Stat 154: Elementary Statistics

Jongyun Jung

Minnesota State University, Mankato

Ch 2: **Descriptive Statistics**

jongyun.jung@mnsu.edu

January 15, 2019

Overview

Organizing and Displaying Data

- **Raw data**: when data are collected in original form, they are called raw data.
- By using the raw data, we can draw conclusions and make inferences about events.
- However, the raw data must be represented in some meaningful way. We can use **frequency table**.

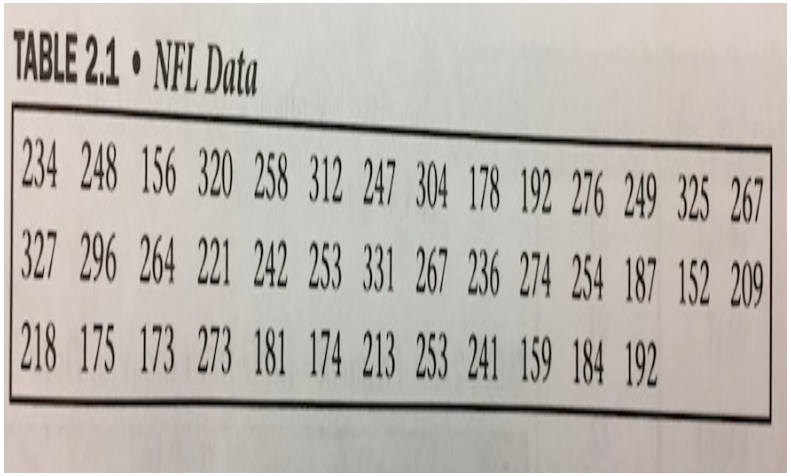


TABLE 2.1 • *NFL Data*

234	248	156	320	258	312	247	304	178	192	276	249	325	267
327	296	264	221	242	253	331	267	236	274	254	187	152	209
218	175	173	273	181	174	213	253	241	159	184	192		

Figure: 40 National Football League (NFL) player's weights (in pounds)

Frequency Table

- **Frequency Table**: a table including different classes (or groups) and their respective frequencies (or counts within the classes)
- A frequency table can be constructed using the following steps:
  1. Step 1 Determine the number of classes or groups depending on the sample size n. A rule of thumb is $c \approx ln(n)$, where c is the number of classes, and 'ln' stands for the natural logarithm.
  2. Step 2 Determine the range of the data as *Range* = *Maximum* - *Minimum*.
  3. Step 3 Determine the width or length of the class intervals as *Width* = *Range* / c. Consider a slightly higher value of the *Width* than it actually is so that when the classes are formed they include all the data values.
  4. Step 4 Determine the classes by introducing an extra decimal position so that each data value belongs to exactly one class.
  5. Step 5 Use tally marks to put all data values in different classes and count the tallies to find the frequencies for each class.

Frequency Table - Cont.

- **Frequency Table**: Example: Construct the frequency table for the NFL data in table 2.1
- A frequency table can be constructed using the following steps:
  1. Step 1 $c = ln(n) = ln(40) = 4$.
  2. Step 2 *Range* = *Maximum* - *Minimum*. 331 - 152 = 179
  3. Step 3 *Class Width* = *Range* / c = 179 / 4 $\approx$ 45.
  4. Step 4 The classes are 152-196, 197-241, 242-286 and 287-331.
  5. Step 5 The class boundaries are 151.5-196.5, 196.5-241.5, 241.5-286.5 and 286.5-331.5

Class Limits	Class Boundaries	Tally	Frequency
152–196	151.5–196.5	𝍸 𝍸 ||	12
197–241	196.5–241.5	𝍸 ||	7
242–286	241.5–286.5	𝍸 𝍸 ||||	14
287–331	286.5–331.5	𝍸 ||	7

Figure: Frequency Table from the NFL Data in Table 2.1

Frequency Table - Cont.

- **Lower class limits**: the smallest numbers that can belongs to different classes. In NFL example, the values 152, 197, 242, and 287 are the lower class limits.
- **Upper class limits**: the largest numbers that can belongs to different classes. In NFL example, the values 196, 241, 286, and 331 are the upper class limits.
- **Class boundaries**: the numbers used to separate the classes but without the gaps.
  - To obtain the class boundaries, find the gaps created by class limits.
  - Subtract the half of the gap to each of the lower limits and add the half of the gap to each of the upplimits.
- In NFL example, the gap between class limits is 1.
- So we subtract 0.5 to each of the lower limits.
- And add 0.5 to each of the upper limits
- Then, 151.5, 196.5, 241.5, 286.5 and 331.5 as the class boundaries.

Frequency Table - Cont.

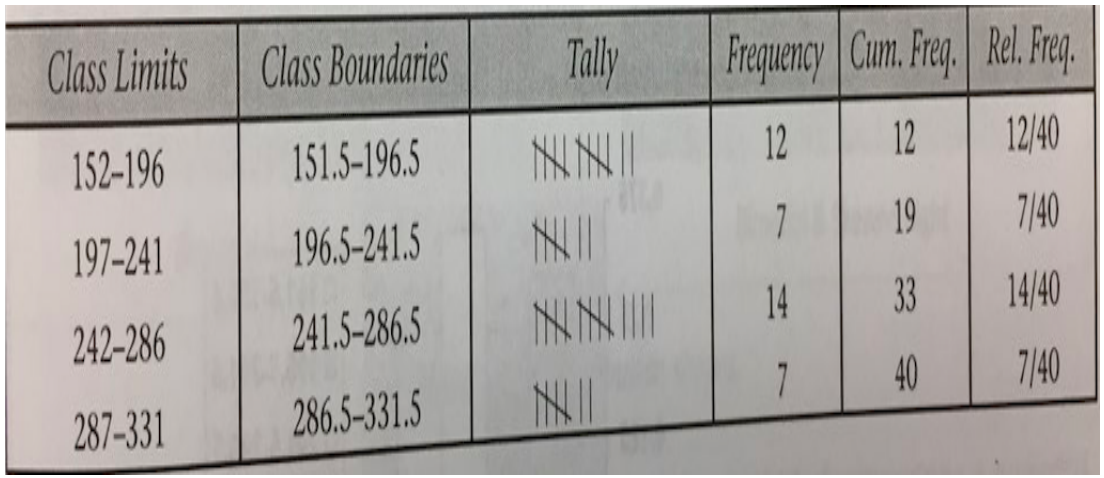
- **Class midpoints**: midpoints of each class. It can be found by adding the lower class boundary to the upper class boundary and dividing the total by 2.
  - The class midpoints in NFL example are 174, 219, 264 and 309.
- **Class width**: the difference between two consecutive lower class limits or two consecutive lower class boundaries.
  - The class midpoints in NFL example are 45.

Additional Guidelines for Constructing a Frequency Table

1. Usually there are between 5 to 20 classes to avoid having too few or too many classes.
2. The classes are mutually exclusive (or non-overlapping).
3. The classes are continuous.
4. The classes are exhaustive (accommodate all the data).
5. The classes are equal in width.

Cumulative Frequency Tables

- **Cumulative frequency** is the sum of the frequencies for that class and all previous classes.



Class Limits	Class Boundaries	Tally	Frequency	Cum. Freq.	Rel. Freq.
152–196	151.5–196.5	𝍸	12	12	12/40
197–241	196.5–241.5	𝍸	7	19	7/40
242–286	241.5–286.5	𝍸 ||||	14	33	14/40
287–331	286.5–331.5	𝍸	7	40	7/40

Figure: The Cumulative Frequency (Cum. Freq.) and the Relative Frequency (Rel. Freq.) Table from the NFL Data

Relative Frequency

- **Relative frequency** are obtained by dividing the class frequency by the total number of observations, which is the sum of all frequencies.

$$RelativeFrequency = \frac{Frequency}{TotalFrequency}$$

$$\text{R.F.} = \frac{f}{n}, \text{ where } n = \sum f$$

Class Limits	Class Boundaries	Tally	Frequency	Cum. Freq.	Rel. Freq.
152–196	151.5–196.5	𝍸 𝍸 |	12	12	12/40
197–241	196.5–241.5	𝍸 |	7	19	7/40
242–286	241.5–286.5	𝍸 𝍸 ||||	14	33	14/40
287–331	286.5–331.5	𝍸 |	7	40	7/40

Figure: The Cumulative Frequency (Cum. Freq.) and the Relative Frequency (Rel. Freq.) Table from the NFL Data

Graphical Displays

- It is easier for most people to understand the meaning of data presented graphically than data presented numerically in tables of frequencies.
- Stem-Leaf Display, Bar Graph, Pie Chart, Histogram, Relative Frequency Histogram, Box-plot, etc

Histogram

- It is a graph of bars for a variable having at least ordinal measurements.
- It is a graph of bars in which the horizontal scale represents classes of the data value and the vertical scale represent frequencies.

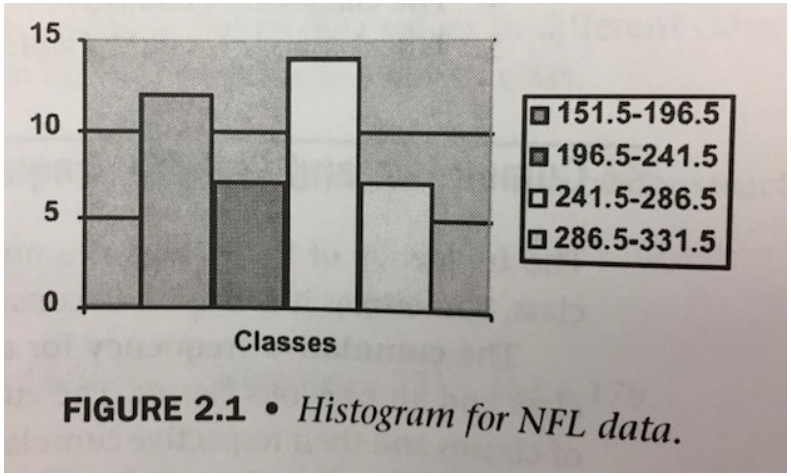


Figure: Histogram for NFL data

Relative Frequency Histogram

- It is a histogram with the vertical scale as relative frequencies instead of as actual frequencies.

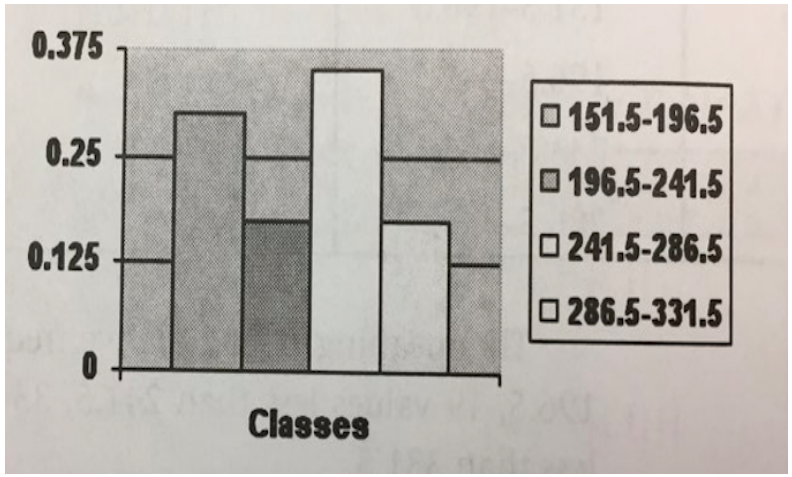


Figure: Relative frequency Histogram for NFL data

Distribution Shapes

- Shapes of the distributions help in determining the appropriate statistical methods to be used to analyze the data.
- Some of the shapes are uniform symmetrical, symmetrical, right or left skewed, bell shaped and symmetric, bimodal, etc.

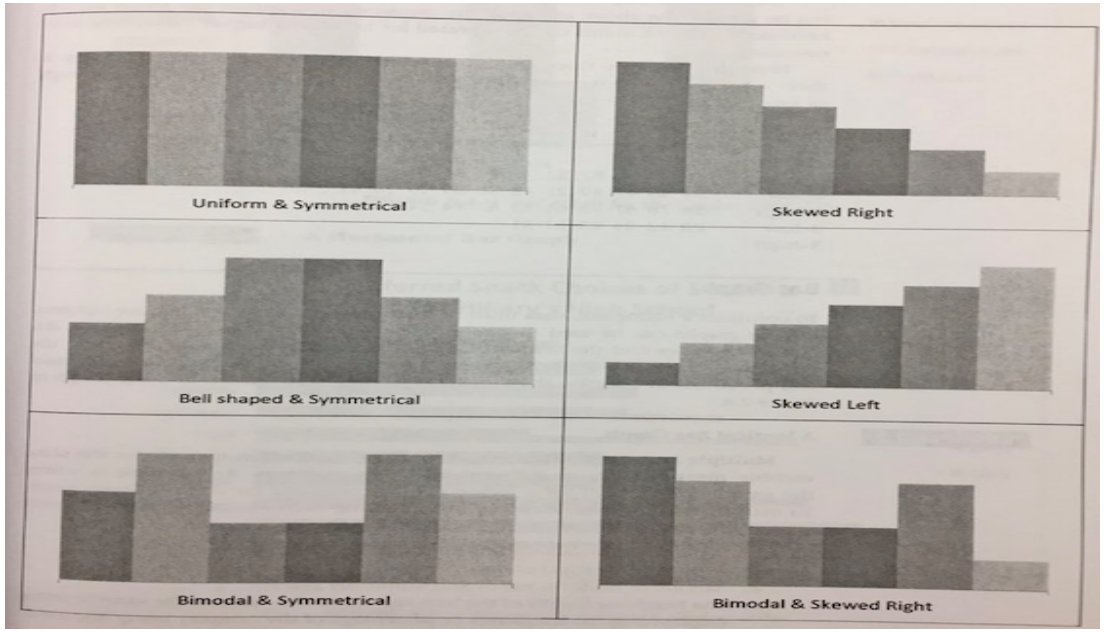
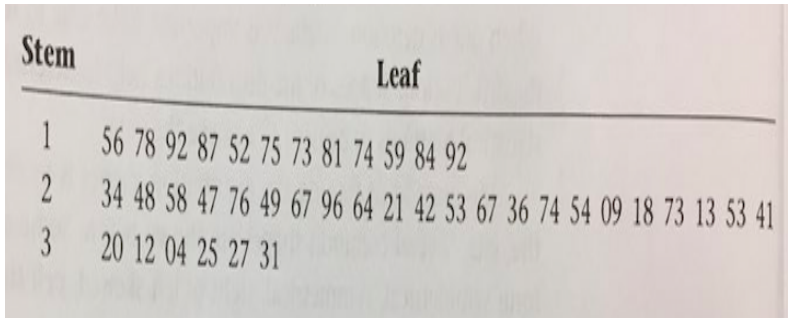


Figure: Distribution Shapes

Stem-leaf display

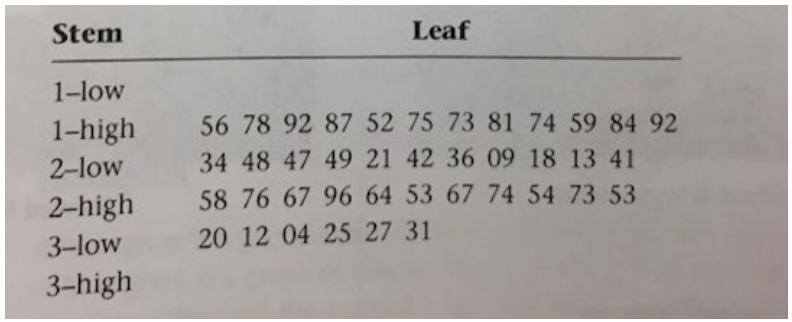
- the simplest method of summarizing a numerical data set when the number of observations in the data is not too large.



Stem	Leaf
1	56 78 92 87 52 75 73 81 74 59 84 92
2	34 48 58 47 76 49 67 96 64 21 42 53 67 36 74 54 09 18 73 13 53 41
3	20 12 04 25 27 31

Figure: Stem-leaf display

- If there is a need to have more groups, each group could be divided as low and high.



Stem	Leaf
1–low	
1–high	56 78 92 87 52 75 73 81 74 59 84 92
2–low	34 48 47 49 21 42 36 09 18 13 41
2–high	58 76 67 96 64 53 67 74 54 73 53
3–low	20 12 04 25 27 31
3–high	

Figure: Stem-leaf display

Bar Graph

- If the data is quantitative, then we can use a histogram. However, if the data is qualitative (categorical), bar graphs can be used to disply.

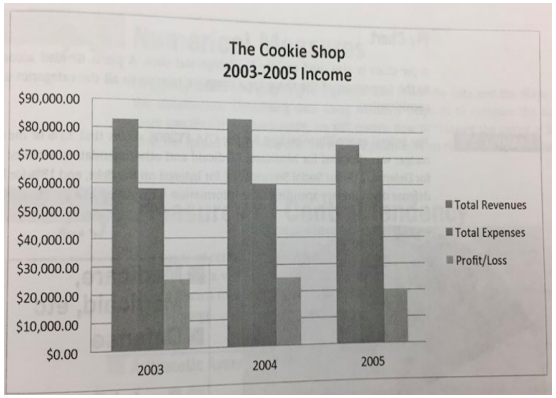


Figure: A Vertical Bar Graph for the Cookie Shop data

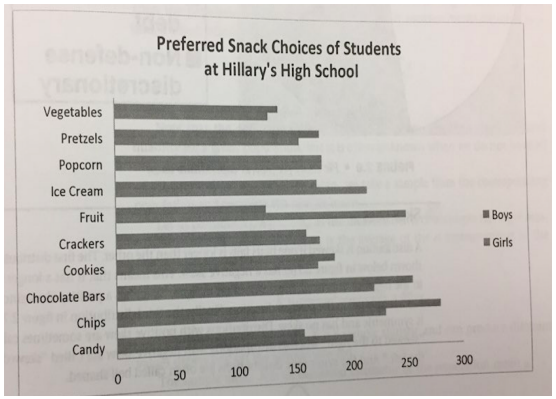


Figure: A Horizontal Bar Graph for the snack choices data

Pie Graph

- It is also used mainly for categorical data.
- A pie is divided according to the proportion of the share for a category relative to all the categories under consideration.

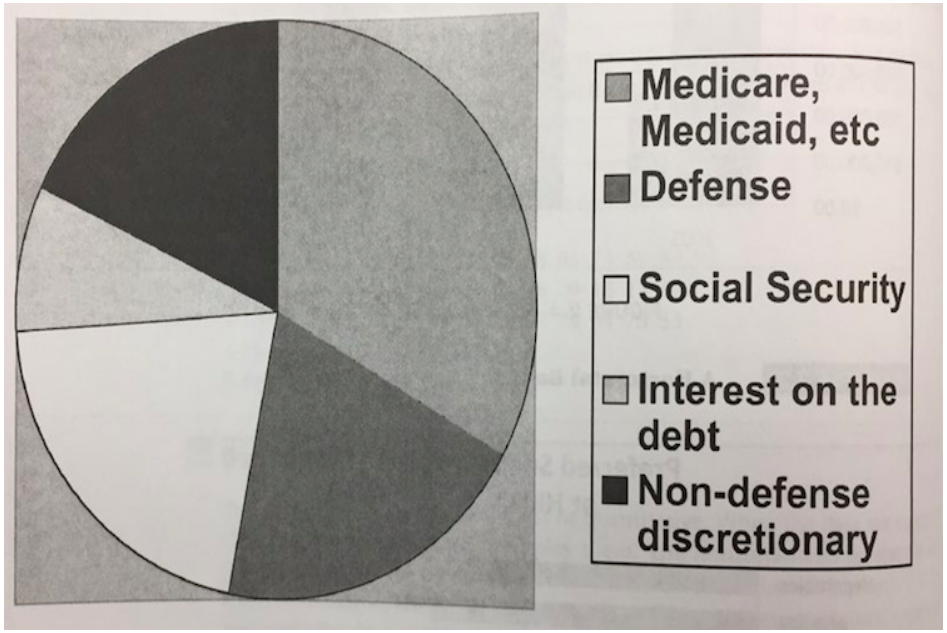


Figure: Pie Chart for federal expenditure data

Skewness

- A distribution is skewed if one of its tails is longer than the other.
- Distributions with positive skew are sometimes called "skewed to the right" whereas distributions with negative skew are called "skewed to the left".
- The symmetric distributions are often called bell shaped.

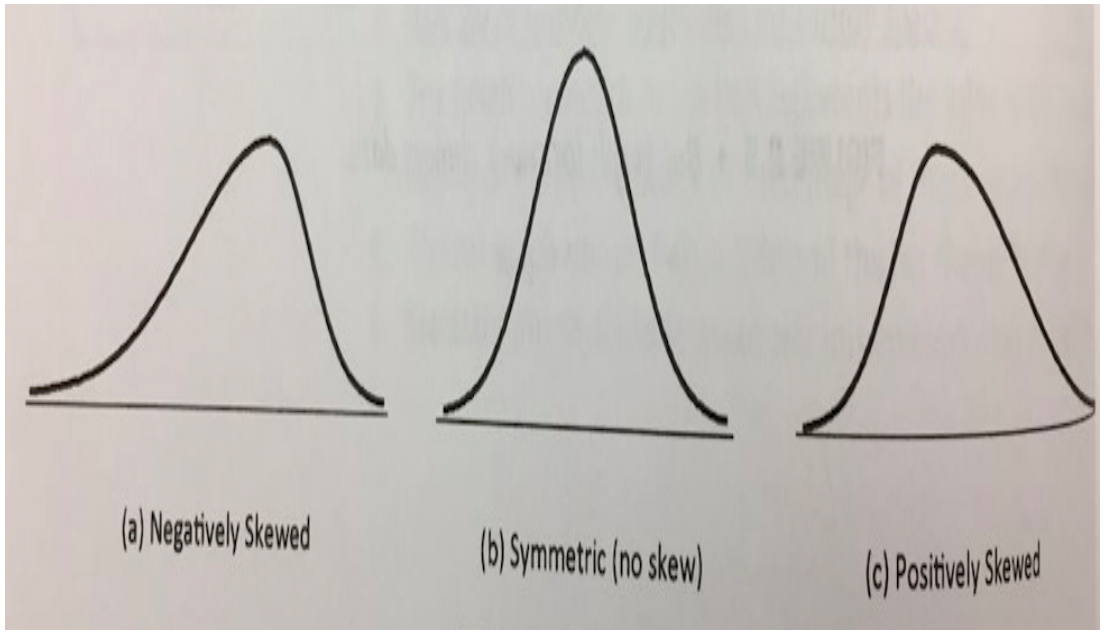


Figure: Skewness

References

Mezbahur Rahman, Deepak Sanjel, Han Wu. Statistics Introduction, Revised Printing

KendallHunt