Chapter 2 Classroom Notes

Section 2.1 Organizing Qualitative Data

Objectives

1. Organize Qualitative Data in Tables

2. Construct Bar Graphs

3. Construct Pie Charts

When data is collected from a survey or designed experiment, they must be organized into a manageable form. Data that is not organized is referred to as **raw data.**

Ways to Organize Data
- Tables
- Graphs
- Numerical Summaries (Chapter 3)

❶ Organize Qualitative Data in Tables

A **frequency distribution** lists each category of data and the number of occurrences for each category of data.

The **relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$Relative\ frequency = \frac{frequency}{sum\ of\ all\ frequencies}$$

A **relative frequency distribution** lists each category of data together with the relative frequency.

Example 1 Organizing Qualitative Data into a Frequency Distribution

In StatCrunch, join the Group "SullyStats".   Do this by selecting Explore > Groups. Search the group above and join.   Go into the group and download the data set "Cereal Brands".

Construct a frequency distribution and relative frequency distribution of manufacturer (mfr).
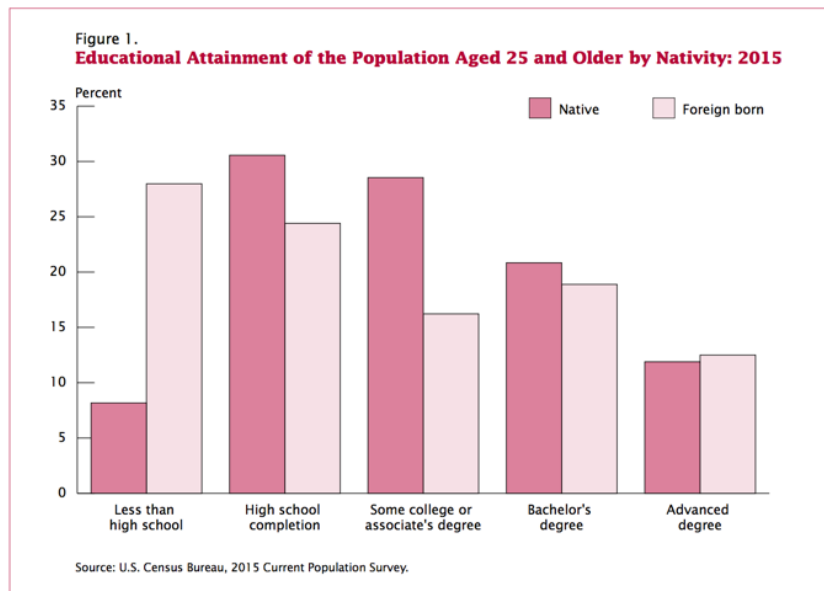
A **bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.

Example   Constructing a Frequency and Relative Frequency Bar Graph

Use the Cereal Brands data to construct a frequency and relative frequency bar graph of manufacturer.

Side-by-Side Bar Graphs

Suppose we want to compare the educational attainment of males and females. We could draw a **side-by-side bar graph** to compare the data for the two genders. The data sets should be compared by using relative frequencies rather than frequencies, because different sample or population sizes make comparisons using frequencies difficult or misleading.

Figure 1.
Educational Attainment of the Population Aged 25 and Older by Nativity: 2015

Source: U.S. Census Bureau, 2015 Current Population Survey.

Example Constructing Side-by-Side Bar Graphs

In StatCrunch (SullyStats Group), open the data set "Educational Attainment". Draw a side-by-side relative frequency bar graph of the data. What does the graph suggest?

Row	Educational Attainment	Male	Female	RelFreq_Male	RelFreq_Female
1	Less than High School	11337	11204	0.10866898	0.099506199
2	High School Diploma	31296	31216	0.29998275	0.27723898
3	Some College, no degree	16930	18525	0.16227978	0.16452627
4	Associate's degree	9651	12659	0.0925081	0.11242851
5	Bachelor's degree	21906	24357	0.20997642	0.21632207
6	Master's degree	9049	11543	0.086737726	0.10251696
7	Professional degree	1765	1407	0.016918122	0.012496003
8	Doctoral degree	2392	1685	0.022928129	0.014965008
9					
10	Sum	104326	112596		
11					

❸ Construct Pie Charts

A **pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

Example   Drawing a Pie Chart

Draw a pie chart of "manufacturer" for the cereal brand data.

2.2 Organizing Quantitative Data: The Popular Displays
Objectives

1. Organize discrete data in tables

2. Construct histograms of discrete data

3. Organize continuous data in tables

4. Construct histograms of continuous data

5. Draw dot plots

6. Identify the shape of a distribution

The first step in summarizing quantitative data is to determine whether the data are discrete or continuous. If the data are discrete and there are relatively few different values of the variable, the categories of data (**classes**) will be the observations (as in qualitative data). If the data are discrete, but there are many different values of the variables, or if the data are continuous, the categories of data (the **classes**) must be created using intervals of numbers.

❶ Organize Discrete Data in Tables

Example 1   Constructing a Frequency and Relative Frequency Distribution from Discrete Data

Open the "SullivanStatsSurveyI.txt" file in StatCrunch. Construct a frequency and relative frequency distribution of the variable "Number of Televisions".

A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.

Example 2   Drawing a Histogram for Discrete Data

Construct a frequency and relative frequency histogram of the "Number of Televisions" data from Example 1.

❸ Organize Continuous Data in Tables

Classes are categories into which data are grouped. When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, we must create classes by using intervals of numbers.

See the table on the next page.

The **lower class limit** of a class is the smallest value within the class while the **upper class limit** of a class is the largest value within the class. The lower class limit of first class is 25. The lower class limit of the second class is 35. The upper class limit of the first class is 34. The **class width** is the difference between consecutive lower class limits. The class width of the data given above is 35 − 25 = 10.

The data are continuous. So, the class 25 to 34 actually represents 25 to 34.999…, or 25 up to every value less than 35.

Classes do not overlap. This table is **open ended** because the last class has no upper class limit.

Table 1.

Educational Attainment of the Population Aged 25 and Older by Age, Sex, Race and Hispanic Origin, and Other Selected Characteristics

(Numbers in thousands)

Characteristic	Total	High school graduate or more		Some college or more		Associate's degree or more		Bachelor's degree or more		Advanced degree	
		Percent	Margin of error[1] (±)	Percent	Margin of error[1] (±)	Percent	Margin of error[1] (±)	Percent	Margin of error[1] (±)	Percent	Margin of error[1] (±)
Population 25 and older	212,132	88.4	0.3	58.9	0.5	42.3	0.5	32.5	0.5	12.0	0.3
Age											
25 to 34	43,006	90.5	0.6	65.0	0.9	46.5	0.9	36.1	1.0	10.9	0.6
35 to 44	39,919	88.7	0.5	62.8	0.9	46.7	1.0	36.3	1.0	13.8	0.7
45 to 64	83,213	89.4	0.4	59.0	0.7	42.6	0.7	32.0	0.7	12.1	0.5
65 and older...................	45,994	84.3	0.7	49.7	0.9	34.1	0.9	26.7	0.8	11.3	0.7
Sex											
Male.........................	101,888	88.0	0.4	57.6	0.7	41.2	0.7	32.3	0.6	12.0	0.4
Female.......................	110,245	88.8	0.3	60.1	0.6	43.4	0.6	32.7	0.6	12.0	0.4
Race and Hispanic origin											
White alone	168,420	88.8	0.3	59.2	0.6	42.8	0.6	32.8	0.6	12.1	0.3
Non-Hispanic White alone	140,638	93.3	0.3	63.8	0.6	46.9	0.7	36.2	0.7	13.5	0.4
Black alone	25,420	87.0	0.9	52.9	1.4	32.4	1.4	22.5	1.2	8.2	0.7
Asian alone	12,331	89.1	1.2	70.0	1.9	60.4	2.0	53.9	2.0	21.4	1.5
Hispanic (of any race)	31,020	66.7	1.1	36.8	1.0	22.7	0.9	15.5	0.7	4.7	0.4
Nativity Status											
Native born	175,519	91.8	0.3	61.3	0.5	43.3	0.6	32.7	0.6	11.9	0.3
Foreign born	36,613	72.0	1.0	47.6	1.1	37.6	1.1	31.4	1.1	12.5	0.7
Disability Status											
With a disability	28,052	78.6	0.9	41.6	1.2	24.9	1.0	16.7	0.9	5.7	0.5
Without a disability..............	183,351	89.9	0.3	61.5	0.5	45.0	0.6	34.9	0.5	12.9	0.3

Example Organize Continuous Data into a Frequency and Relative Frequency Distribution

Open the "Cereal Brands" data in StatCrunch. Construct a frequency and relative frequency distribution of calories using a lower class limit of the first class equal to 40 and a class width of 10.

Example Drawing a Histogram

Draw a frequency and relative frequency histogram of calories for the Cereal Brand data. Use a lower class limit of the first class equal to 40 and a class width of 10.

Exploration In StatCrunch, open the Cereal Brand data.   Now select Applets > Histogram with sliders.   Under "Select column:" choose "calories". Click Compute! Use the slider to adjust the starting point and Bin width (aka Class Width). Describe how the histogram changes.

❻ Draw Dot Plots

A **dot plot** is drawn by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.

Example   Drawing a Dot Plot

Draw a dot plot of the "Number of Televisions" data from the SullivanStatsSurveyI.txt data set.

❼ Identify the Shape of a Distribution

**Uniform distribution -** the frequency of each value of the variable is evenly spread out across the values of the variable

Bell-shaped distribution - the highest frequency occurs in the middle and frequencies tail off to the left and right of the middle

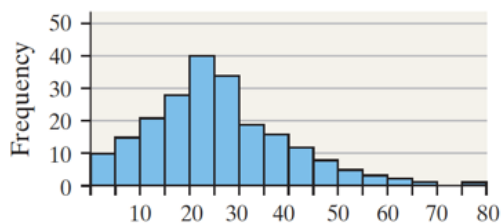**Skewed right -** the tail to the right of the peak is longer than the tail to the left of the peak

Skewed left - the tail to the left of the peak is longer than the tail to the right of the peak.
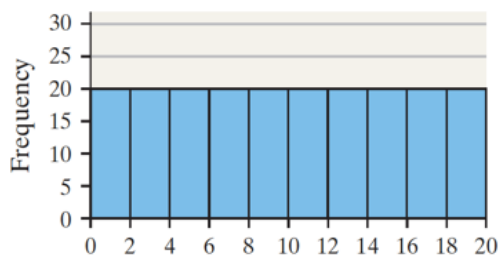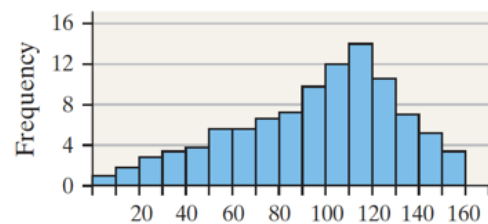

(a) Uniform (symmetric)
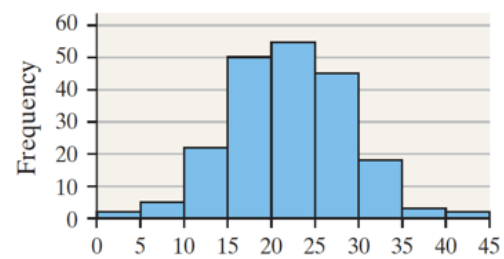
(b) Bell-shaped (symmetric)

(c) Skewed Right

(d) Skewed Left

Example Identify Distribution Shape

Draw a histogram of the variable "Cell phone bill" in the SullivanStatsSurveyI.txt data set. Identify the shape of the distribution.