Univariate Summaries

Ryan Miller

Outline

1. Why study statistics?
    - Motivating examples and rationale
2. The structure of data
    - Definitions and examples
3. Categorical variables
    - Numerical summaries and graphs
4. Quantitative (numeric) variables
    - Numerical summaries and graphs

Why do we need data?

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

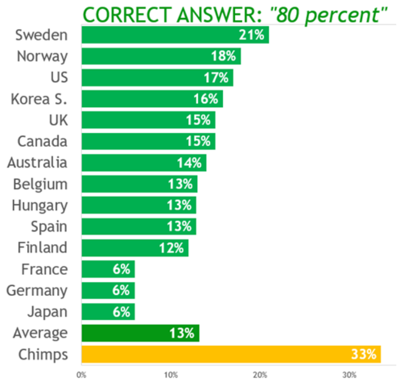
A) 20%
B) 50%
C) 80%

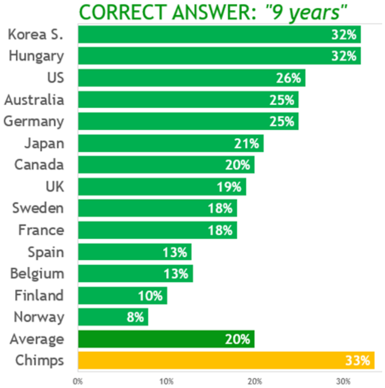
Question 2: Worldwide, 30-year-old men have an average of 10 years of schooling. What is the world average for women of the same age?

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

Why do we need data?

Here's what the data say about these two questions:





Source: Allan Rossman's JSM talk

Data and the field of statistics

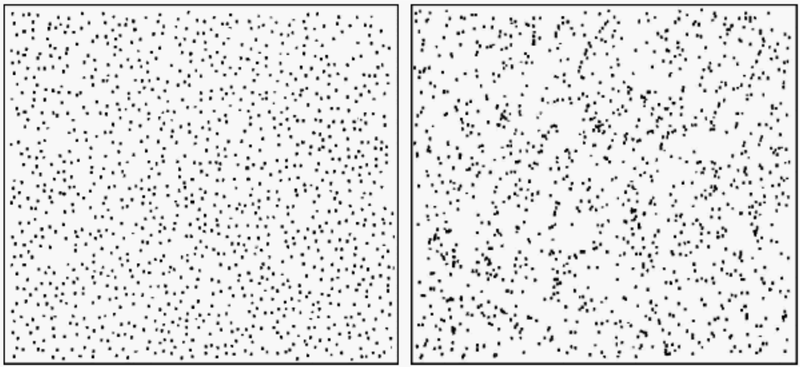
- Humans are bad at *objectively* assessing trends in the world around them
  - We often focus too much on rare/unusual events (ie: what's in the news)
  - We rely upon small samples (ie: personal experiences and anecdotes)
  - We tend to seek out evidence that confirms our prior beliefs

Data and the field of statistics

- Humans are bad at *objectively* assessing trends in the world around them
    - We often focus too much on rare/unusual events (ie: what's in the news)
    - We rely upon small samples (ie: personal experiences and anecdotes)
    - We tend to seek out evidence that confirms our prior beliefs
- The field of *Statistics* (of which *Biostatistics* is a sub-discipline) is the science of collecting, describing, and analyzing data
    - The tools of statistics enable us to make better, more accurate, conclusions in the face of uncertainty
    - Biostatistics focuses on the applications of statistics within the realm of biological and health sciences

Why do we need “statistics”?

One of these panels is *randomly generated*, the other contains an *underlying pattern*, which is which?



The structure of data

To work in any field, you must learn 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*)

The structure of data

To work in any field, you must learn 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)

Practice

Click here, or go to the “Data” section of our website, to download the “Happy Planet” dataset

1) What are the cases in this dataset?
2) What type of variable is “Region”?
3) What type of variable is “Population”?

Practice (solution)

1) The cases are countries.
2) “Region” is a nominal categorical variable (don’t be fooled by it being recorded using numeric values)
3) “Population” is a discrete quantitative variable (don’t be fooled by it being recording using decimal places)

Gray areas

Why classify variables?

- Helps us determine the proper methods to use in an analysis
- For example, we might report averages for quantitative variables, and proportions for categorical variables

Gray areas

Why classify variables?

- Helps us determine the proper methods to use in an analysis
- For example, we might report averages for quantitative variables, and proportions for categorical variables

In practice, how strict should we be?

- Rely on your best judgment rather than strict rules
- It might make sense to treat “graduation year” as categorical (if the number of unique numeric values is small)
- Similarly, a Likert scale (ie: Disagree, . . . , Neutral, . . . , Agree) might be better treated as quantitative

Summarizing data

Presenting raw data is rarely useful, humans aren't good at processing that type of information

- Shown here are SIDS (sudden infant death syndrome) cases in the Group Health Cooperative of Puget Sound (Seattle) health system between 1972 and 1983 following diphtheria-tetanus-pertussis (DTP) immunization
- Can you describe any noteworthy trends in these data?

	Sex	Days
1	F	60
2	M	78
3	M	80
4	M	77
5	F	87
6	M	115
7	M	175
8	F	56
9	F	60
10	M	114
11	M	81
12	M	58
13	M	103
14	M	134
15	M	46
16	F	53

Summarizing a single categorical variable

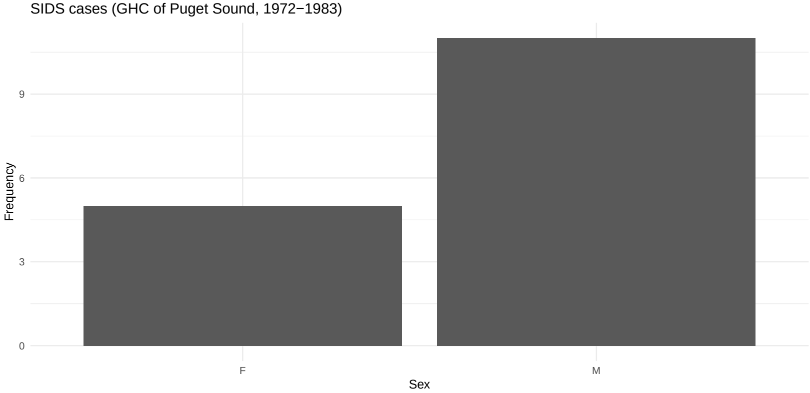
- A single categorical variable, such as “Sex” in the SIDS data, can be *summarized* using:
  - **Frequencies** - counts of how many cases belong to each category
  - **Proportions** - the fraction of the total number of cases that belong to each category (ie: Proportion $= \frac{\text{Cases in category}}{\text{Total cases}}$)

Table 1: This is a simple example of a ‘one-way frequency table’

Sex	Frequency
F	5
M	11

Graphs for a single categorical variable

A single categorical variable can be *graphed* using a **bar chart**:



Practice

StatKey is statistical software that we will use extensively in this course. It is web-based, completely free, and hosted online at: https://www.lock5stat.com/StatKey/

1) On the Statkey homepage, click on “One Categorical Variable” in the “Descriptive Statistics and Graphs” panel
2) Click on “Upload file” and load the Happy Planet data (it should be a .csv file in your downloads)
3) Choose the “Region” variable and observe the output StatKey provides
4) Interpret the proportion for the 4th region (ie: what does this summary say about these data?)

1	2	3	4	5	6	7
Latin America	Western Nations	Middle East	Africa	South Asia	East Asia	Former Soviet States

Practice (solution)

- #1-3 must be done on StatKey
- The proportion of countries in Region 4 is 0.231, this tells us that 23.1% of the countries in the Happy Planet data are located in Africa, the most prevalent region.

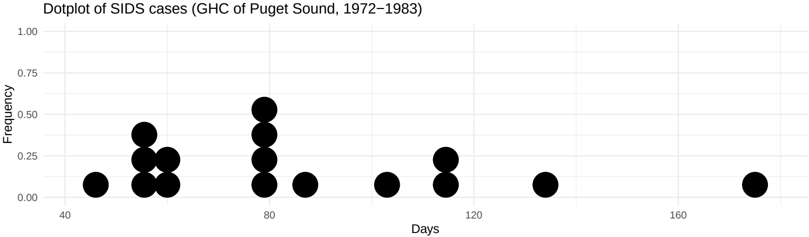
Summarizing a single quantitative variable

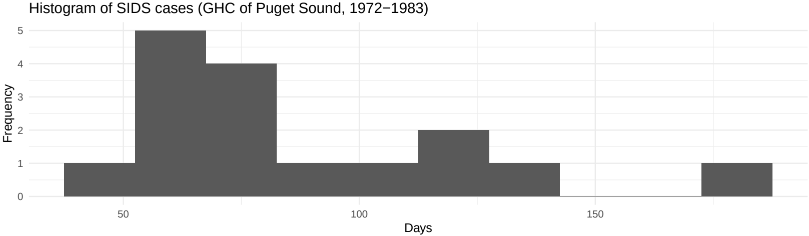
Summarizing a single quantitative is variable is more complex, as there are *three important aspects* of the variable's values we should consider:

1. **Shape** - what is the general shape of the *distribution* of the variable
2. **Center** - were do values of the variable appear to be centered around
3. **Spread** - to what extent do the values of the variable vary

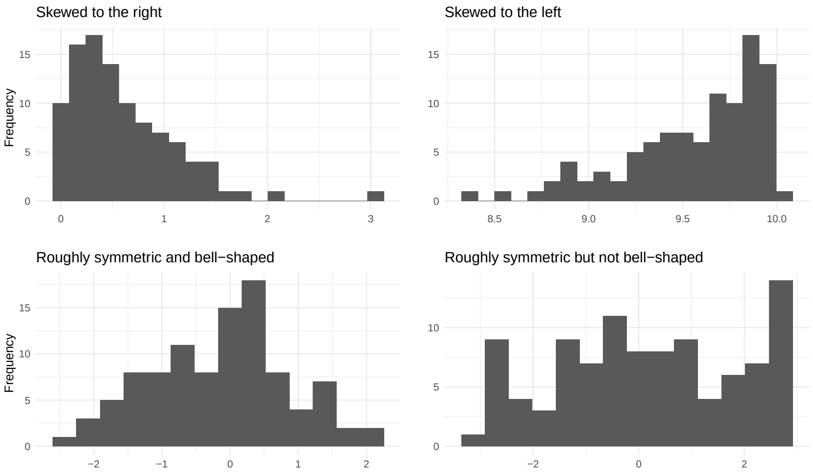
Shape (one quantitative variable)

The shape of a single quantitative variable should be assessed graphically using either a **dotplot** or a **histogram**





Important shapes (one quantitative variable)



Center (one quantitative variable)

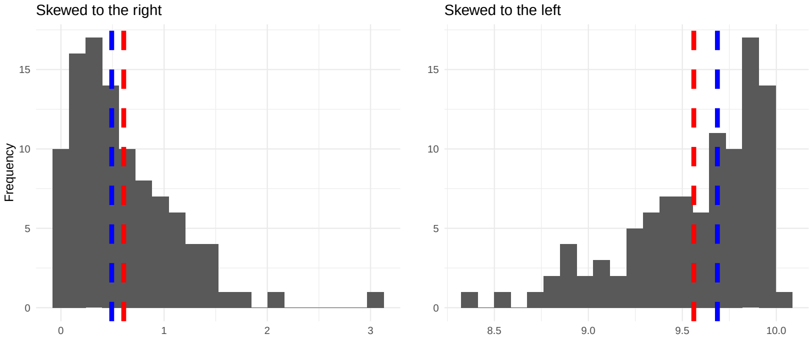
Important ways to summarize a variable's center:

- **Mean** - the arithmetic average of a variable (see Ch 2 of our textbook for a mathematical formula)
- **Median** - the middle value if the data were arranged in ascending order (see Ch 2 for a more detailed walk through)

The mean tends to be impacted by skew and outliers, while the median is considered to be *resistant*, or *robust*

Center vs. shape (one quantitative variable)

The mean (shown in red) is pulled in the direction of skew or outliers more so than the median (shown in blue)



Spread (one quantitative variable)

Important ways to summarize a variable's spread:

- **Standard deviation** - the average deviation (distance) of individual data-points from the center of the distribution (see our Ch 2.3 of our textbook for the mathematical formula)

Spread (one quantitative variable)

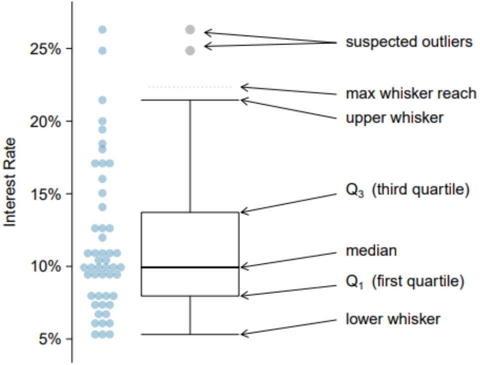
Important ways to summarize a variable's spread:

- **Standard deviation** - the average deviation (distance) of individual data-points from the center of the distribution (see our Ch 2.3 of our textbook for the mathematical formula)
- **Percentiles** - the P^{th} percentile is the value of a variable that is greater than P percent of the data
- **Range** - the difference in the data's maximum and minimum values
- **Interquartile Range (IQR)** - the difference in the 75^{th} and 25^{th} percentiles of the data (also called Q3 and Q1 respectively)

The standard deviation and range are *greatly* influenced by outliers, while the IQR is resistant/robust.

Boxplots and five number summaries

Boxplots are a type of graphical presentation that show robust measures of center and spread (shape cannot be definitively inferred)



The set of summary statistics contained in a boxplot (min, Q1, median, Q3, max) are called a **five number summary**

Practice

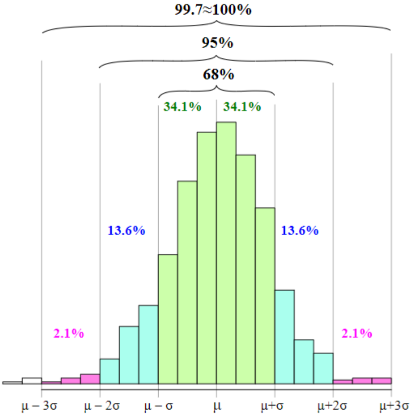
1) On the Statkey homepage, click on "One Quantitative Variable" button under "Descriptive Statistics and Graphs" panel
2) Upload the Happy Planet Dataset (if not already done) and select the variable "LifeExpectancy"
3) Describe the *shape*, *center*, and *spread* of the variable "LifeExpectancy"

Practice (solution)

- Shape - the distribution is left skewed (ie: there's a long tail of countries with low life expectancy)
- Center - the mean is 67.8 years, and the median is 71.5 years; these can be interpreted as the life expectancy of the typical/average country
- Spread - the standard deviation is 11.0 years, and the IQR s 14.15 years; both of these indicate a fairly large amount of variability across countries

The 68-95-99 Rule

The combination of *shape* and *spread* are particularly important due to the 68-95-99 rule (*only* for bell-shaped distributions)



This rule of thumb suggests that 95% of data-points are within 1 standard deviation of a variable's mean value!

Conclusion

1. Why study statistics?
   - To learn how to use data to make informed decisions in the face of uncertainty
2. The structure of data
   - Typically organized in spreadsheet form, where rows represent cases (units of observation) and columns represent variables (either categorical or quantitative)
3. Categorical variables
   - Summarized using frequencies and proportions, graphed using bar charts
4. Quantitative (numeric) variables
   - Graphed using dotplots, histograms, and boxplots
   - Described in terms of *shape* (ie: skew/symmetry), *center* (ie: mean/median), and *spread* (ie: standard deviation/IQR/range)