Graphical Methods for Describing Data Distributions

STAT 250

Lecture 1B

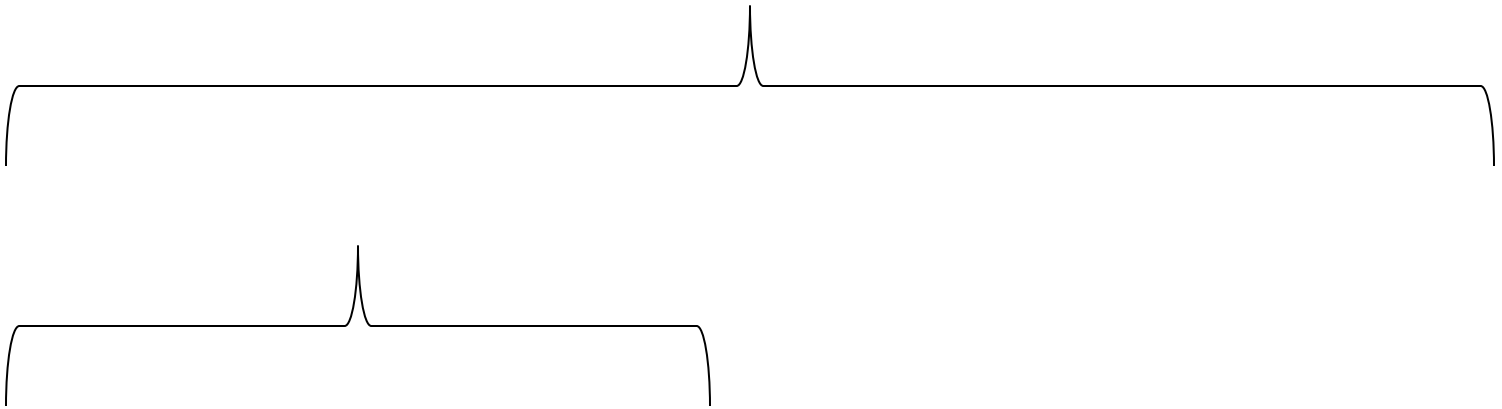
Learning outcomes

- Identify types of data
- Identify types of variables
- Choose appropriate graphical displays for each data/variable type
- Constructing graphical displays
- Describing key features of a distribution based on the graphical display

Data and Variables

- **Variable** : any characteristic whose value may change from one individual or unit to another
- **Data:** The values for a variable from individual observations
  - **Univariate** – consist of observations on a single variable made on individuals in a sample or population
  - **Bivariate** - data that consist of pairs of numbers from two variables for each individual in a sample or population
  - **Multivariate** - data that consist of observations on two or more variables

P.E. Student Data



Height Weight Sit-ups

Mile time

Jump distance

Types of Variables

- Categorical variables (Qualitative) - Consist of categorical responses
  - Examples

- Numerical variables (Quantitative) - Observations or measurements take on numerical values

Two Types of Numeric Variables

- Discrete numeric - Isolated points along a number line, usually **counts** of items. Can list possible values.
  - Examples –

- Continuous numeric - Variable that can be any value in a given interval, usually measurements of something
  - Examples –

- What about age or weight?

Identify the type of variable:

- the color of cars in a lot
- the number of calculators owned by students at your college
- the zip code of an individual
- the amount of time it takes students to drive to school
- the appraised value of homes in your city

Graphical Displays

Graphical Display	Variable Type	Data Type	Purpose
Bar Chart	Univariate	Categorical	Display data distribution
Comparative Bar Chart	Univariate for 2 or more groups	Categorical	Compare 2 or more groups
Dotplot	Univariate	Numerical	Display data distribution
Comparative dotplot	Univariate for 2 or more groups	Numerical	Compare 2 or more groups
Stem-and-leaf display	Univariate	Numerical	Display data distribution
Comparative stem-and-leaf	Univariate for 2 groups	Numerical	Compare 2 or more groups
Histogram	Univariate	Numerical	Display data distribution
Scatterplot	Bivariate	Numerical	Investigate relationship between 2 variables
Time series plot	Univariate, collected over time	Numerical	Investigate trend over time

Distributions

- The **distribution** of a variable describes how often the possible responses occur.
- A **frequency distribution** of a categorical variable is a listing of all categories along with their frequencies (_____)
- A **relative frequency distribution** is a listing of all categories along with their relative frequencies, expressed as a _____ or _____.

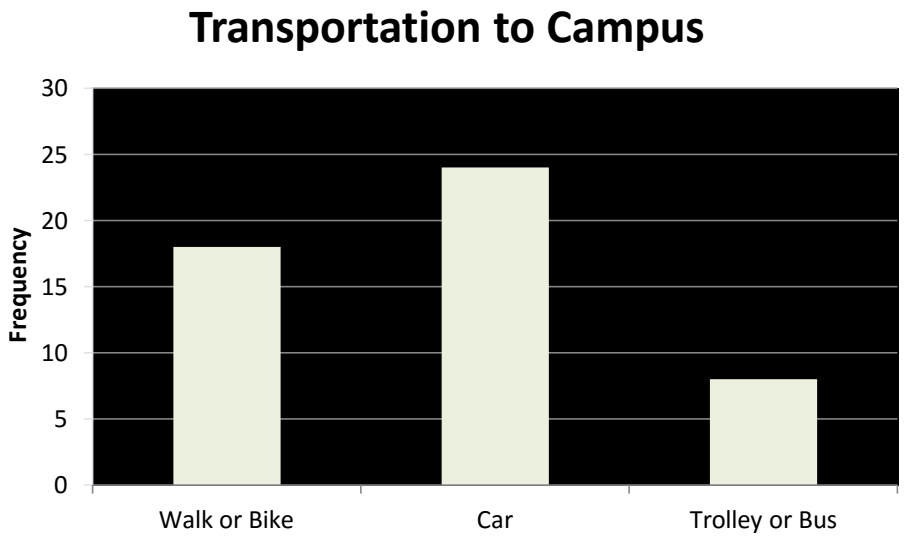
Distribution Example

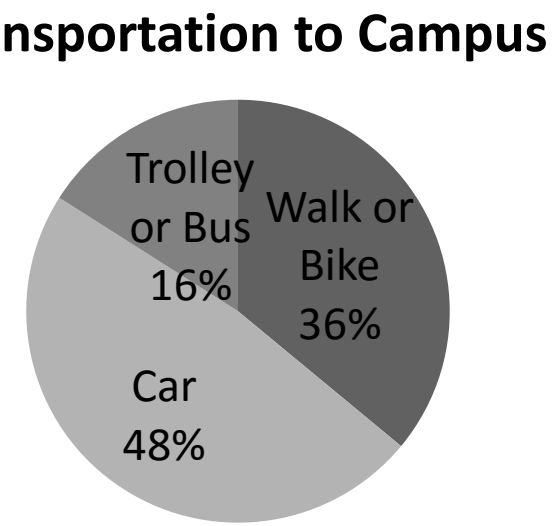
Transportation	Walk or Bike	Car	Trolley or Bus
Frequency	18	24	8
Relative Frequency			

Transportation	Walk or Bike	Car	Trolley or Bus
Male Frequency	5	12	3
Male Relative Freq			
Female Frequency	13	12	5
Female Relative Freq			

Bar Charts

Bar Charts are useful for summarizing one categorical variable.



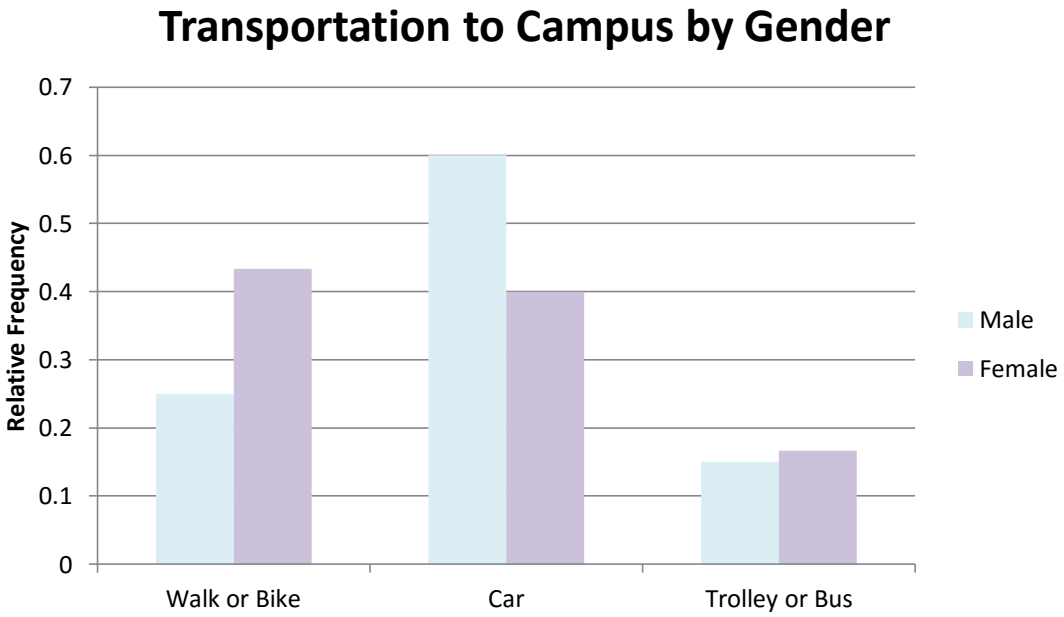


How would a relative frequency chart differ?

Comparative Bar Charts

Comparative Bar Charts are particularly useful for comparing the distribution of a categorical variable across groups.

- Use relative frequencies as group size may differ



Displaying Numerical Data

- Small data sets
  - Dot plot
  - Stem-and-leaf display
- Larger data sets
  - Histogram
  - Box Plot (later in course)

Dotplots

To create a dotplot:

- Draw a horizontal number line that covers the range from the smallest to the largest data value.
- Place a dot above the number line located at each observation's data value.
- When there are multiple observations with the same value, the dots are stacked vertically.

Illustration – year of minting of pennies

1971 1975 1977 1978 1979 1980 1982 1984 1986 1986 1988 1988 1992 1996 1996 1998 1999 2000 2003 2003 2004 2005 2006 2007 2007 2007 2007 2008 2008 2009 2009 2010 2011 2013

Stem-and-leaf Displays

To make a stemplot:

- **Sort the data.** Separate each observation into a **leaf**, which is the final digit, and the **stem**, which consists of the remaining digits.
- Write the stems in a column in consecutive order, with the smallest at the top, and draw a vertical line to the right of this column.
- Write each leaf in the row to the right of its stem, with the leaves in consecutive order, smallest to largest.

Illustration: Make a stem-and-leaf plot of points scored by the basketball team in the 2010-2011 season

51, 55, 56, 56, 60, 61, 62, 63, 64, 66, 66, 68, 68, 68, 69, 70, 71, 71, 71, 72, 74, 77, 77, 77, 79, 79, 80, 81, 83, 85, 85, 87, 88, 90, 92, 93, 96

Variations on the stem-and-leaf

- **Split stems -** If there are too many observations in a stem, you can split all of the stems into equal sized pieces.
  - For instance, with the scores, write each first digit (5-9) twice, and write leaves 0-4 after the first and leaves 5-9 after the second.
- **Comparative Stem and Leaf Displays** - To compare two distributions, use the same stems, but write the leaves for the second distribution on the left side.
  - For instance, we could have put opponents' scores on the other side of the stem-and-leaf plot in the example.
  - Be sure to label each side.

Histograms

To create a histogram:

- Decide how many equally spaced intervals to use for the horizontal axis. Between 6 and 15 is usually appropriate. Break the range of your data up into these intervals.
- Decide whether to use frequencies or relative frequencies on the vertical axis.
- Determine the frequency or relative frequency of data values in each interval and draw a bar with the corresponding height. If a value is on a boundary, count it in the interval that begins with that value.

San Diego Rainfall Histogram

Year	Rain	Year	Rain	Year	Rain
2002	3.3	1968	7.86	1969	11.48
2007	3.85	2012	7.9	1982	11.85
2014	5.09	1971	8.03	1991	12.31
1996	5.18	1981	8.13	1988	12.44
2004	5.18	2001	8.57	1992	12.48
2006	5.36	1965	8.81	2011	12.7
1984	5.37	2009	9.15	1966	14.76
2000	5.75	1977	9.18	1979	14.93
1989	5.88	1987	9.3	1986	14.95
1972	6.12	1985	9.6	1980	15.62
1970	6.33	1994	9.93	1995	17.13
1999	6.5	1976	10.14	1998	17.16
2013	6.55	2003	10.31	1978	17.3
1974	6.59	2010	10.6	1993	18.26
2008	7.2	1975	10.64	1983	18.49
1990	7.62	1967	10.86	2005	22.6
1997	7.73	1973	10.99		

Inches	Count	Rel. Freq.
2-3.99		
4-5.99		
6-7.99		
8-9.99		
10-11.99		
12-13.99		
14-15.99		
16-17.99		
18-19.99		
20-21.99		
22-23.99		

Describing Univariate Numeric Data

What to look for in dotplots, stem-and-leaf displays, and histograms

- A representative or typical value (center) in the data set
- The extent to which the data values spread out
- The nature of the distribution (shape) along the number line
- The presence of unusual values (gaps and outliers)

Shape

Terms we use to describe the shape of a distribution include

- **Symmetric** – the distribution looks similar on both sides. One symmetric shape we will encounter frequently is a **bell-shape**.
- **Skewed to the right** – there is more data on the left side of the distribution, the _____ tail is longer.
- **Skewed to the left** – there is more data on the right side of the distribution, the _____ tail is longer.
- The **mode** of a dataset is the most frequent value. If there is one prominent peak the shape is **unimodal**. If there are two prominent peaks, the shape is_____.

Describe the shapes of the distributions of coin dates, basketball scores, and rainfall. What shape do you expect for incomes? For grades?

Displaying Bivariate Numerical Data

- How to construct a **scatterplot**:
  - Draw horizontal and vertical axes. Label the horizontal axis and include an appropriate scale for the *x*-variable. Label the vertical axis and include an appropriate scale for the *y*-variable.
  - For each (*x*, *y*) pair in the data set, add a dot in the appropriate location in the display.
- What to look for: Relationship between *x* and *y*

Scatterplot Example

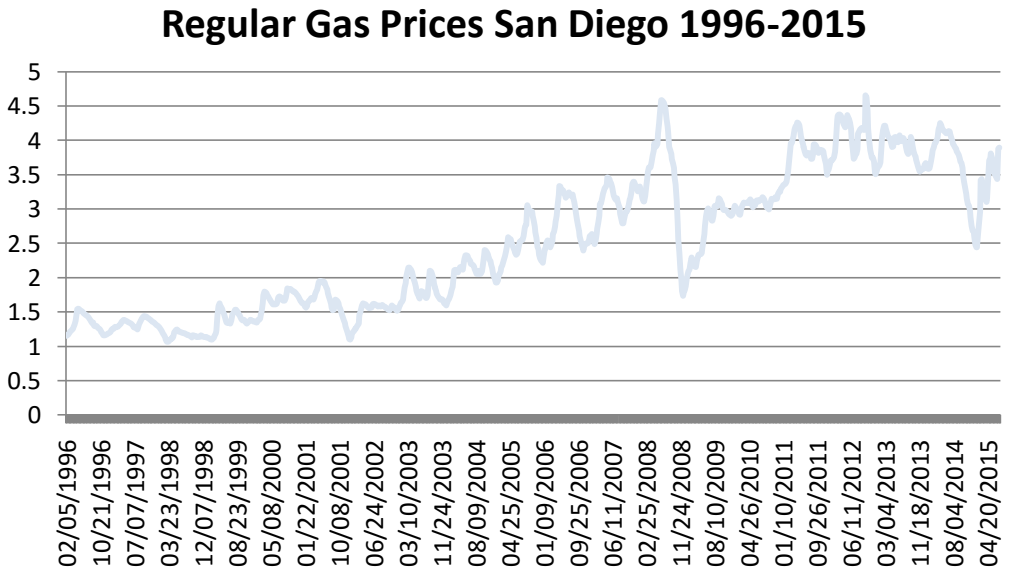
The accompanying table gives the cost (in dollars) and an overall quality rating for 10 different brands of men's athletic shoes (www.consumerreports.org).

Cost	65	45	45	80	110	110	30	80	110	70
Rating	71	70	62	59	58	57	56	52	51	51

Is there a relationship between x = cost and y = quality rating?

Time Plots

- Bivariate data with time and a numeric variable
- How to construct
  - Draw horizontal and vertical axes. Label the horizontal axis and include an appropriate scale for the *x*-variable. Label the vertical axis and include an appropriate scale for the *y*-variable.
  - For each (*x*, *y*) pair in the data set, add a dot in the appropriate location
  - Connect dots in order



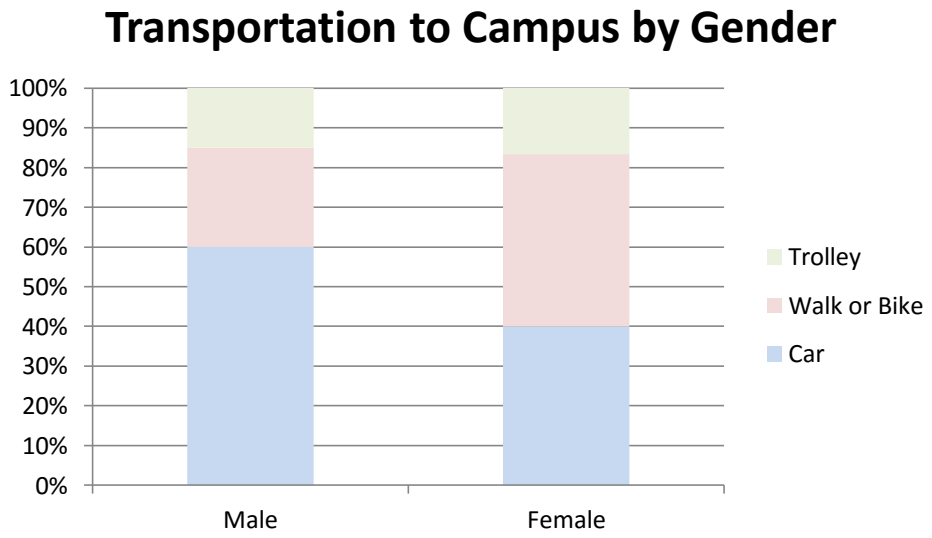
Other Charts

Pie (Circle) Chart

- When to Use: Categorical data
- How to construct
  - A circle is used to represent the whole data set.
  - “Slices” of the pie represent the categories
  - The size of a particular category’s slice is proportional to its frequency or relative frequency.
- Most effective for summarizing data sets when there are not too many categories

Segmented (or Stacked) Bar Charts

- When to Use: Categorical data
- How to construct
  - Use a rectangular bar rather than a circle to represent the entire data set.
  - The bar is divided into segments, with different segments representing different categories.
  - The area of the segment is proportional to the relative frequency for the particular category.



Avoid these Common Mistakes

- Areas should be proportional to frequency, relative frequency, or magnitude of the number being represented.

- Watch out for unequal time spacing in time series plots.

- Be sure you have the right type of plot for your data type

Broken Axes

- Be cautious of graphs with broken axes (axes that don't start at 0).
- Not starting at zero is OK for a scatterplot; does not result in a misleading picture of the relationship of bivariate data.
- In time series plots, broken axes can sometimes exaggerate the magnitude of change over time.
- In bar charts and histograms, the vertical axis should **NEVER** be broken. This violates the "proportional area" principle.

Scatterplots

- Be careful how you interpret patterns in scatterplots.
  - Consider the following scatterplot showing the relationship between the number of Methodist ministers in New England and the amount of Cuban rum imported into Boston from 1860 to 1940 (Education.com). r = .999973