Data Basics

Ryan Miller

Grinnell College
Statistics

Motivation

Question 1: What percentage of the world's 1-year-old children have been vaccinated against at least one disease?

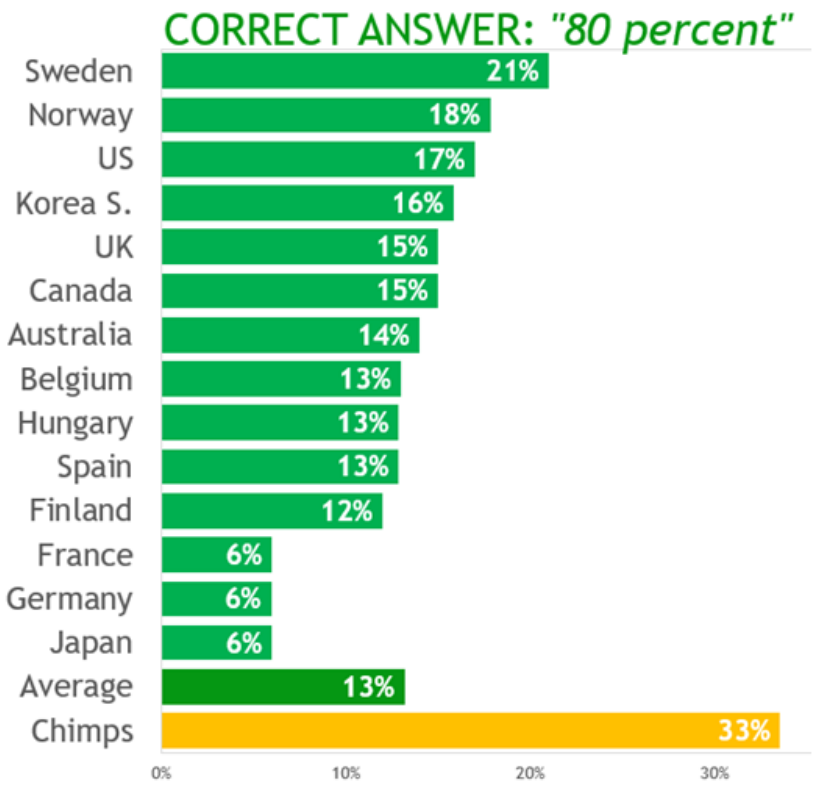
A) 20%
B) 50%
C) 80%

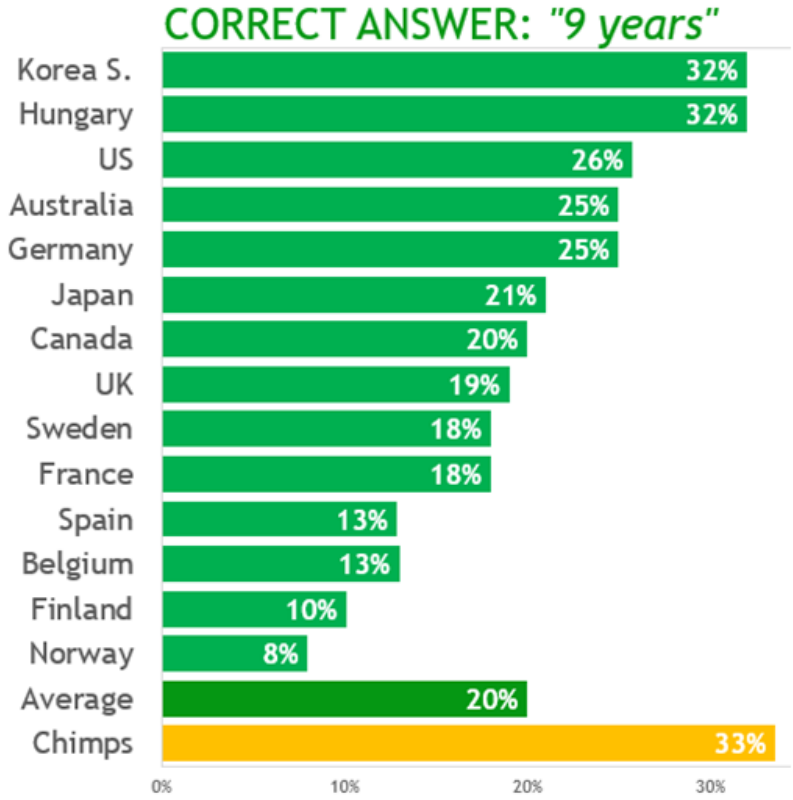
Question 2: Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

A) 3 years
B) 6 years
C) 9 years

Bonus Question: Why do we need data?

Answers





Humans are bad at estimating things and prone to all sorts of biases. Relying upon data can help us overcome these challenges.

Data Basics

- **Data** is defined as "a collection of discrete or continuous values that convey information"
  - This is a very broad definition, we'll largely restrict our attention to "tidy data" or "tabular data"
- A "tidy" data set is organized such that each row represents an *observation/case* and each column represents a *variable*
  - An **observation** or **case** is defined to be a single unit of analysis (ie: person, subject, etc.)
  - A **variable** is any characteristic or attribute that is recorded for each case

Tidy data

Below is an example of "tidy data" obtained from the Washington Post's database of police shootings:

name	date	age	race	armed	city	state	body_camera
Tim Elliot	1/2/2015	53	A	armed	Shelton	WA	FALSE
Lewis Lee Lembke	1/2/2015	47	W	armed	Aloha	OR	FALSE
John Paul Quintero	1/3/2015	23	H	unarmed	Wichita	KS	FALSE
Matthew Hoffman	1/4/2015	32	W	armed	San Francisco	CA	FALSE
Michael Rodriguez	1/4/2015	39	H	armed	Evans	CO	FALSE
Kenneth Joe Brown	1/4/2015	18	W	armed	Guthrie	OK	FALSE
Kenneth Arnold Buck	1/5/2015	22	H	armed	Chandler	AZ	FALSE
Brock Nichols	1/6/2015	35	W	armed	Assaria	KS	FALSE
Autumn Steele	1/6/2015	34	W	unarmed	Burlington	IA	TRUE
Leslie Sapp III	1/6/2015	47	B	armed	Knoxville	PA	FALSE
Patrick Wetter	1/6/2015	25	W	armed	Stockton	CA	FALSE
Ron Sneed	1/7/2015	31	B	armed	Freeport	TX	FALSE

In this data set, the cases are individual people who were killed by the police, and the variables describe characteristics of each case.

Types of variables

There are different types of variables, and the statistical methods we use will differ by variable type.

- **Categorical Variables** divide the cases into *groups*
  - **Binary** - two mutually exclusive categories
  - **Nominal** - many groups with no natural ordering
  - **Ordinal** - groups with a natural order
- **Quantitative Variables** record a *numeric* value for each case
  - **Discrete** - countable (ie: integers)
  - **Continuous** - uncountable (ie: real numbers)

For which of these types could you calculate an average? Are there any where you could calculate a median but *not* an average?

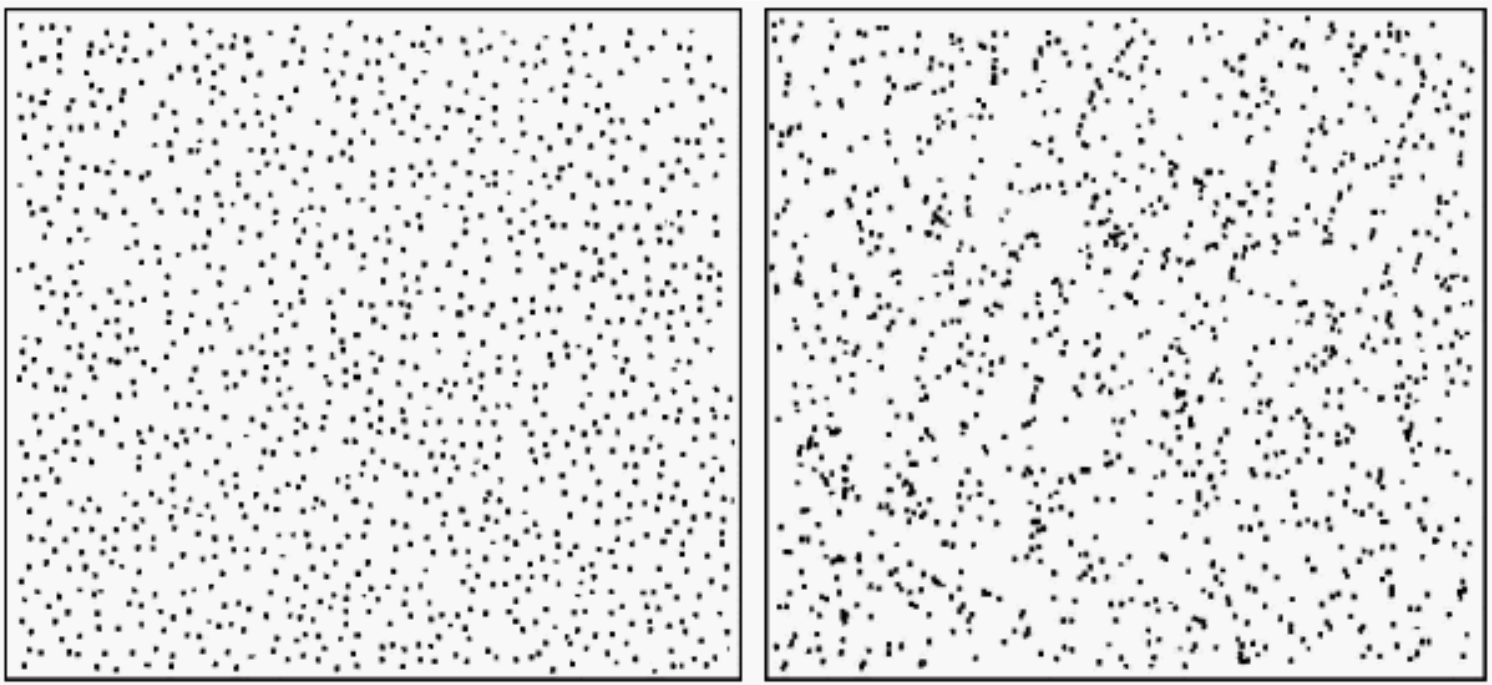
Grey areas

Sometimes a very might technically fit the description of a certain type, but it's more useful to analyze it another way.

- Suppose I collect data on the expected graduation date of every student in this class
  - What type of variable is this? Why might analyzing it another way be useful?
- Suppose I collect survey data using a series of 7-point Likert scales to measure different personality traits
  - What type of variable is this? Why might analyzing it another way be useful?

Data vs. statistics

- Statistics (as a field) is about more than just data, it is about the *uncertainty* present in data
  - To foreshadow why we need statistics, which of these do you think reflects a real biological pattern and which one is randomly generated?



Conclusion

After today's lecture and lab you should be able to:

1. Define and identify *data*, *cases*, and *variables*
2. Classify variables as *categorical* or *quantitative* (numeric)
3. Read data stored in a CSV file on the web into `R` and report basic attributes such as the number of cases/variables it contains