PS 211: Introduction to Experimental Design

Spring 2025 · Section C1

Lecture 3: Frequency distributions & visual displays of data

Updates and reminders

- You should now have R and R Studio installed on your computer.
- Please see me or Juneau *right away* if you have any issues.
- Homework 1 has been posted and is due (via Blackboard) on Monday, September 15 at 11:59 PM.
  - Homework 1 mostly covers material from Lectures 1 - 3, with a small amount of material from Lecture 4.
  - You should be able to do most of it after today's lecture.
- In today's lecture, we will also go over instructions for using R Markdown to complete Homework 1.
- My office hours tomorrow (Sept. 10) are cancelled, but if you need to meet, please Slack me and we can find an alternate time.

Review: Two branches of statistics

Descriptive statistics

- Descriptive stats allow us to summarize the characteristics or properties of a distribution of data.
  - They enable us to describe our *sample* or *population*.

- There are many correct ways to present the same data.
  - Our job is to choose the ways that are most useful.

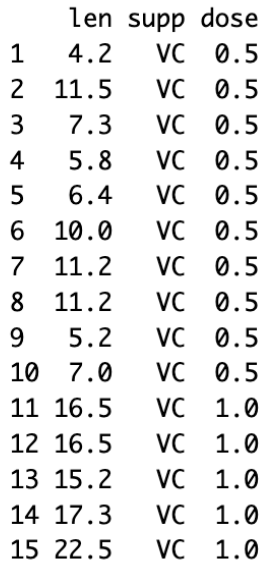
How should we describe data?

Raw data

- The original measurements or observations collected in a study.
- Have not been transformed, summarized, or analyzed.

Why is presenting raw data limiting?

Answer: It’s hard to see patterns or make comparisons when looking at a list of numbers. It’s generally helpful to organize or summarize them in some way.



```
    len supp dose
1   4.2   VC  0.5
2  11.5   VC  0.5
3   7.3   VC  0.5
4   5.8   VC  0.5
5   6.4   VC  0.5
6  10.0   VC  0.5
7  11.2   VC  0.5
8  11.2   VC  0.5
9   5.2   VC  0.5
10  7.0   VC  0.5
11 16.5   VC  1.0
12 16.5   VC  1.0
13 15.2   VC  1.0
14 17.3   VC  1.0
15 22.5   VC  1.0
```

How should we make sense of raw data?

One way: Frequency distributions

- Displays count or proportion of each value (or range of values) in a dataset.
- Options include:
  - Frequency tables
  - Grouped frequency tables
  - Histograms

Frequency tables

- A frequency table is a visual representation of a data set that shows how often (frequently) each value occurred.

- Values are listed in one column, and the count of individual scores with that value are listed in the second column.
  - All possible values are listed, even if the count is 0.

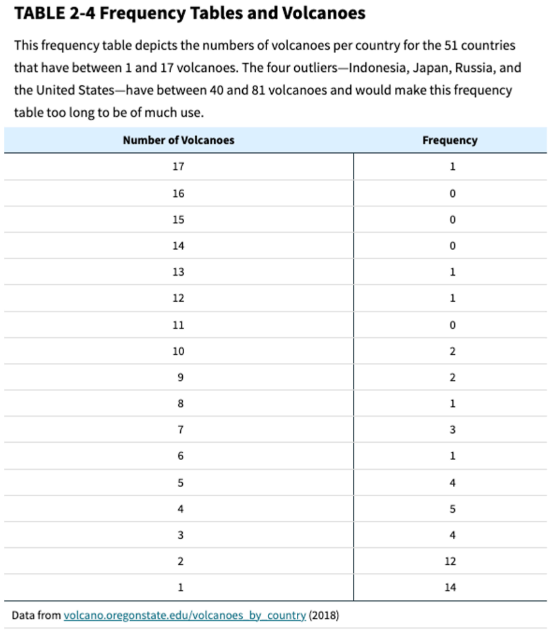


TABLE 2-4 Frequency Tables and Volcanoes

This frequency table depicts the numbers of volcanoes per country for the 51 countries that have between 1 and 17 volcanoes. The four outliers—Indonesia, Japan, Russia, and the United States—have between 40 and 81 volcanoes and would make this frequency table too long to be of much use.

Number of Volcanoes	Frequency
17	1
16	0
15	0
14	0
13	1
12	1
11	0
10	2
9	2
8	1
7	3
6	1
5	4
4	5
3	4
2	12
1	14

Data from volcano.oregonstate.edu/volcanoes_by_country (2018)

Creating a frequency table

General steps

- Create two columns
  - First column: list all possible values of the variable
  - Second column: tally the number of times each value occurs

In R:

- The 'dplyr' package can be used to create frequency tables easily.

```r
library(dplyr)

data %>%
  group_by(variable) %>%
  summarize(count = n())
```

For example:

```r
volcano_data %>%
  group_by(number_of_volcanoes) %>%
  summarize(count = n())
```

Including percentages

General steps

- Create two columns
  - First column: list all possible values of the variable
  - Second column: tally the number of times each value occurs
  - **Third column**: calculate the percentage of the total for each value
    - Percentage = (Count / Total number of observations) * 100

In R:

- The 'dplyr' package can be used to create frequency tables easily.

```r
library(dplyr)

data %>%
  group_by(variable) %>%
  summarize(count = n()) %>%
  mutate(percentage = (count / sum(count)) * 100)
```

For example:

```r
volcano_data %>%
  group_by(number_of_volcanoes) %>%
  summarize(count = n()) %>%
  mutate(percentage = (count / sum(count)) * 100)
```

Grouped frequency tables

- Sometimes frequency tables can be limited
  - What if data cover a wide range of values?
  - What if there are many unique values? (i.e., for continuous variables)
    - Listing all the possible values can basically be equivalent to displaying the raw data.

Grouped frequency tables:

- Show data spanning a specific interval, rather than individual values.
- Intervals (or "bins") are created to group values together.

Creating a grouped frequency table

General steps

- Determine the range of values in the dataset (max - min).
- Decide on the number of intervals (or "bins") to create.
- Determine the bottom of the lowest interval.
- Create intervals of equal width to cover the entire range of data.
- Tally the number of observations that fall within each interval.

Grouped frequency tables in R:

The 'cut' function in R can be used to create intervals.

```r
library(dplyr)

data %>%
  mutate(interval =
        cut(variable,
        breaks = seq(min, max, by = bin_width))) %>%
  group_by(interval) %>%
  summarize(count = n())
```

Converting a frequency table to a grouped frequency table

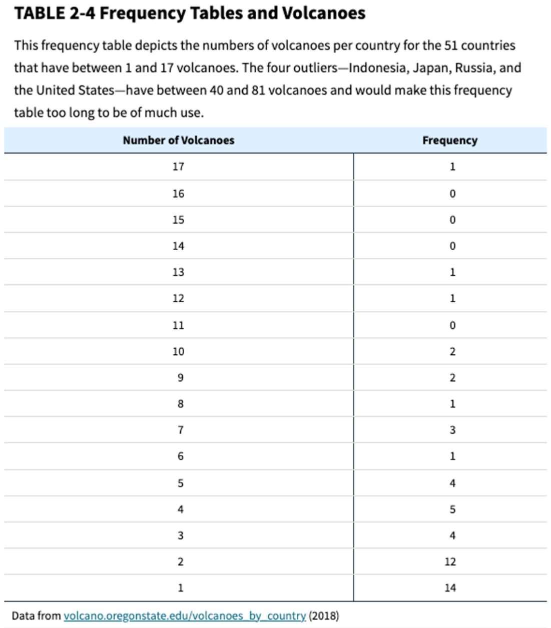


TABLE 2-4 Frequency Tables and Volcanoes

This frequency table depicts the numbers of volcanoes per country for the 51 countries that have between 1 and 17 volcanoes. The four outliers—Indonesia, Japan, Russia, and the United States—have between 40 and 81 volcanoes and would make this frequency table too long to be of much use.

Number of Volcanoes	Frequency
17	1
16	0
15	0
14	0
13	1
12	1
11	0
10	2
9	2
8	1
7	3
6	1
5	4
4	5
3	4
2	12
1	14

Data from volcano.oregonstate.edu/volcanoes_by_country (2018)

How would you convert this table to a grouped frequency table with intervals of width 5?

Answer: Create intervals like 0-4, 5-9, 10-14, etc., and tally the counts for each interval.

What is an advantage of using a grouped frequency table? What is a disadvantage?

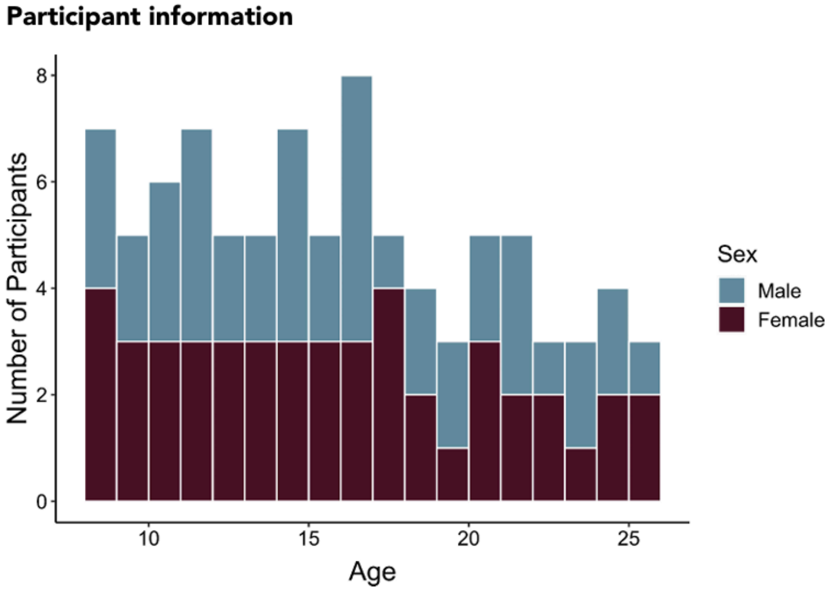
Answer: Advantage: Easier to see patterns in data with many unique values. Disadvantage: Loss of detail about individual values.

Histograms

- Graphs are often more effective than tables for seeing patterns in data.
- A **histogram** is a graphical representation of a grouped frequency table.
- The x-axis shows the intervals (or "bins") of values.
- The y-axis shows the frequency (or count) of observations in each interval.
- A histogram *looks* like a bar graph, but the y-axis always represents frequency or count, not a separate variable.

Histograms in the wild

I made this in R for an actual paper.



Creating a histogram in R

- ggplot2 is a popular R library for creating graphs, including histograms.
- Your assignments will use it throughout the semester.
- The 'syntax' (meaning the way you write the code) can be a bit tricky at first, but you’ll get the hang of it!
- Basic idea:
    - Tell R which dataset to use with `ggplot()`.
    - Use `aes()` to "map" the variable you want on the x-axis.
    - Use `geom_histogram()` to create the histogram.

```r
library(ggplot2)

ggplot(data, aes(x = variable)) +
  geom_histogram(binwidth = 5)
```

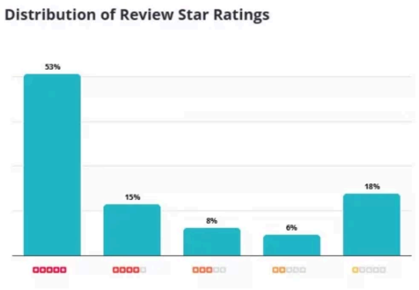
R histogram demo

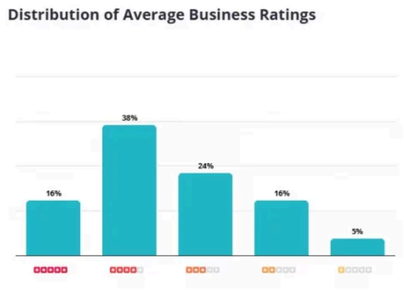
(R Demo)

- Open R Studio and create a new script.
- Load the ggplot2 library with `library(ggplot2)`.
- Use the `ggplot()` function to specify the dataset and mapping.
- Add `geom_histogram()` to create the histogram.
- Run the script to generate the histogram.

Distributions

- Histograms also let us see the **distribution** of a variable.
- A distribution describes how values of a variable are spread or clustered.
- The shape of a distribution provides key insights into the characteristics of the data.

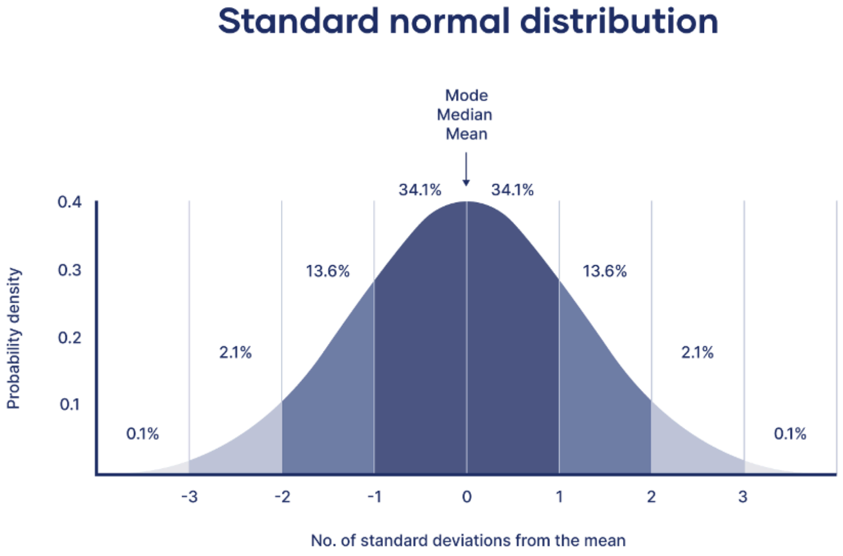

Distribution of Review Star Ratings


Distribution of Average Business Ratings

Note: Some people say histograms are for continuous variables only and that the bars should touch to indicate this. I don't think this is a hard-and-fast rule, but it's something to be aware of.

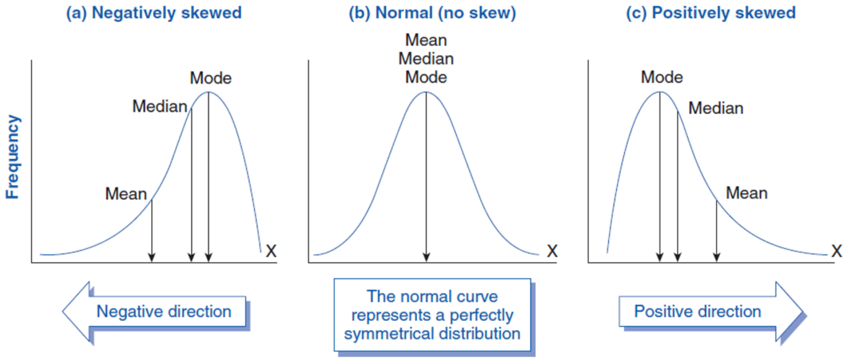
Normal distributions

- A **normal distribution** is a specific type of distribution that is symmetric and bell-shaped.
- We are going to discuss normal distributions much more extensively later in the course.
- For now, just know that many variables in nature and social science tend to follow a normal distribution.



Skew

- When data are not symmetrically distributed, we say they are **skewed**.
- The ends of the distribution are called the "tails."
- **Positively skewed** (or right-skewed) distributions have a long tail on the right side.
- **Negatively skewed** (or left-skewed) distributions have a long tail on the left side.

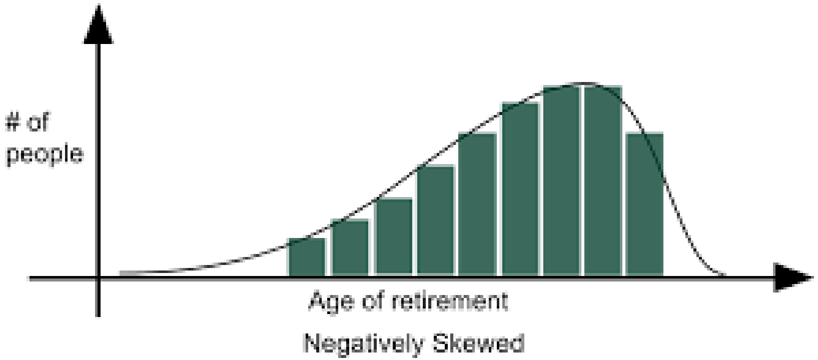


Positive skew and floor effects

- **Positive skew**: The tail of the distribution extends to the right.
- Sometimes a positive skew can indicate a **floor effect**.
- A **floor effect** occurs when a large number of observations cluster at the lower end of the scale, with few observations at the higher end.

Negative skew and ceiling effects

- **Negative skew**: The tail of the distribution extends to the left.
- Sometimes a negative skew can indicate a **ceiling effect**.
- A **ceiling effect** occurs when a large number of observations cluster at the higher end of the scale, with few observations at the lower end.



Practice: Skew

Consider these three variables: finishing times in a marathon of recreational runners, number of university dining hall meals eaten in a day, and scores on a scale of extroversion from a randomly sampled population.

Which variable do you think would be most likely to show a positive skew?

Answer: Finishing times in a marathon are likely to show a positive skew, as many runners may finish within a certain time range, but a few may take much longer.

Which variable do you think would be most likely to be normally distributed?

Answer: Scores on a scale of extroversion are often normally distributed in a randomly sampled population, as most people tend to fall in the middle range of extroversion, with fewer people being extremely introverted or extremely extroverted.

More practice

Can nominal variables have a skewed distribution? Why or why not?

Answer: No, nominal variables cannot have a skewed distribution because they represent categories without any inherent order or ranking. Skewness applies to the distribution of ordinal or continuous variables, where the data can be arranged along a scale.

You want to visualize the distribution of ages in a sample of adults. Would you use a frequency table, a grouped frequency table, or a histogram?

Answer: Either a grouped frequency table or a histogram would be appropriate for visualizing the distribution of ages in a sample of adults. A grouped frequency table would summarize the data into age intervals, while a histogram would provide a graphical representation of the same information.

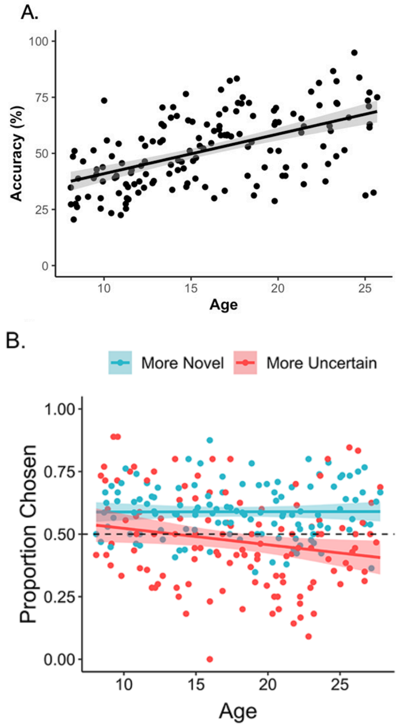
Moving beyond histograms: More on data visualization

- Histograms (and frequency tables) are just one of many ways to visualize data.
- There are **many** other types of graphs that can be useful for different purposes.
- We are going to go through some of them now.
- We will also discuss some general principles of effective data visualization.

Bar graphs

Scatter plots

- Scatter plots are used to visualize the relationship between two variables.
  - Usually continuous variables, but can also be used with ordinal variables.
- Each point on the graph represents an observation, with its position determined by the values of the two variables.
- Trendlines can be added to show the best-fitting line through the data points, indicating the overall direction of the relationship.
  - We will discuss how to calculate and interpret trendlines later in the course.



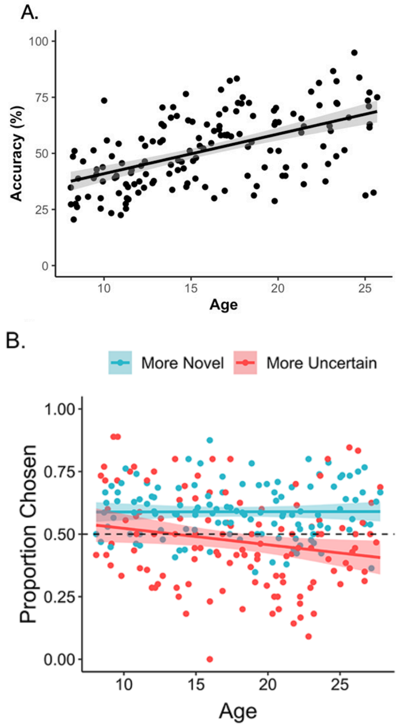
Scatter plots

What is one advantage of using a scatter plot?

Answer: One advantage of using a scatter plot is that it can show the relationship between two variables, making it easy to identify trends and correlations.

What is one disadvantage of using a scatter plot?

Answer: A disadvantage is that it can be difficult to interpret when there are many overlapping points, especially without a trendline.



Line graphs

Box plots

Other visualizations

- There are other ways to visualize data, including:
  - Violin plots
  - Pie charts
  - Heatmaps
  - And many more!

Principles of effective data visualization

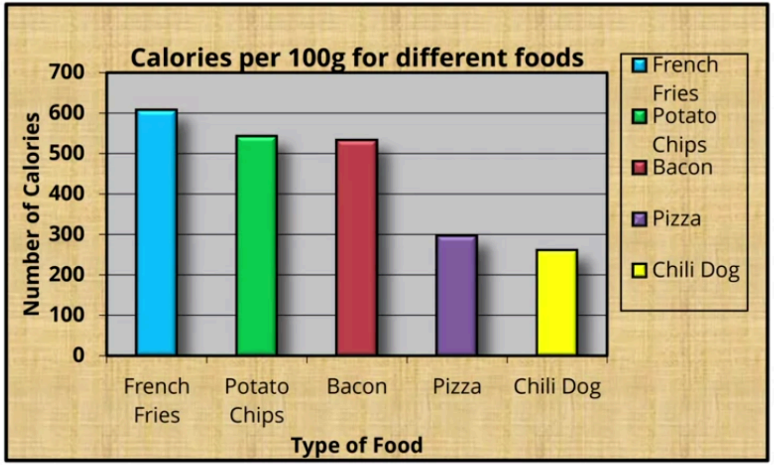
- Choose the right type of graph for your data and the message you want to convey.
- Keep it simple and avoid unnecessary elements that can distract from the main message.

K.I.S.S. - Keep It Simple Stupid
Great advice... Hurts my feelings every time

Principles of effective data visualization (continued)

- Use clear labels, titles, and legends to help the audience understand the graph.
- When possible, display the full distribution of data rather than just summary statistics.
    - This is increasingly become a requirement at many journals.

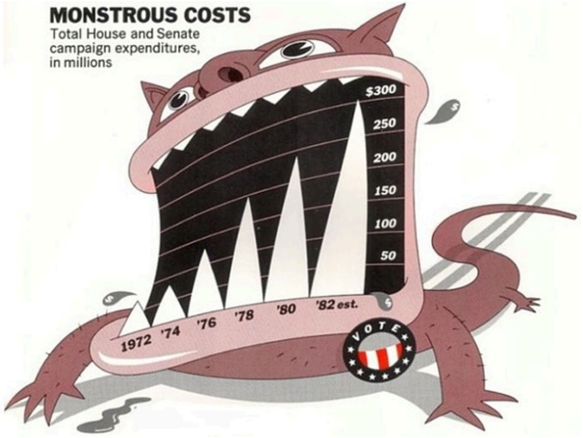
Keep it simple!



What is wrong with this graph?

Answer: The graph is overly complicated and includes unnecessary elements (chart junk) that distract from the main message. As one example: There is no reason for the bars to be different colors.

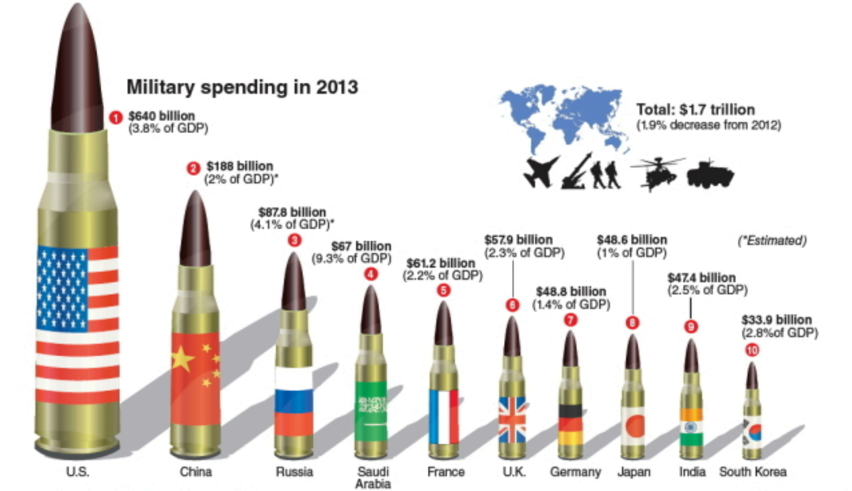
Keep it simple!



What is wrong with this graph?

Answer: Almost everything. The 3D effect distorts the data and the visuals are distracting.

Keep it simple!



Source: Stockholm International Peace Research Institute

Graphic by Han Chang-duk

What is wrong with this graph?

Answer: The visuals are distracting, the grid lines are unnecessary, there is no y-axis label, it is unclear how the image size relates to the data.

Practice: Choosing the right graph

For each of the following scenarios, decide which type of graph would be most appropriate (bar graph, scatter plot, or line graph) and explain your choice.

Visualizing the relation between hours studied and exam scores for a group of students.

Answer: Scatter plot - because both variables (hours studied and exam scores) are continuous, and we want to see the relationship between them.

Visualizing the average monthly temperatures over a year in a specific city.

Answer: Line graph - because we are looking at changes over time (monthly temperatures).

Visualizing the number of students in different majors at a university.

Answer: Bar graph - because we are comparing counts across different categories (majors).

Axes matter!

Axes can be extremely misleading.

- The choice of scale and range on the axes can dramatically affect how data are perceived.
- Always check the axes to ensure they accurately represent the data.

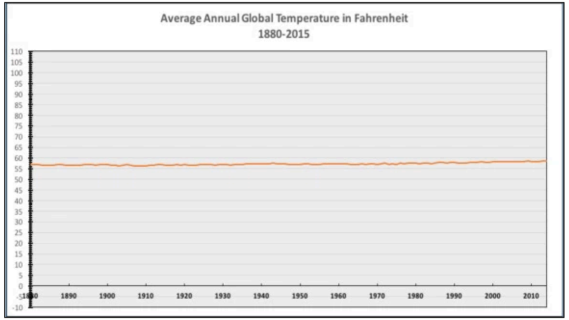
Good website for examples:

https://callingbullshit.org/tools/tools_misleading_axes.html

Axes matter: Examples

- Here are some examples of how axes can be manipulated to mislead viewers:

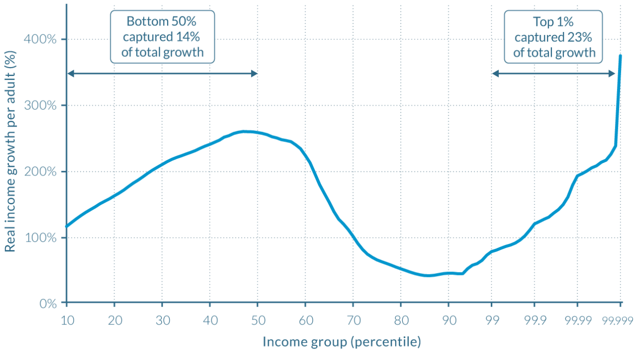
What is misleading about each of these graphs?



Answer: The y-axis covers a very large range, making differences appear non-existent.

Note: This is part of why we need statistical tests to determine if differences are "real" or not.

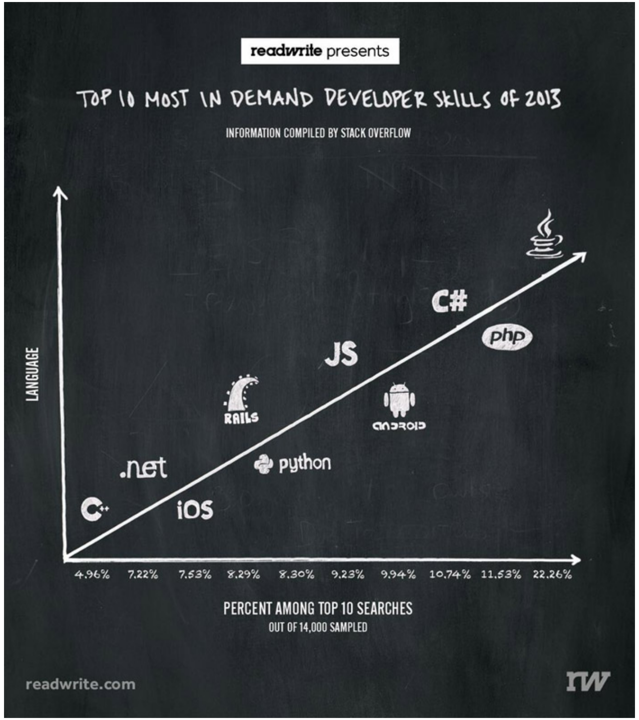
Axes matter: Examples (continued)



Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

Answer: The x-axis changes the scale, making the top 1% look more like the top 20%.



Answer: The y axis is completely meaningless.

That's all for data visualization!

We will continually revisit data visualization throughout the course.

Getting set up with R Markdown for Homework 1

(Demo)