Introduction

Grinnell College

January 24, 2024

STA-209

A brief outline of the class

1. Describe data and variable relationships
   - Univariate and bivariate relationships (numerical and graphical)
   - Multivariate relationships (confounding)
2. Estimation
   - Populations vs Samples
   - Confidence intervals
3. Hypothesis Testing
   - z-test
   - t-test
   - Chi-square tests
4. Statistical Models
   - Regression

What are you learning today?

Why do we need statistics?

Could you describe the statistical framework to your mother?

What is an observation and how do we describe its characteristics?

What types of variables are there, and when is each appropriate?

Two questions

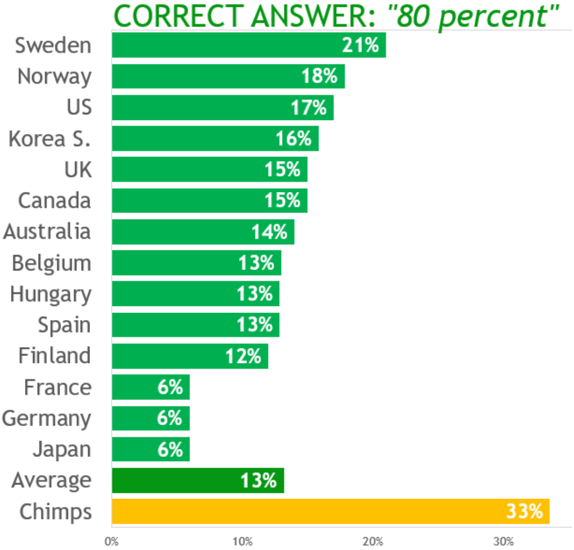
Question 1: What percentage of the world's 1-year-old children have been vaccinatd against at leat one disease?

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

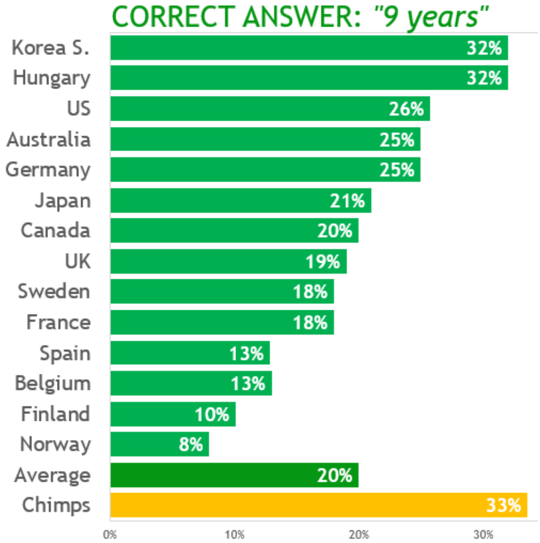
Question 2: Worldwide, 30-year-old men have 10 years of schooling, on average. How many years do women of the same age have?

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

Vaccination

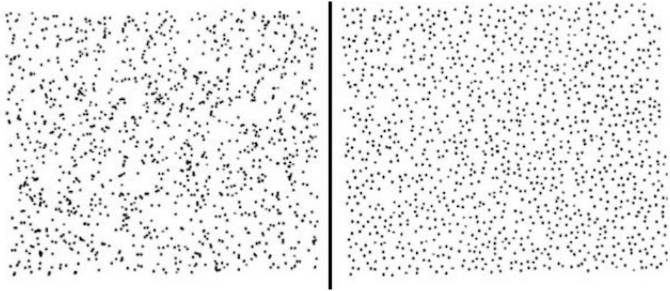


School


CORRECT ANSWER: "9 years"
Korea S.
32%
Hungary
32%
US
26%
Australia
25%
Germany
25%
Japan
21%
Canada
20%
UK
19%
Sweden
18%
France
18%
Spain
13%
Belgium
13%
Finland
10%
Norway
8%
Average
20%
Chimps
33%
0%
10%
20%
30%

Dots

Which of these boxes do you think reflects true randomness, and which of these seems artificially contrived?



Why do we need statistics?

Human beings are great at identifying patterns

- Cognitive biases
- Poor understanding of uncertainty

Statistics as a discipline is about the *quantification of uncertainty*.

1. Construct a hypothesis
2. Collect data
3. Consider evidence
4. Draw conclusions

Populations and Parameters

A **population** is a constrained set of events or subjects about which we wish to ask a scientific question

A **parameter** is a *quantifiable* attribute of a population. It is often assumed to be a fixed or immutable quality within the bounds set by the population

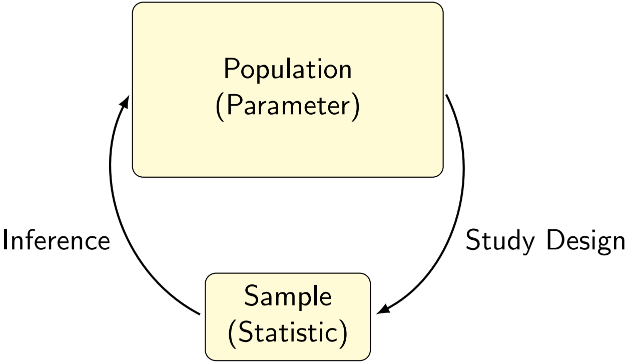
To determine the value of a parameter within a population with certainty is to conduct a **census**

Samples and Statistics

A **sample** is (often) a much smaller, *randomly collected* subset of a larger population

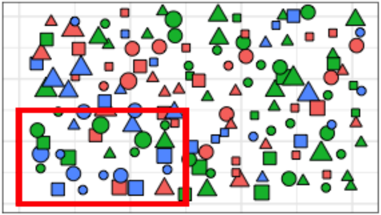
A **statistic** is an *estimate* of a parameter derived from data collected within the sample

The Statistical Framework

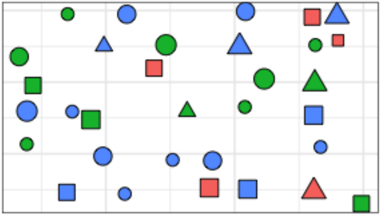


Population and Samples

Population



Sample



An example

Suppose we are interested in determining the average height of students currently enrolled at Grinnell College

Does it matter *how many* students we sample?

Does it matter *which* students we sample?

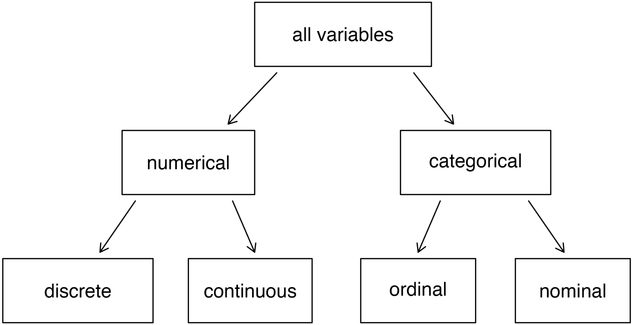
How much confidence do we have that our estimate of the average height is close to correct?

Some definitions

In this course we will primarily be working with data derived from *observations*, our most basic unit of study. Characteristics of an observation are known as **variables**. Variables typically come in one of two types:

1. **Quantitative Variable:** Typically data that is stored in the form of *numbers*, and is numerical in nature
   - Continuous data i.e., height and weight
   - Discrete data i.e., points scored in a game
2. **Categorical Variable:** variables that are naturally divided into *groups*
   - Binary
   - Nominal
   - Ordinal

Variables


all variables
numerical
categorical
discrete
continuous
ordinal
nominal

Gray areas

The type of variable dictates how we analyze it:

- We often use the **mean** or **average** to analyze quantitative variables
- We often use **proportions** or **percentages** to analyze categorical variables

Sometimes there are situations in which a variable is technically one type, but it may be more useful to analyze it as another:

Gray areas

Take a few minutes to discuss these questions with your group whether these might be used as quantitative or categorical variables:

1. Grades for a statistics class
2. A Likert Scale with five levels, measuring pain from "None at all" to "Extreme"
3. The year of birth for people enrolled in STA-209

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

Summarizing Data

Data collection has made remarkable progress in the last decades, giving us a greater quantity of data than most could ever dream of

In it's raw form, it becomes nearly incomprehensible to comprehend

Typically, we present either *numerical or visual summaries* of our data to facilitate interpretation

Next lecture we will consider both **univariate** visual summaries of a single variable, as well as **bivariate** visual summaries demonstrating the relationship between two variables.

Knowledge Check

Why do we need statistics?

Could you describe the statistical framework to your mother?

What is an observation and how do we describe its characteristics?

What types of variables are there, and when is each appropriate?

Summary

Statistics is a domain agnostic tool that allows us to make quantitative statements about a population

Most data that we encounter will be categorical or quantitative in nature

Next Time:

- Introduction to R
- Read 1.2 and 1.3 from IMS

Sources

IMS textbook
Professor Miller's course notes