COLLEGE OF CHARLESTON

Week 1

Math 104: Elementary Statistics

Chapter 1
Introduction to Statistics

“Half of marriages end in divorce”?

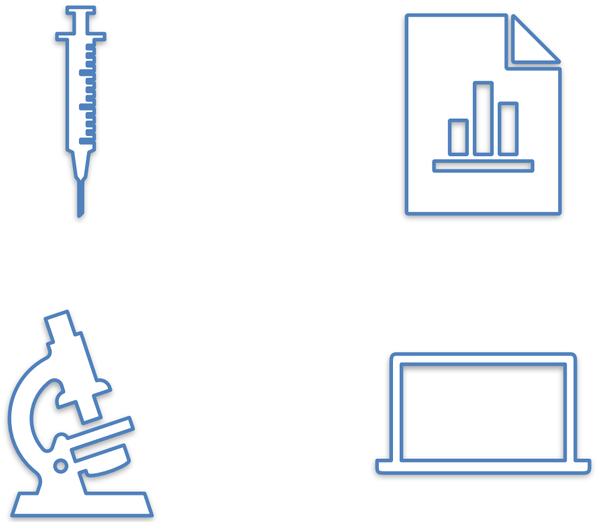
“The figure is based on a simple — and flawed — calculation: the annual marriage rate per 1,000 people compared with the annual divorce rate. In 2003, for example, the most recent year for which data is available, there were 7.5 marriages per 1,000 people and 3.8 divorces, according to the National Center for Health Statistics.

“But researchers say that this is misleading because the people who are divorcing in any given year are not the same as those who are marrying, and that the statistic is virtually useless in understanding divorce rates. In fact, they say, studies find that the divorce rate in the United States has never reached one in every two marriages, and new research suggests that, with rates now declining, it probably never will.”

Dan Hurley, “Divorce Rate: It's Not as High as You Think”, *New York Times*

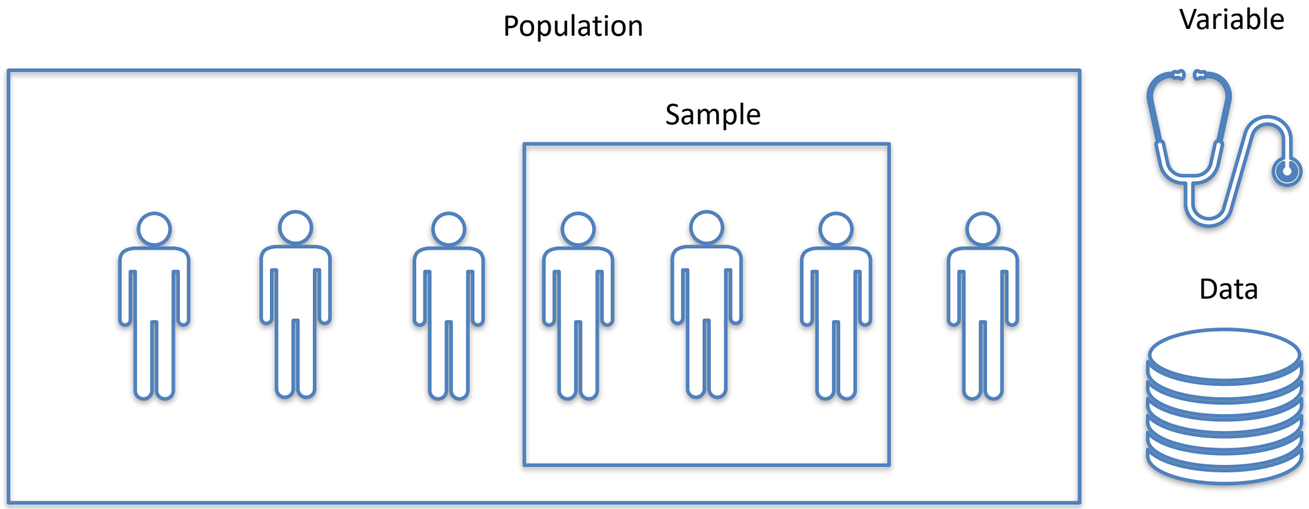
Why YOU should care about statistics

- Statistics is about data
- Misinformation abundant
  - fake news
  - deep fakes
- Applications of statistics
  - Healthcare
  - Business
  - Scientific research
  - AI



What is statistics?

- My definition: *the science of statistics*
- A **statistic** is a numerical description of **data samples** of a **population**.
  - **population**, a group of interest
  - **sample**, subset of population
  - **variable**, value or characteristic
  - **data**, measurements or observations about one or more variables
  - **parameter**, numerical description of a population
  - **sample**, subset of a population

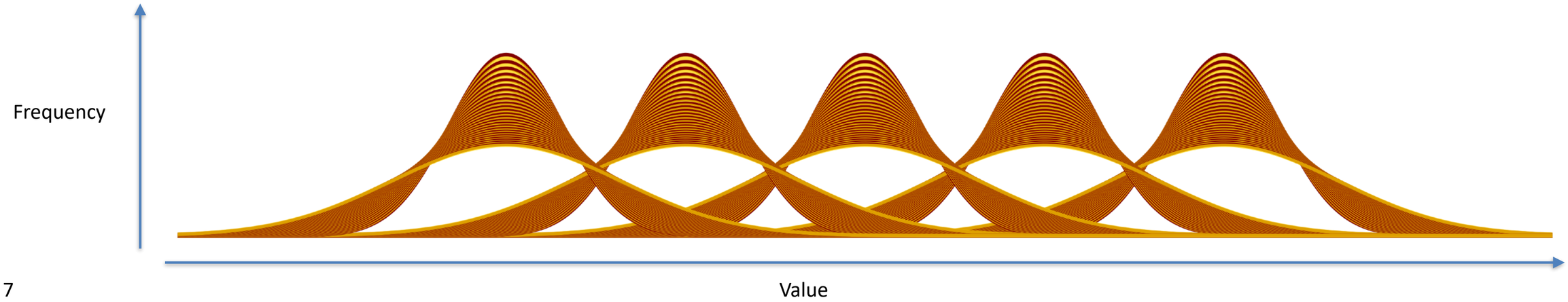


Branches of statistics

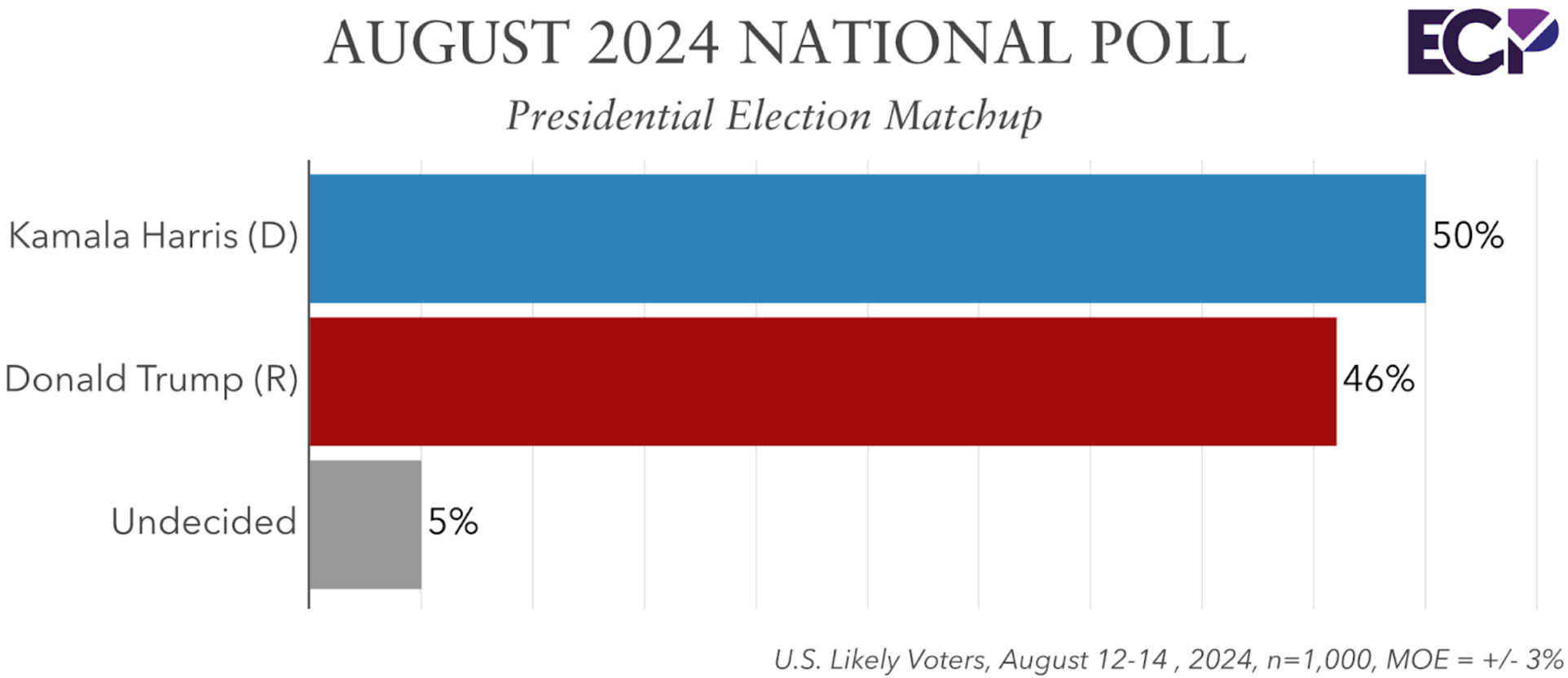
- **Descriptive Statistics**
  - Summarize data
  - Methods
    - Numerical descriptions of data (e.g. mean, standard deviation)
    - Graphical descriptions of data (e.g. histograms, box plots)
- **Inferential Statistics**
  - Learn population parameters (e.g. mean of a population)
  - Methods
    - Hypothesis testing
    - Confidence intervals
    - Regression analysis

Mathematics of statistics

- **probability**, the mathematics of chance
- **distribution**,
  - the frequency the various value of a variable occurs in a population
  - the probability of various values occurring
- **normal distribution**: the “bell curve”
- Why “normal”?
  - found in nature
  - sums of samples with arbitrary distributions often exhibit normal distribution

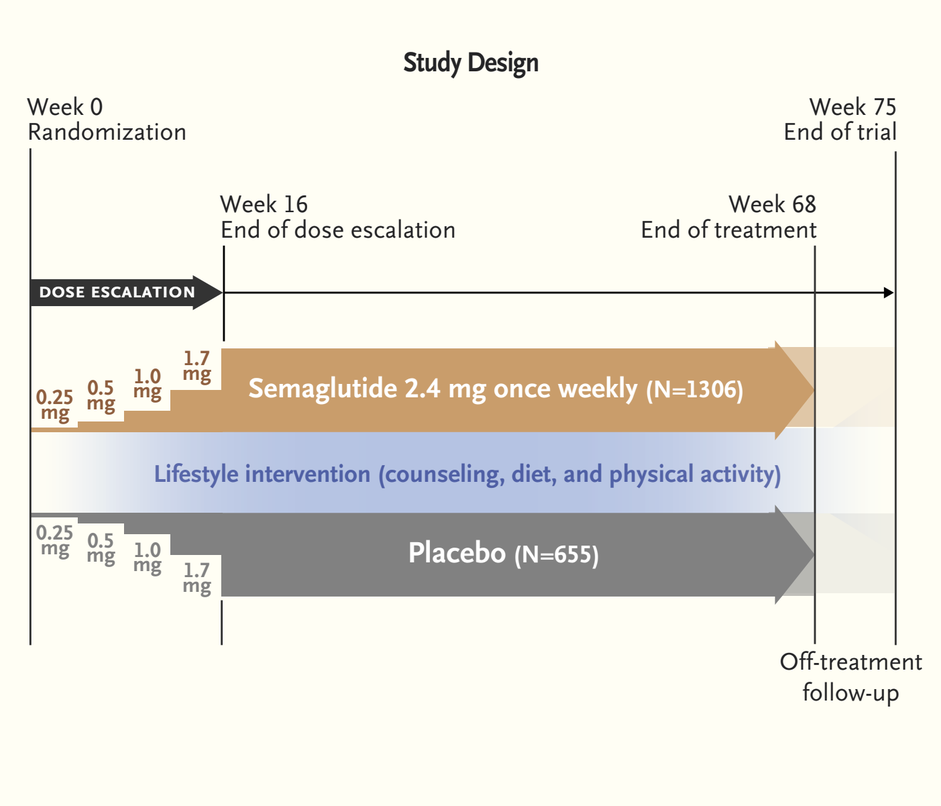


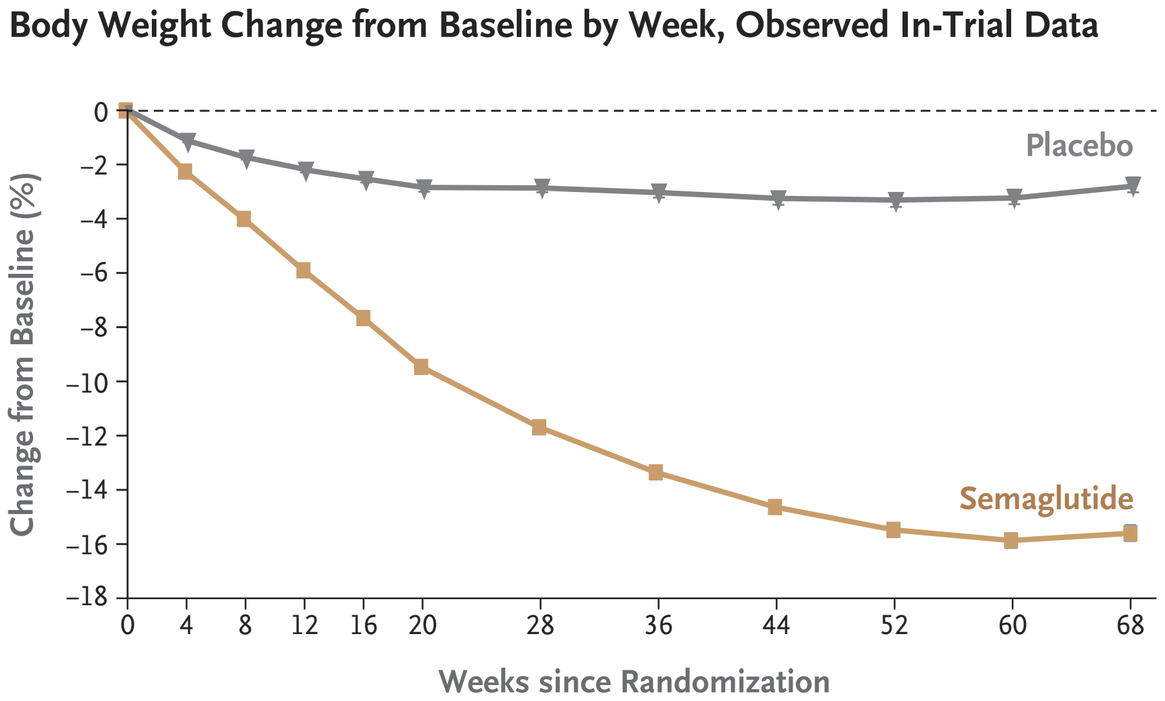
Case Study: Election Polls



Source: https://emersoncollegepolling.com/august-2024-national-poll-harris-50-trump-46/

Case Study: Clinical Trial





Source: https://www.nejm.org/doi/full/10.1056/NEJMoa2032183

1.2 Data Classification

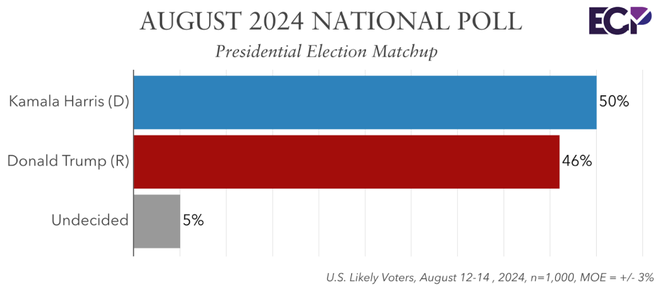
Types of data

- **Qualitative** data
  - labels of samples
  - often count the number of samples with the same label, or **class**
- **Quantitative** data
  - Numerical measurements of samples
  - **discrete**, only takes on particular values
  - **continuous**, take on any value (at least in an interval)
- Levels of measurement
  - **nominal**, labels or names for each sample
  - **ordinal**, relative ordering of each sample
  - **interval**, numbers for each sample and zero is a placeholder
  - **ratio**, numbers and zero indicates absence of something

Chapter 2
Graphical Descriptions of Data

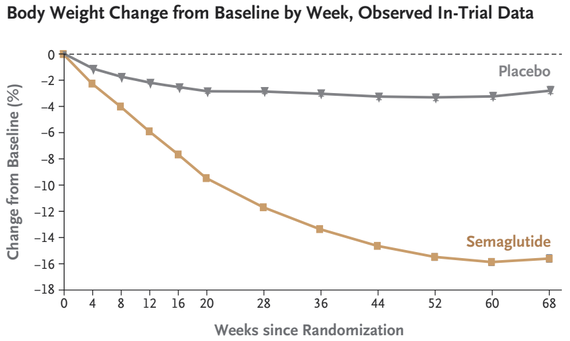
Why visualize data?

- Simplify complex data
- Highlight key insights
- Enhance communication
- Facilitate decision making
- Reveal hidden patterns



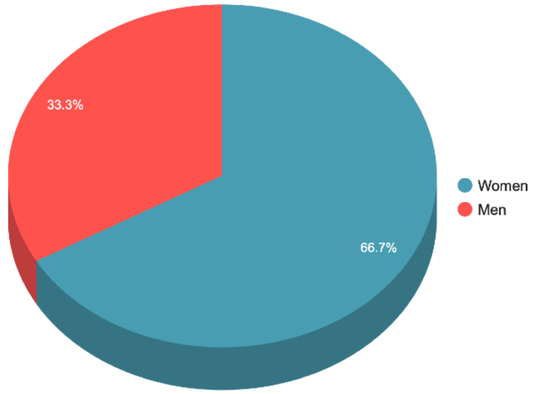
Bar plot

Source: https://emersoncollegepolling.com/august-2024-national-poll-harris-50-trump-46/



Line graph

Source: https://www.nejm.org/doi/full/10.1056/NEJMoa2032183



Pie chart

Source: https://www.collegefactual.com/colleges/college-of-charleston/student-life/diversity/chart-undergraduate-gender-diversity.html

2.1 Frequency Distributions

Why frequency distributions?

- Describe the number of occurrences of ranges of values in a sample
- Useful in drawing graphs

How to compute frequency distributions

- Step 1: decide how many classes should be in the distribution
- Step 2: choose the appropriate class width
- Step 3: find the class limits
- Step 4: determine the frequency of each class

Example: Salaries

- Step 1: decide classes
  - Let's choose 7 classes
- Step 2: calculate class width
  - (82,700-50,700)/7 = 32,000/7 ≈ 5,000
- Step 3: calculate class limits
  - Lower limit = 85,000 (maximum to the nearest 5,000)
  - Upper limit = 50,000 (minimum to the nearest 5,000)

Highest Early-Career Salaries with a Bachelor's Degree (in an Ordered Array)				
50,700	53,400	54,700	60,700	60,800
62,700	63,200	63,900	64,000	65,400
67,900	68,900	69,900	70,700	70,800
71,100	71,300	71,400	71,800	71,900
72,600	72,600	74,000	79,600	82,700

Example: Salaries

- Step 4: Calculate frequencies

Highest Early-Career Salaries with a Bachelor's Degree	
Class	Frequency
50,000 - 54,999	
55,000 - 59,999	
60,000 - 64,999	
65,000 - 69,999	
70,000 - 74,999	
75,000 - 79,999	
80,000 - 84,999	

Highest Early-Career Salaries with a Bachelor's Degree				
50,700	53,400	54,700	60,700	60,800
62,700	63,200	63,900	64,000	65,400
67,900	68,900	69,900	70,700	70,800
71,100	71,300	71,400	71,800	71,900
72,600	72,600	74,000	79,600	82,700

Example: Salaries

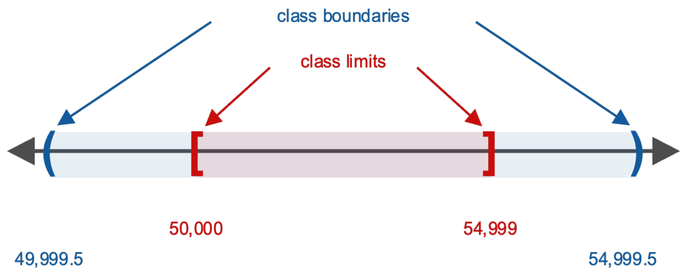
- **Frequency table**

Highest Early-Career Salaries with a Bachelor's Degree	
Class	**Frequency**
50,000 - 54,999	3
55,000 - 59,999	0
60,000 - 64,999	6
65,000 - 69,999	4
70,000 - 74,999	10
75,000 - 79,999	1
80,000 - 84,999	1

Class boundaries

Look at the first and second classes. The upper limit of class one is 54,999. The lower limit of class two is 55,000. Thus, the class boundary between the first two classes is calculated as follows.

$$\frac{(54{,}999 + 55{,}000)}{2} = 54{,}999.5$$

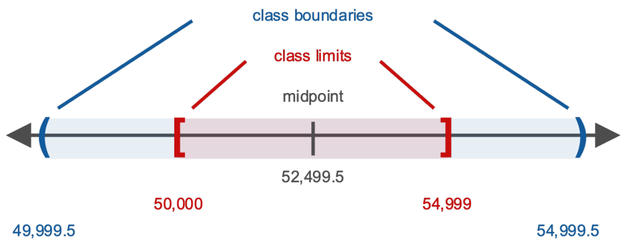


Highest Early-Career Salaries with a Bachelor's Degree		
Class	**Frequency**	**Class Boundaries**
50,000 – 54,999	3	49,999.5 – 54,999.5
55,000 – 59,999	0	54,999.5 – 59,999.5
60,000 – 64,999	6	59,999.5 – 64,999.5
65,000 – 69,999	4	64,999.5 – 69,999.5
70,000 – 74,999	10	69,999.5 – 74,999.5
75,000 – 79,999	1	74,999.5 – 79,999.5
80,000 – 84,999	1	79,999.5 – 84,999.5

Class midpoints

The midpoint is the sum of the class limits divided by two. For the first class, the midpoint is calculated as follows.

$$\frac{(50{,}000 + 54{,}999)}{2} = 52{,}499.5$$



Highest Early-Career Salaries with a Bachelor's Degree (with Class Midpoints)		
Class	Frequency	Class Midpoints
50,000 – 54,999	3	52,499.5
55,000 – 59,999	0	57,499.5
60,000 – 64,999	6	62,499.5
65,000 – 69,999	4	67,499.5
70,000 – 74,999	10	72,499.5
75,000 – 79,999	1	77,499.5
80,000 – 84,999	1	82,499.5

Other notions of class frequency

- The **relative frequency** is the fraction or percentage of the data set that falls into a particular class, given by the formula: Relative Frequency $= \frac{f}{n}$ where f is the class frequency and n is the sample size given by $n = \sum f_i$ and f_i is the frequency of Class i
  - What is the sum $\sum \frac{f_i}{n}$?
- The **cumulative frequency** is the sum of the frequencies of a given class and all previous classes.
  - What is the cumulative frequency of the last class?

Relative frequency

Highest Early-Career Salaries with a Bachelor's Degree (with Relative Frequencies)		
Class	**Frequency**	**Relative Frequency**
50,000 – 54,999	3	3/25=012=12%
55,000 – 59,999	0	0/25=0.0=0%
60,000 – 64,999	6	6/25=0.24=24%
65,000 – 69,999	4	4/25=0.16=16%
70,000 – 74,999	10	10/25=0.4=40%
75,000 – 79,999	1	1/25=0.04=4%
80,000 – 84,999	1	1/25=0.04=4%

Cumulative frequency

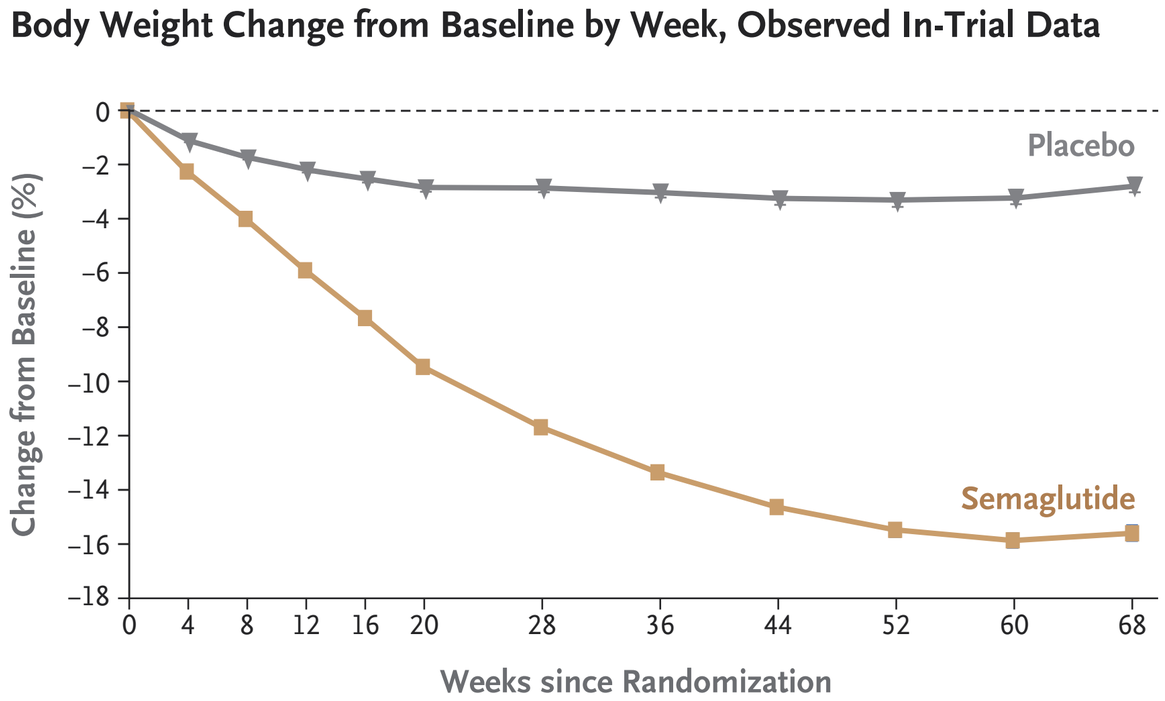
Highest Early-Career Salaries with a Bachelor's Degree (with Cumulative Frequencies)		
Class	**Frequency**	**Cumulative Frequency**
50,000-54,999	3	3
55,000-59,999	0	3 (3+0)
60,000-64,999	6	9 (3+0+6)
65,000-69,999	4	13 (3+0+6+4)
70,000-74,999	10	23 (3+0+6+4+10)
75,000-79,999	1	24 (3+0+6+4+10+1)
80,000-84,999	1	25 (3+0+6+4+10+1+1)

2.3 Graphical Displays of Data

Graphs

- A **graph** is a picture of the data that allows us to view patterns at a glance
- A **legend** is a description of how each data category is identified in the graph
- Graphs should have a
  - Title
  - Labels on the axis (x-axis and y-axis)
  - Source
  - Data
  - Legend (if applicable)

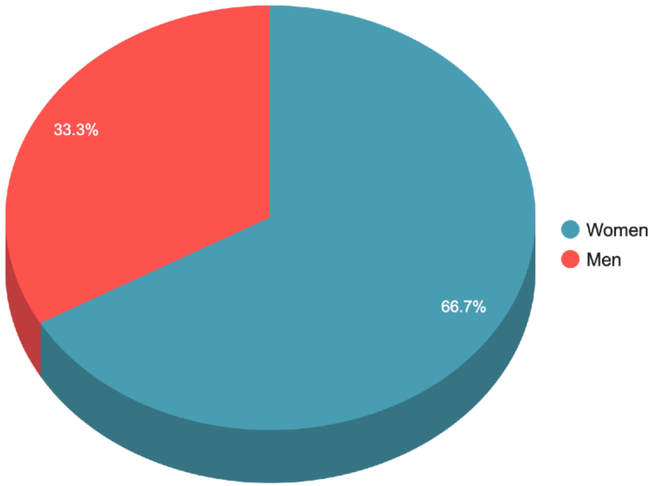
Can you spot the legend?



Source: https://www.nejm.org/doi/full/10.1056/NEJMoa2032183

Pie Charts

- A **pie chart** is a circular graph used for qualitative data that depicts parts of a whole and shows how large each category is in relation to the whole
- Pie charts display nominal data



Source: https://www.collegefactual.com/colleges/college-of-charleston/student-life/diversity/chart-undergraduate-gender-diversity.html

How to draw a pie chart

- Calculate the relative frequency for each category
- Divide the "pie" into "slices" whose area is proportional to the relative frequency
  - Multiple the relative frequency by 360 degrees

Housing Types for Students in a Statistics Class	
Type of Housing	**Number of Students**
Apartment	20
Dorm	10
House	5
Sorority/Fraternity House	5
Total	40

Relative Frequencies

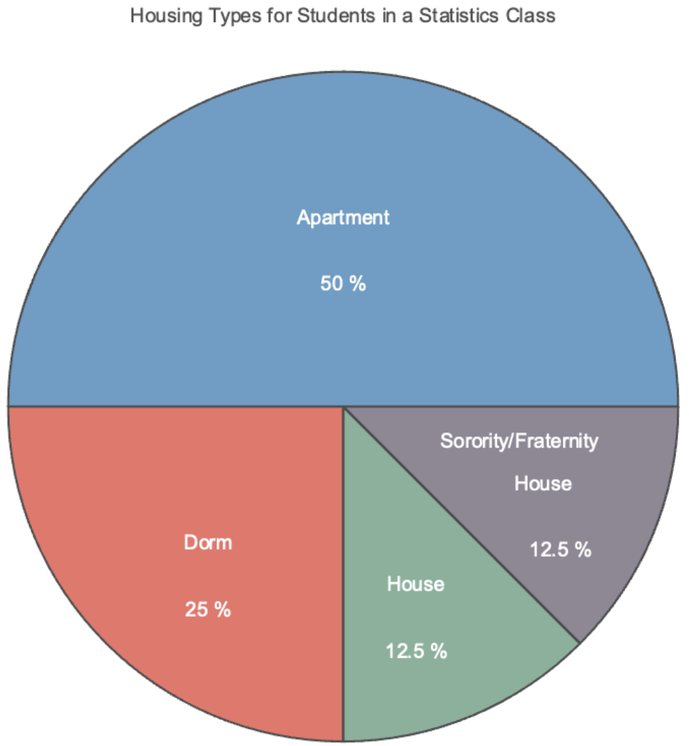
Apartment: $\frac{f}{n} = \frac{20}{40} = 0.5 = 50\%$

Dorm: $\frac{f}{n} = \frac{10}{40} = 0.25 = 25\%$

House: $\frac{f}{n} = \frac{5}{40} = 0.125 = 12.5\%$

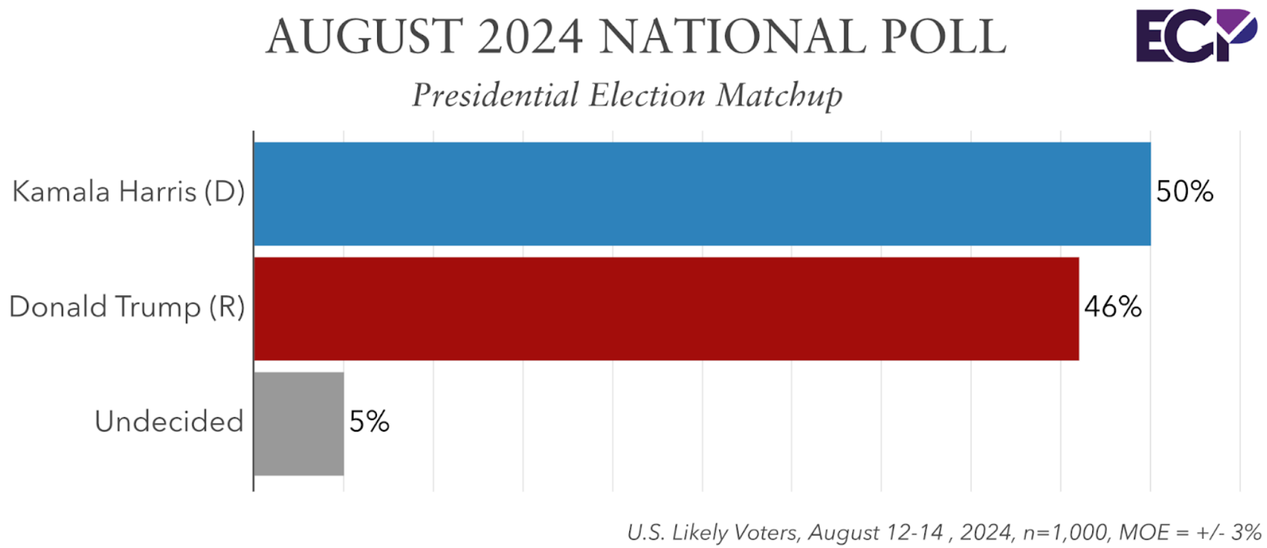
Sorority/Fraternity: $\frac{f}{n} = \frac{5}{40} = 0.125 = \underline{12.5\%}$

100%



Bar graphs

- A **bar graph** is a graph that uses bars to represent the amount of data in each category.
- Bar graphs also display nominal data

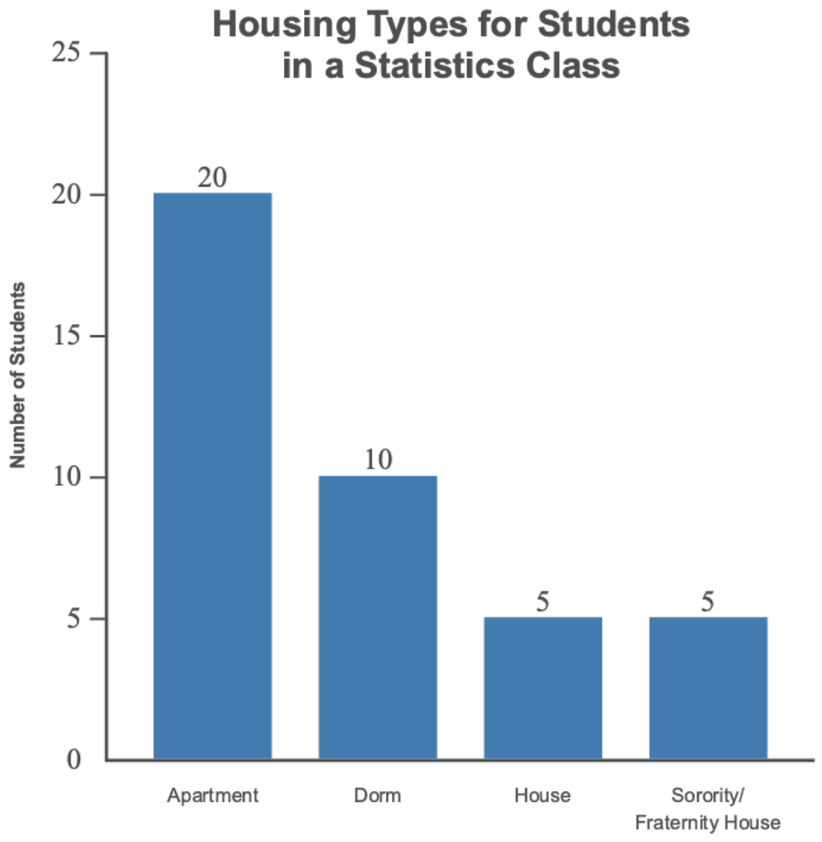


Source: https://emersoncollegepolling.com/august-2024-national-poll-harris-50-trump-46/

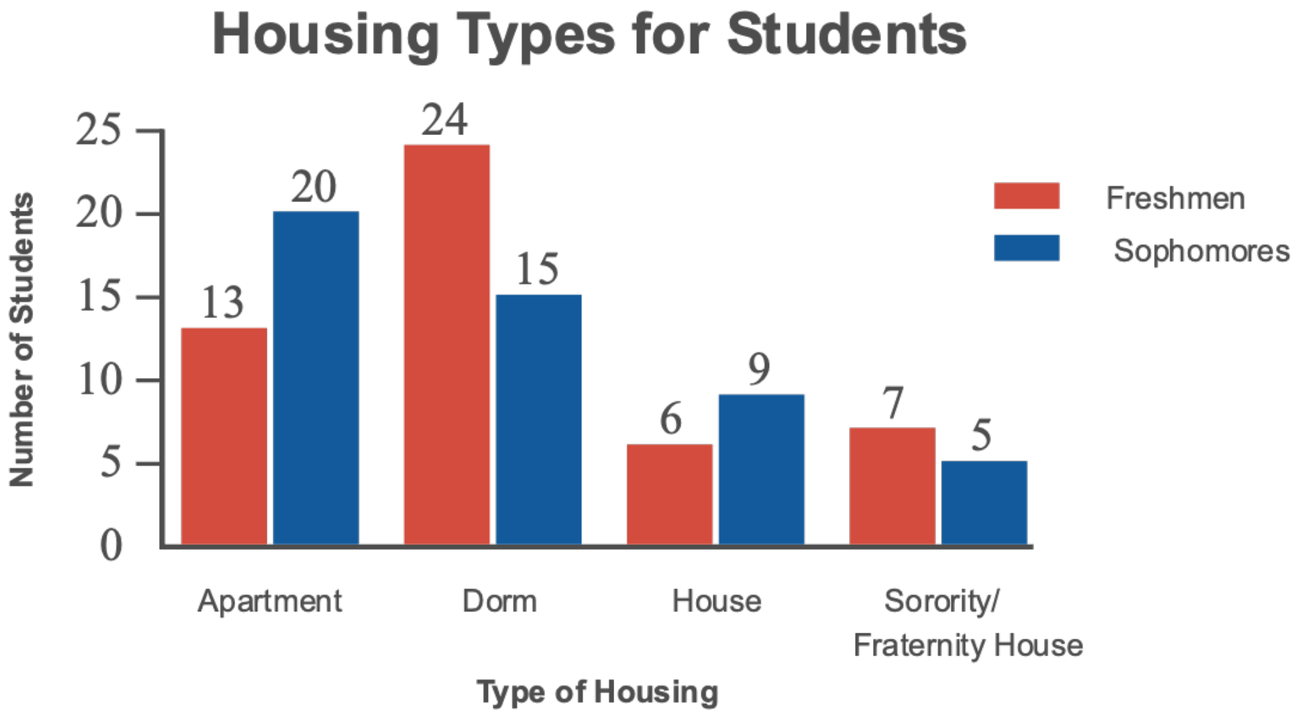
How to draw a bar graph

- Label the x-axis with the categories and label the y-axis with a numerical scale
- Over each category draw a bar of equal height to the number of occurrences of the category

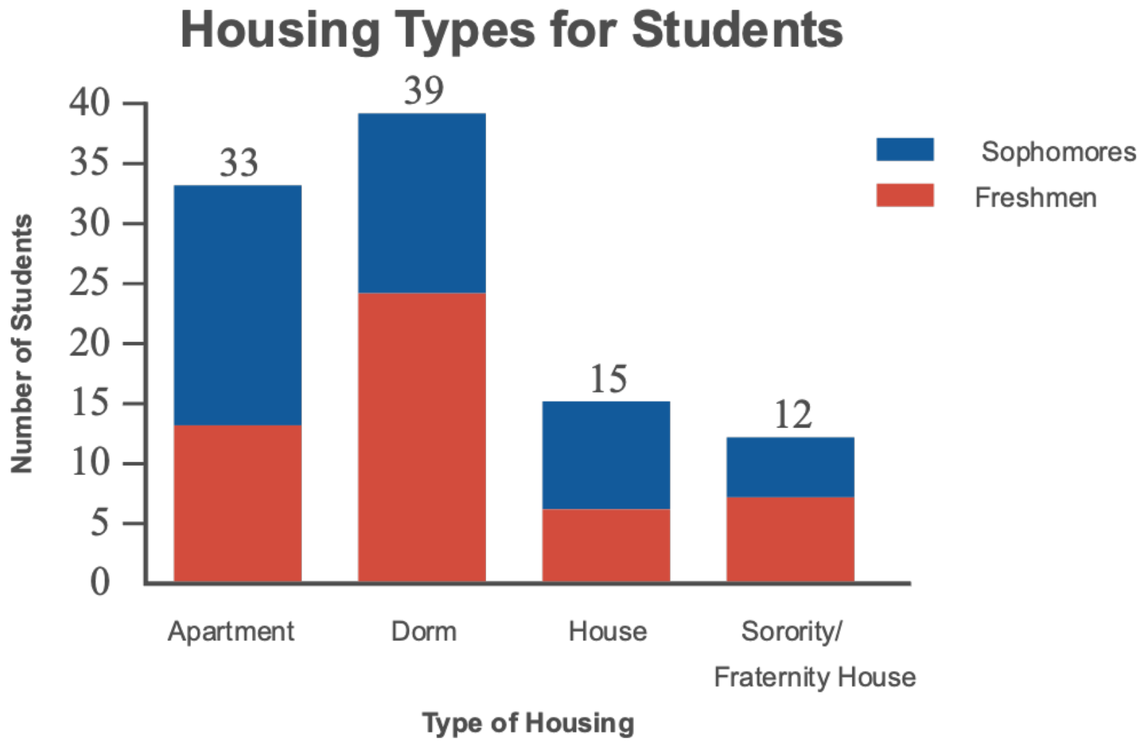
Housing Types for Students in a Statistics Class	
Type of Housing	**Number of Students**
Apartment	20
Dorm	10
House	5
Sorority/Fraternity House	5
Total	40



Variations of bar graphs



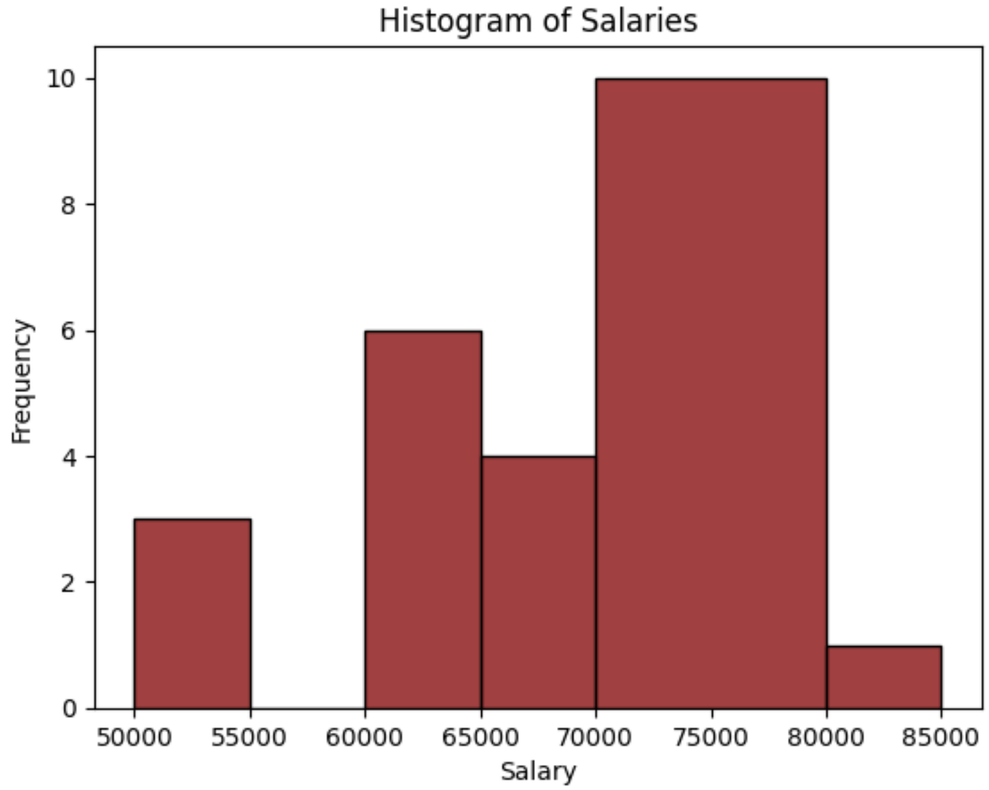
side-by-side bar graph



stacked bar graph

Histograms

- A histogram displays the frequency distribution of quantitative data
- **Discuss the benefits/drawbacks of left v.s. right with your neighbor**



versus

Early-Career Salaries				
50,700	53,400	54,700	60,700	60,800
62,700	63,200	63,900	64,000	65,400
67,900	68,900	69,900	70,700	70,800
71,100	71,300	71,400	71,800	71,900
72,600	72,600	74,000	79,600	82,700

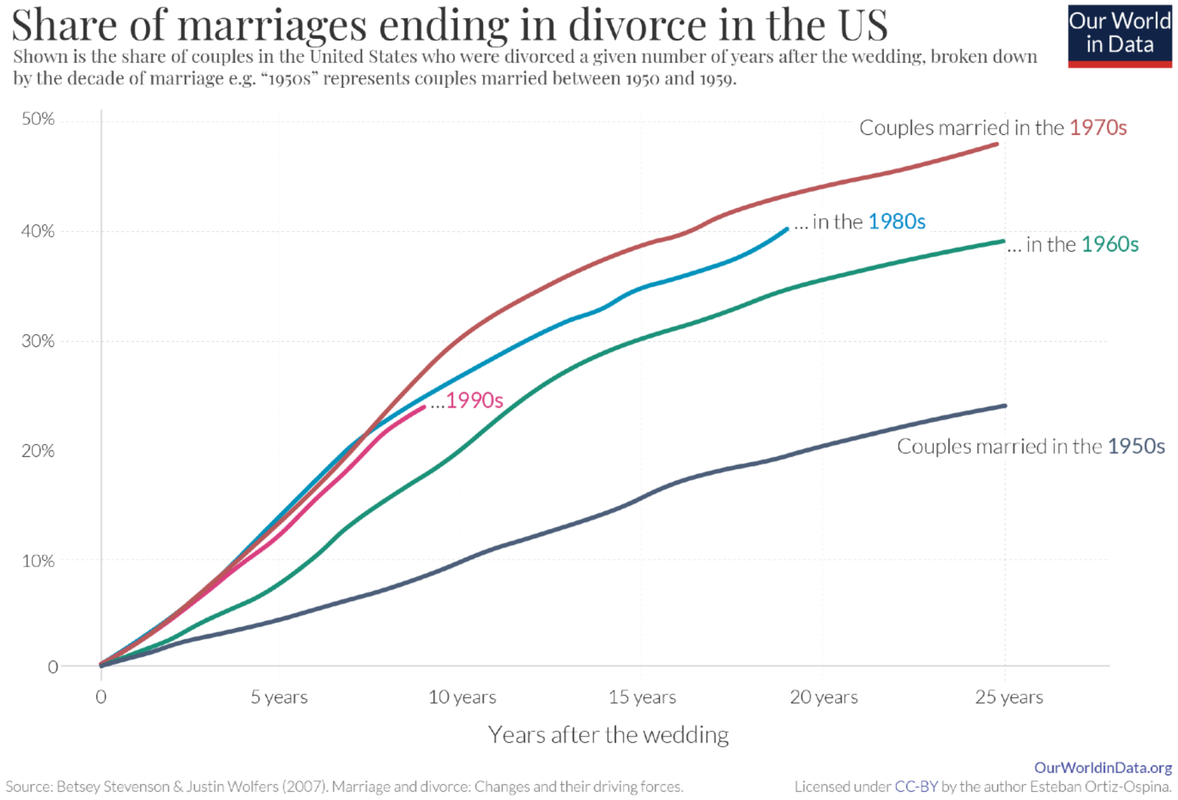
How to draw a histogram

Solution

To construct the histogram, begin by marking the class boundaries along the x-axis. The width of each bar will be the class width, from the lower class boundary to the upper class boundary. The height of the bar is the frequency of the class, just as it is for any bar graph. For this example, the first class would have a bar with a width of 5,000 from 49,999.5 to 54,999.5, and a height of 3. Repeat this process for each class, remembering that the bars in a histogram should touch.

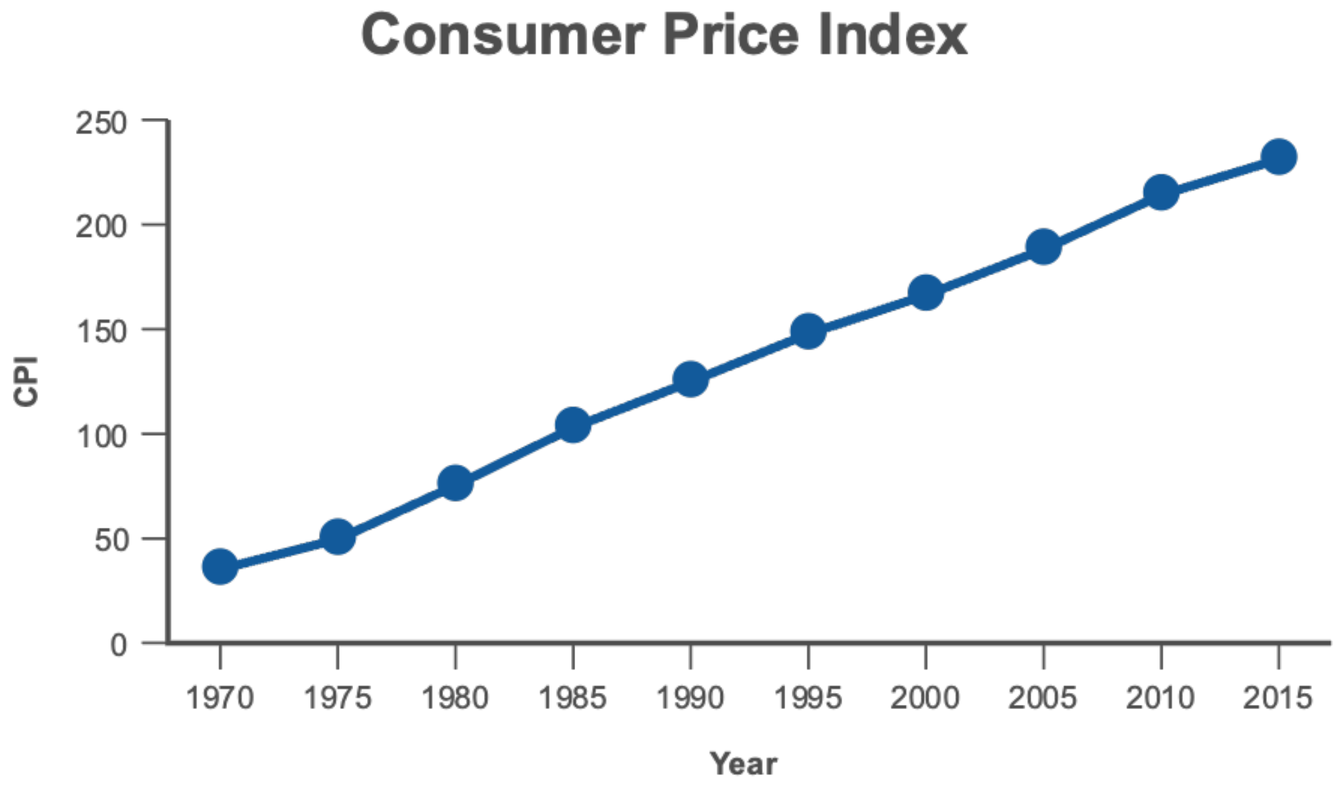
Line graphs

- A **line graph** is a graph that uses points to represent data values at a particular time in history and then joins the points with line segments. It displays changes in a quantitative variable over time.



How to draw a line graph

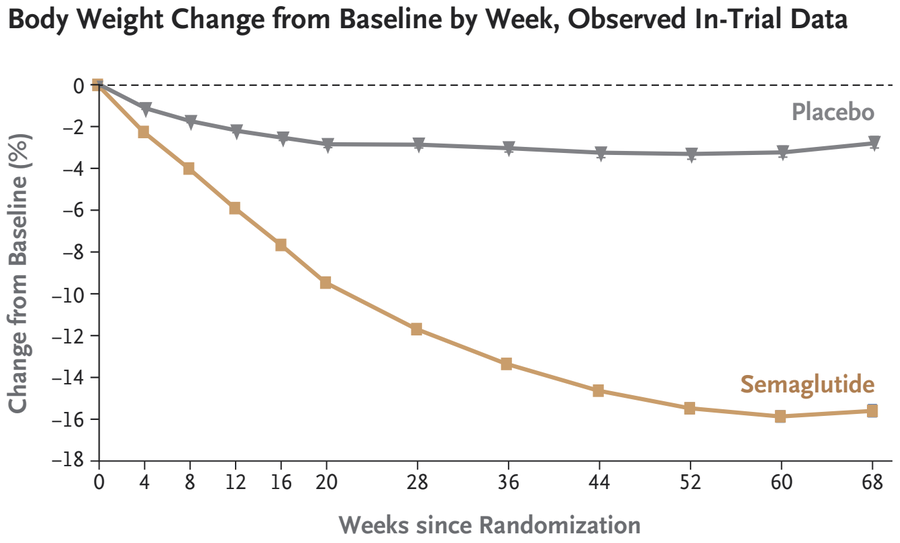
Consumer Price Index in January	
Year	CPI
1970	37.8
1975	52.1
1980	77.8
1985	105.5
1990	127.4
1995	150.3
2000	168.8
2005	190.7
2010	216.7
2015	233.7



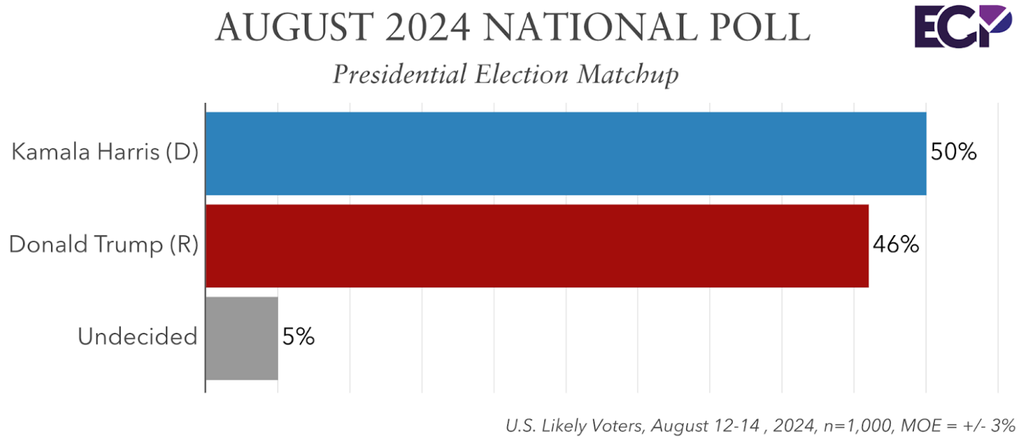
2.3 Analyzing Graphs

Broad categories of graphs

- A **time-series graph** is a line graph that is used to display a variable whose value changes over time
- A **cross-sectional graph** displays information collected only at one point in time (or an average over a short window of time)



Source: https://www.nejm.org/doi/full/10.1056/NEJMoa2032183

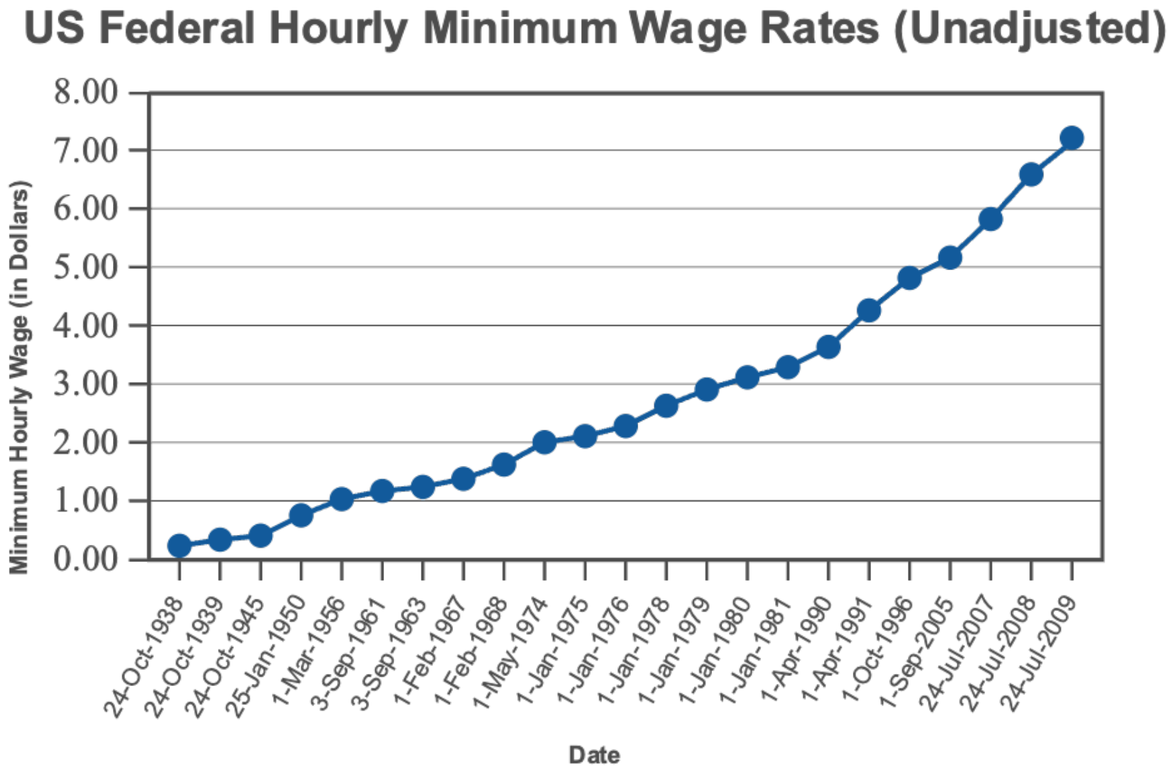


Source: https://emersoncollegepolling.com/august-2024-national-poll-harris-50-trump-46/

What's wrong with this graph?

Example 2.3.1: Scaling of Graphs

Consider the following graph on US federal minimum hourly wage rates, unadjusted for inflation, for various times between 1938 and 2009, when the minimum wage reached its current rate of $7.25 an hour. What errors can you find in the graph? How should they be fixed?



The shape of a graph is

- The shape of a graph
  - **uniform**, if the frequency of each class is relatively the same
  - **symmetric**, if a vertical line can be drawn to give approximate mirror images
  - **skewed** if the majority of the data falls on the left or right side of the distribution
    - skewed to the **right** if a majority of data falls on the left side
    - skewed to the **left** if a majority of the data falls on the right sid
- An **outlier** is a data value that falls outside the shape of the distribution

Case Study: Bitcoin Transactions

- Bitcoin users rate the trustworthiness of other users
- What do you notice about the data?
  - The rating -10 is an outlier
  - Not symmetric (or asymmetric)
  - Skewed to the left (most of data on the right)