

In an _____ the researchers cause an intervention and then collect data to investigate possible causal relationships.

OpenIntro Statistics Exercise 1.4: Buteyko method

The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.

Identify the main research question of the study.

Was this an experimental or observational study? Is there potential for a causal conclusion?

Who are the subjects (observational units) in this study, and how many are included?

OpenIntro Statistics Exercise 1.3 Air pollution and birth outcomes

Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM10) in g/m3. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM10 and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.

Identify the main research question of the study.

Was this an experimental or observational study? Is there potential for a causal conclusion?

Who are the subjects in this study, and how many are included?

Self checkin

How are you feeling about today's lecture? What do you feel you understand the best from today's content and what do you feel you understand the least? Was any of it already familiar to you?