Data and Statistics

Ryan Miller

Data

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

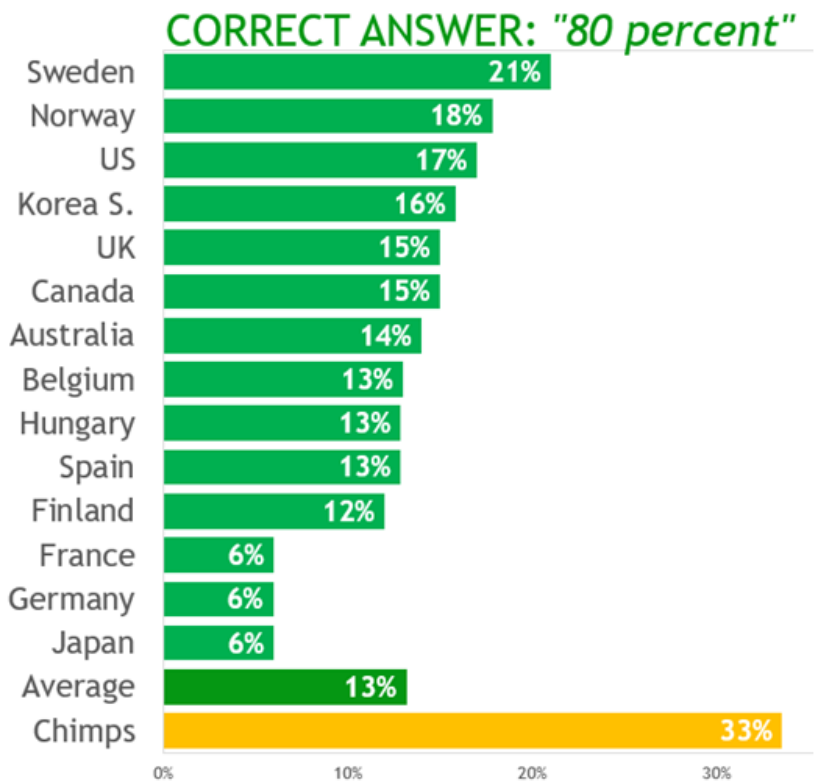
A) 20%
B) 50%
C) 80%

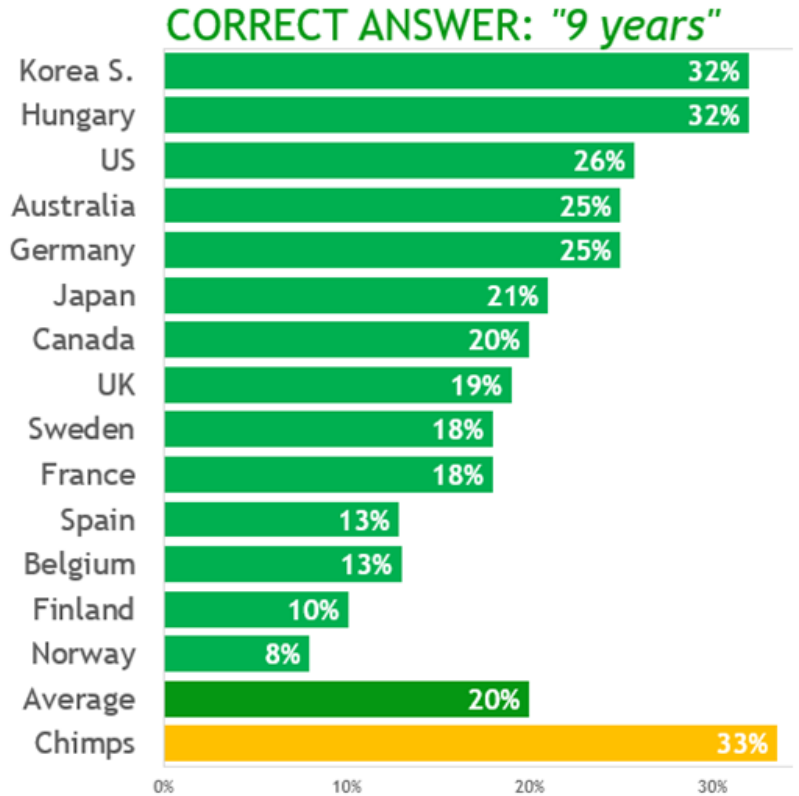
Question 2: Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

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

Data

Here's what the data show:





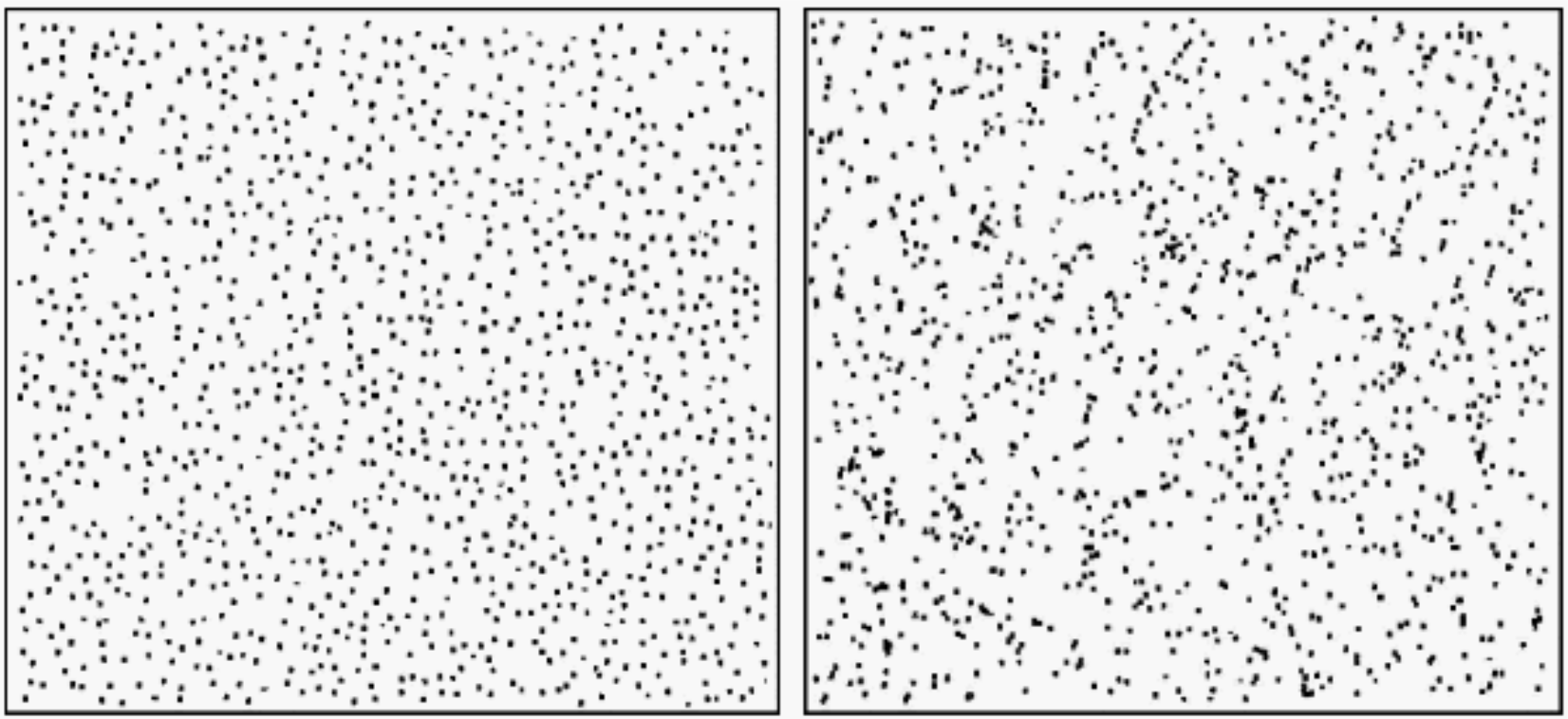
Source: Allan Rossman's JSM talk

Data

- The world has made remarkable progress in the last 20 years
  - Due to biases and a lack of exposure to quality data, most people aren't away of this
- Data empowers us to *objectively understand reality*

What about statistics?

- In most situations simply having data isn't enough, humans are too good at finding non-existent patterns
  - Which panel do you think displays randomly generated data?



What about statistics?

- Statistics is often defined as the science of *understanding uncertainty*
  - More specifically, it's a way of thinking combined with collection of tools and methods that can be used to understand uncertainty in order to make judgements about the world

What about statistics?

- Statistics is often defined as the science of *understanding uncertainty*
  - More specifically, it's a way of thinking combined with collection of tools and methods that can be used to understand uncertainty in order to make judgements about the world
  - ie: What can we learn from experiment that used only 30 people? What can we learn from an poll of 1000 registered voters?

What about statistics?

- Statistics is often defined as the science of *understanding uncertainty*
  - More specifically, it's a way of thinking combined with collection of tools and methods that can be used to understand uncertainty in order to make judgements about the world
  - ie: What can we learn from experiment that used only 30 people? What can we learn from an poll of 1000 registered voters?
- But before we can get to answer these questions, we need to learn the vocabulary of Statisticians

Vocabulary

- **Case**: the subject/object/unit of observation
  - Usually data is organized so that each case is represented by a *row* (but not always!)
- **Variable**: any characteristic that is recorded for each case (generally stored in a *column*)

Vocabulary

- **Case**: the subject/object/unit of observation
  - Usually data is organized so that each case is represented by a *row* (but not always!)
- **Variable**: any characteristic that is recorded for each case (generally stored in a *column*)
- **Categorical Variable**: a variable that divides the cases into *groups*
  - **Nominal**: many categories with no natural ordering
  - **Binary**: two exclusive categories
  - **Ordinal**: categories with a natural order
- **Quantitative Variable**: a variable that records a *numeric* value for each case
  - **Discrete**: countable (ie: integers)
  - **Continuous**: uncountable (ie: real numbers)

Practice

1) Download and open the "Happy Planet" dataset from our course website or this link
2) Identify the cases
3) What type of variable is "Population"?
4) What type of variable is "Region"?

Practice (solution)

- Each case is a country
- "Population" is a quantitative variable, it is measured in millions of people (a numeric entity)
- "Region" is categorical variable, it divides the cases into 7 geographic groups (categories)

Grey Areas

Sometimes there are situations where a variable is technically one type, but it more useful to analyze it as if it were another. For example:

- "Year" might be a discrete quantitative variable, but if the data only contain 2 or 3 years we might treat it is as categorical
- A Likert Scale question is be an ordinal categorical variable, but we might translate it into numeric scores and treat it is a quantitative

Grey Areas

Sometimes there are situations where a variable is technically one type, but it more useful to analyze it as if it were another. For example:

- "Year" might be a discrete quantitative variable, but if the data only contain 2 or 3 years we might treat it is as categorical
- A Likert Scale question is be an ordinal categorical variable, but we might translate it into numeric scores and treat it is a quantitative

"An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem." - John Tukey (Statistician, 1915-2000)