Introduction

Ryan Miller

Two questions

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

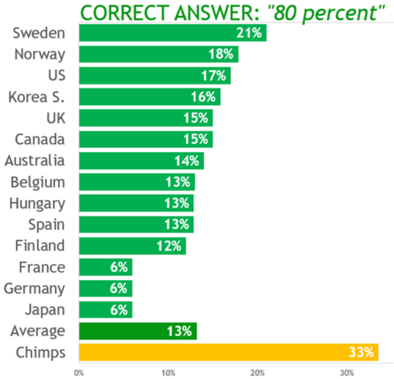
Another question

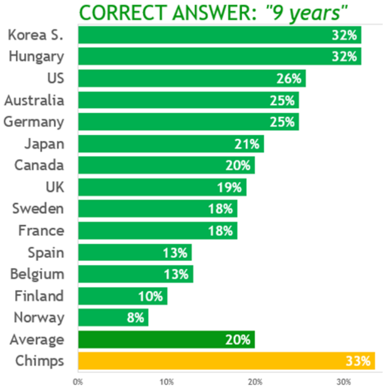
Why do we need data?

1. Humans are great at coming up with non-existent patterns
2. Humans are bad at estimating things
3. Humans are subject to all sorts of biases

Two questions

Here's what the data says about the two questions you answered earlier:





Source: Allan Rossman's JSM talk

Why do we need data?

- Over the last 20 years, the world has made remarkable progress in a great number of areas
  - Due to biases and a lack of exposure to quality data, most people's perceptions don't match reality
- Collecting, analyzing, and interpretting data provides us the power to *objectively understand reality*
  - In other words, data provides a way of overcoming biases and more accurately understanding the world

Why do we need statistics?

- The original Starburst candy comes in 4 flavors, strawberry (pink), orange (orange), lemon (yellow), and cherry (red)
  - I like red/pink and dislike yellow
  - 4 of 12 pieces (33%) in the last package I bought were yellow!
  - Should I believe that 1/3 of *all* starbursts are lemon? Is the company ripping me off?
- A polling company surveys 1,200 registered voters and finds that 52% support candidate "X"
  - Do the majority of voters support candidate "X"?

Why do we need statistics?

Statistics (as a discipline) is all about *understanding uncertainty*

- What can we learn from 1 package of Starbursts? What can we learn from a sample of 1,200 registered voters?

Some relevent quotes:

> *"Statistical thinking will one day be as necessary a qualification for efficient citizenship as the ability to read and write." H.G. Wells (Sci-Fi writer, 1866-1946)*

> *"Many use statistics in the same way that a drunk uses lamp-posts, for support rather than illumination." Andrew Lang (Scottish scholar, 1844-1912)*

In this class . . .

A brief outline of this class:

1. Describing data and variable relationships
   - univariate and bivariate summaries (numeric and graphical)
   - multivariate relationships (confounding variables)
2. Estimation
   - populations and samples
   - confidence intervals
3. Hypothesis Testing
   - one-sample and two-sample tests
   - Chi-squared tests
4. Statistical models
   - regression

The first steps

To work within any field, you must know its 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)

Software

Modern statistics, like many fields, has become increasingly reliant on computing. This class will involve a mixture of "pencil and paper" work, and analyses done using Minitab. Today we will:

1. Practice loading data into Minitab
2. Identify cases and types of variable
3. Practice some basic data manipulations using Minitab

Happy Planet Data Dictionary

- **Country**: Name of the country
- **Region**: Code for the region, 1 = Latin America, 2 = Western Nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = Former Communist Countries
- **Happiness**: 0 to 10 score from Gallop World Poll data
- **LifeExpectancy**: Average life expectancy (years) from UN Department of Economic and Social Affairs
- **Footprint**: A measure of ecological footprint from *The Edition of the Global Footprint Networks National Footprint Accounts*, higher numbers indicate greater environmental impact
- **HLY**: Happy Life Years - a combined measure of life expectancy and well-being
- **HPI**: Happy Plant Index - a 0-100 score
- **HPIRank**: HPI rank of the country
- **GDPperCapita**: Gross Domestic Product per capita
- **HDI**: Human Development Index from the UN Human Development Report Office
- **Population**: Population (in millions)

Practice - Cases and variables

Click this link to download the “Happy Planet” data. After loading it into Minitab, discuss the following with your group:

1. What is a case for these data?
2. What type of variable is “Region”?
3. Could these data be re-organized so that each region is a case?
4. What type of variable is “HPIRank”?

Practice - Solutions

1. A country
2. A nominal categorical variable
3. Yes, but it would require summarization
4. An ordinal categorical variable (each increase in rank doesn't mean the same thing)

Grey areas

Suppose I collect data on this class and I record your expected graduation year (ie: 2020, 2021, . . .)

- Is this variable quantitative or categorical?
- *Should we analyze* this variable as quantitative or categorical?
  - We often use the **mean**, or **average**, to analyze a quantitative variable
  - We often use **proportions**, or **percentages**, to analyze a categorical variable

Take a minute or two to discuss these questions with your group

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:

- The variable "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 variable might 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)

Practice - Transforming variables

Often it is necessary to create or transform variables, in Minitab this is done by:

1. Right clicking on a new column
2. Selecting “Formula” -> “Assign Formula to Column”

For the Happy Planet Data:

1. Create a new variable “TotalGDP” that transforms “GDPperCapita” into the country’s total GDP
2. Create a binary variable “IsAsia”, which records “Asia” if the country is located in South Asia (Region 5) or East Asia (Region 6), and “NotAsia” if the country is located in any other region.

Conclusion

After today you should:

1. Be comfortable with the terminology related to cases and variables
2. Be comfortable with loading and manipulating data in Minitab

If you need more guidance:

- Read Ch 1.1 and 1.2 of the text