Chapter 2 Descriptive Statistics Notes

STA 2023 SECTION 2.1 AND 2.3: FREQUENCY DISTRIBUTIONS AND THEIR GRAPHS

Learning Outcomes:

1) Construct a frequency distribution including limits, midpoints, relative frequencies, cumulative frequencies, and boundaries
2) Interpret frequency histograms.

Notes:

After data is collected, it needs to be **organized** and **described** in order to interpret any results.

The purpose of frequency distribution is to separate data into categories (organize) and to find the number of data values that fall into each category(describe).

____________________: the number of times a category occurs in the data set.

______________________________: a table that presents the frequency of each category.

→ A frequency distribution (or frequency table) presents each category along with its frequency (the number of times that value occurs in a data set).

An example of one is shown below:

Class	Frequency, *f*
1 – 5	5
6 – 10	8
11 – 15	6
16 – 20	8
21 – 25	5
26 – 30	4

In a numerical frequency distribution, each class has a **lower-class limit**, which is the _______________ number that can belong to the class and an **upper-class limit**, which is the __________________ number that can belong to the class. The **class width** is the difference between consecutive lower (or upper) class limits (when a frequency table is given without the data set).

If instead of a frequency distribution table, a real data set (numerical) is given, the calculation of the class width formula is quite different:

$$Class\ Width = \frac{\max value - \min value}{\#\ of\ classes}$$

**always round up to next convenient unit of number.

Chapter 2 Descriptive Statistics Notes

Example 1: The data set lists the cell phone screen times (in minutes) for 30 U.S. adults on a recent day. Construct a frequency distribution that has seven classes.

200	239	155	252	384	165	296	405	303	400
307	241	256	315	330	317	352	266	276	345
238	306	290	271	345	312	293	195	168	342

Range=

Class Width=

Chapter 2 Descriptive Statistics Notes

______________________________: the frequency of a category divided by the sum of all frequencies.

______________________: a table that presents the relative frequency of each category.

_________________________: is the sum of the frequencies of that class and all previous classes.

_________________________: is the sum of the lower and upper limits of the class divided by two. The midpoint is sometimes called the *class mark*.

$$Mid - point = \frac{Lower\ class\ limit + Upper\ class\ limit}{2}$$

$$Relative\ Frequency = \frac{Class\ Frequency}{Sample\ size\ (total\ frequency)}$$

Example 2: Find the relative and cumulative frequencies for the frequency distribution, round relative frequencies to two decimal places.

Credit Card	Frequency	Relative Frequency	Cumulative Frequency
Master Card	11		
Visa	23		
American Express	9		
Discover	7		

Graphical Summaries of a Data

1. **Bar Graph:** a graphical representation of a frequency or relative frequency distribution. A bar graph consists of rectangles of equal width, with one rectangle for each category.
   In a bar graph, the horizontal axis represents _____________and the vertical axis represents _______________

   Example 03: Draw a bar graph based on the example 2 frequency distribution:

2. **Histogram:** A **frequency histogram** uses bars to represent the frequency distribution of a data set. A histogram has the following properties.
   - **a.** The horizontal scale is quantitative and measures the data entries (classes).
   - **b.** The vertical scale measures the frequencies of the classes.
   - **c.** Consecutive bars must touch.

** A RF histogram can be drawn in a similar way where RF will be on the vertical axis and classes will be on the horizontal axis.

Shapes of Histograms or Bar Graphs

i. Symmetric – taller in the middle and the left half is the mirror image of right half

ii. Skewed to the left or negatively skewed – lower on the left → rises to right

iii. Skewed to the right or positively skewed – taller on the left→falls to right

Example 04: Draw a frequency histogram based on the example 1 data set on the cell phone screen times (in minutes) for 30 U.S. adults on a recent day.

200 239 155 252 384 165 296 405 303 400

307 241 256 315 330 317 352 266 276 345

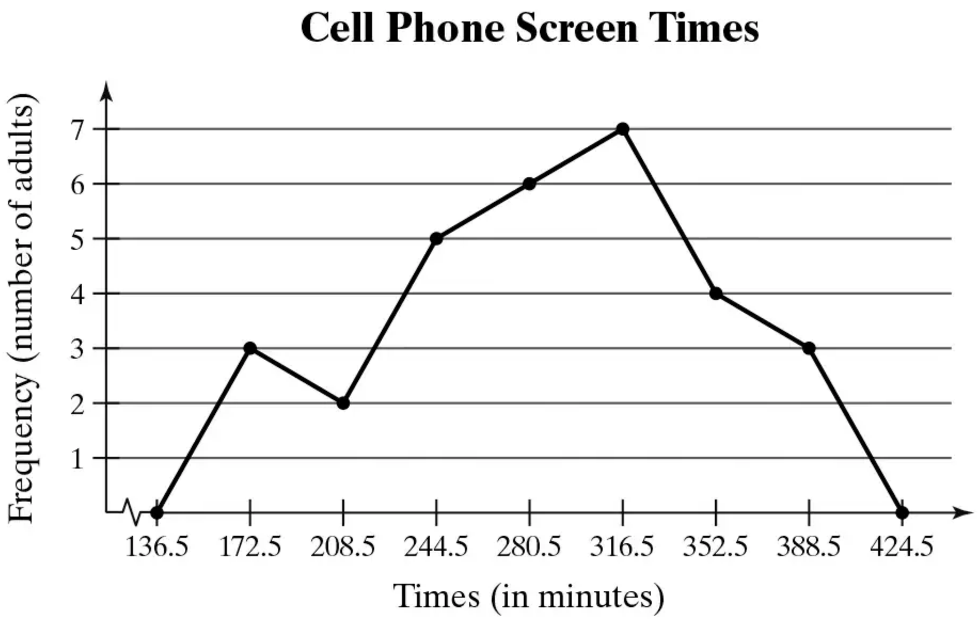
238 306 290 271 345 312 293 195 168 342

Class limit	Frequency
155 - 190	3
191 - 226	2
227 - 262	5
263 -298	6
299 - 334	7
335 - 370	4
371 - 406	3

3. **Frequency polygon:** A **frequency polygon** is a line graph that emphasizes continuous frequency change. To construct a frequency polygon, we denote class midpoints on the horizontal axis and frequency on the vertical axis.

   Draw a frequency polygon based on the example 1 data set on the cell phone screen times (in minutes) for 30 U.S. adults on a recent day.

Class limit	Frequency	Class midpoint
155 - 190	3	172.5
191 - 226	2	208.5
227 - 262	5	244.5
263 -298	6	280.5
299 - 334	7	316.5
335 - 370	4	352.5
371 - 406	3	388.5



4. **Stem and Leaf plots:** are useful for small data sets – gives more detail than a frequency (or relative frequency) distribution
   - The rightmost digit will represent the LEAF (coming off a STEM)
   - The remaining digits will be the STEM
   - if too many values fall in the same stem, you can divide the stem (0 – 4, 5 – 9)
   - LEAVES should be in the order of magnitude
   - Any stem and leaf plot should include a KEY, especially for decimal values.

Tips: Order the data from lowest to highest before drawing a stem and leaf plot.

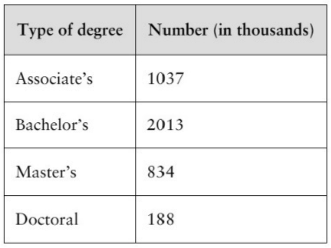
Example 01: A random sample of 30 animated Disney movies was collected and the number of minutes in each was noted. Construct a Stem and Leaf Plot for the 30 animated Disney movies:

70	83	64	76	79	81	71	73	80	92
77	88	69	74	82	81	83	72	78	64
71	120	71	92	89	77	80	79	78	75

5. **Pie Chart:** display relative frequency information using sectors of a circle, with each sector representing a category.

- **Sector for each category = (RF) (360°)**

Example: The number of earned degrees conferred (in thousands) in 2019 is shown in the table at the right. Use a pie chart to organize the data. ***(Source: U.S. National Center for Education Statistics)***



Type of degree	Number (in thousands)
Associate's	1037
Bachelor's	2013
Master's	834
Doctoral	188

Interpretation From the pie chart, you can see that almost one-half of the degrees conferred in 2019 were bachelor's degrees.

Chapter 2 Descriptive Statistics Notes

STA 2023 SECTION 2.3 MEASURES OF CENTRAL TENDENCY

Learning Outcomes:

1) Find the mean, median, and mode of a population and of a sample

Notes:
In previous, we talked about ways to organize data. We will now discuss ways to describe data (aka descriptive statistics).

Numbers that represent what is central or typical for a data set are called **measures of central tendency.**

A Measure of Center is a value that indicates where the center or middle of a data set is located. These values are **all** considered to be averages.

1. **Arithmetic Mean** – usually called the **mean value** – also what is typically known as the average value – can be found by taking the sum of all data values and then dividing that sum by the number of values
   a. Sample Mean:
   b. Population Mean:

- There is an "equal distance" above and below the mean, so data values "balance" around the mean.
- Rounding rule: we are allowed to keep ONE decimal place past what the data has.

2. **Median** – "middle-most" number in a data set when values are placed in order of increasing or decreasing magnitude – separates the data into the lower half and the upper half – does NOT have to be one of the data values

   - If the number of values is odd –

   - If the number of values is even –

Chapter 2 Descriptive Statistics Notes

Example #1: The following values represent amounts (in millions of dollars) spent on media advertising in the USA for ten randomly chosen companies in the first quarter in a year.

295.9	280.3	362.9	394.6	463.5
722.5	286.9	331.2	257.0	340.1

Find the MEAN:

Find the MEDIAN:

If the value of 722.5 was listed incorrectly, how would this affect the mean? The median?

Comparing the properties of the Mean and the Median

- The mean uses every value in the data set. It may be greatly affected by extreme values. Extreme Value –

- The median does not use every value and is therefore not affected by extreme values.
- A statistic or parameter is called a resistant measure if the value is not greatly affected by extreme values in a data set. The _______________ is a resistant measure of center.

Describing the shape of a data set (a histogram) using the mean and median:

SHAPE: RELATIONSHIP:

- Skewed to the RIGHT

- Skewed to the LEFT

- Approximately SYMMETRIC

Chapter 2 Descriptive Statistics Notes

If a data set has a mean of 5.2 and a median of 9.7, describe the shape of the distribution.

*Mean and Median both use numbers to describe the center of numerical data, but do they always have practical meaning?

3. **Mode** – the mode of a data set is the value that appears most frequently – it is considered an average, but not really a measure of center
   - If one value appears most frequently –
   - If 2 values appear with the same maximum frequency –
   - If more than 2 values appear with the same maximum frequency –
   - If no values are repeated –
   - MODE may be used with both quantitative (numerical) **and** qualitative (categorical) data values. Examples:

For the following Stem and Leaf Plot, find the mean, median and mode. Then describe the shape of the distribution.

Stem	Leaf = one
1	2
2	0 7 7 9
3	7 8
4	1 3

Chapter 2 Descriptive Statistics Notes

SECTION 2.4 MEASURES OF VARIATION

Learning Outcomes:

1) Find the range, standard deviation, and variance of a data set

Measures of central tendency can only tell us so much about a data set. Often more information is needed to have a more complete picture of a data set.

The **measures of variation** are used to describe the spread, or variability, of data items in a data set.

Variation – the amount that data values vary from one another or from the mean – a measure of “distance”

1. **Range** – the difference between the maximum value and the minimum value = Max – Min
   - Easy to compute
   - Greatly affected by extreme values – Not a resistant measure of variation

Example: The selling price (in dollars) of a certain stock (Stock A) for 10 randomly chosen days is reported:

26 26 27 28 31 33 33 37 37 37

Find the mean:

Find the median:

Find the mode:

Find the range:

The selling price for a second stock (Stock B) is reported for 10 randomly chosen days:

3 12 18 22 27 37 37 47 52 60

Find the mean:

Find the median:

Find the mode:

Find the range:

2. **Standard Deviation –** a measure of variation of values about the mean – gives an "average distance" that data values are from the mean

Sample Standard Deviation **Population Standard Deviation**

- Units are the same as the data values
- Larger values of $s\ or\ \sigma$ indicate greater amounts of variation among the values
- Rounding rule – we are allowed to keep <u>one decimal place past what the data has</u> (same rule as the mean)
- Not a resistant measure of variation – somewhat affected by extreme values, but not as much as the range

Using and Understanding the Mean and the Standard Deviation: The Empirical Rule

The Empirical Rule – For data that has a bell-shaped distribution:

- Approximately ___________of values fall within ______ standard deviation of the mean
- Approximately ___________of values fall within ______ standard deviations of the mean
- Approximately ________________________________ of values fall within ______ standard deviations of the mean

Note: This rule **cannot** be used if the data has a distribution that is **not** bell-shaped. This rule applies to samples and populations **if** the distribution is bell-shaped.

Chapter 2 Descriptive Statistics Notes

Example: Heights of women have a bell-shaped distribution (normal distribution) with a mean of 161cm and a standard deviation of 7cm.

a. Find the range of values that contains approximately the middle 95% of women's heights.

b. Find the approximate percentage of women who have a height between 154 cm and 168 cm.

Chapter 2 Descriptive Statistics Notes

STA 2023 SECTION 2.5 MEASURES OF POSITION NOTES

Learning Outcomes:

1) Find the first, second, and third quartiles of a data set
2) Find the interquartile range of a data set and identify outliers.
3) Represent a data set graphically using a box-and whisker plot
4) Interpret other fractiles such as percentiles and how to find percentiles for a specific data entry
5) Determine and interpret the standard score (*z*-score)

Notes:

Measures of Central Tendency tell us the central or typical value in the data set**.** Measures of Variation tell us the amount of spread or variability in the data set. Both of these measures give us information about the entire data set.

Measure of Position – a measure of the relative position of a data value – where the value lies relative to other data values in a data set

1. **Z – score** – the number of standard deviations that a given value *x* is above or below the mean
   - There are **no** units of measure

     Sample **Population**

Example: A class took a psychology test that results in a mean score of 84 with a standard deviation of 4. Find the z-score for the following data values:

a. X = 78

b. X = 84

c. X = 96

d. If you scored a 76 on the psychology test and your friend scored a 100 on an English test where the class mean was 120 with a standard deviation of 15, whose score is higher relative to their class?

2. **Quartiles –** breaks the data into 4 groups with ~25% of the values in each group

- The first quartile_______ separates the lowest 25% of the values from the upper 75%. It is the ____________ of the lower half of the data.
- The second quartile__________ separates the lowest 50% of the values from the upper 50%. It is the ____________ of the data.
- The third quartile__________ separates the lowest 75% of the values from the upper 25%. It is the ____________ of the upper half of the data.
- These values can be found on the calculator by using **1-var stats.** (Part of the 5-number summary)

Another measure of **spread (variation)** is the **Interquartile Range.** Where **the range** uses the maximum and minimum values, it gives the range of **all** data values. The **interquartile range** gives the range of the **middle 50% of the data.** It can be found by taking the difference of ______ and _______.
So, **IQR =**

We can use the quartiles to determine if a given data value is an **extreme value or outlier** (considerably larger or smaller than most of the values in the data set). While there are differing methods for defining outliers, we will use The Interquartile Method.

- Find the Interquartile Range
- Compute the outlier boundaries – any value out beyond these boundaries is considered an outlier
  - Lower Outlier Boundary:
  - Upper Outlier Boundary:

Example:

For the given set of data, find the following.

9	15	7	2	4	4	3	4	3	4	25	4	3
12	2	8	3	2	2	6	7	3	10	4	5	4

a. Find the first and third quartiles:

b. Find the Interquartile Range:

c. Find the lower and upper outlier boundaries:

d. List all values that are classified as outliers or extreme values.

The Five Number Summary and Box Plots (Box and Whisker Plots)

The **Five Number Summary** of data consists of the minimum value, the first quartile, the second quartile (median), the third quartile and the maximum value.

A **boxplot** is a graph that presents the Five Number Summary along with the Outlier Boundaries.

Chapter 2 Descriptive Statistics Notes

Determining skewness from a boxplot:

- If the median is closer to Q_1 than to Q_3 or the upper whisker is longer than the lower whisker, data are skewed to the ___________.
- If the median is closer to Q_3 than to Q_1 or the lower whisker is longer than the upper whisker, data are skewed to the ___________.
- If the median is approximately halfway between $Q_1\ and\ Q_3$, and the lower and upper whiskers are approximately equal in length, the data are approximately _____________.

Describe the skewness of the set of data from above.

3. **Percentiles –** divide a data set into 100 groups with ~ 1% in each group – a percentile separates the lowest **p%** of values from the upper (**100 – p)%** of values. Example:

   Important percentiles to remember:

   - $P_{25} = Q_1$
   - $P_{50} = Q_2 = median$
   - $P_{75} = Q_3$