(G230 Course / Geology Dept.)

Descriptive Statistics

Continue within another topic of the Descriptive Statistics:

Measures of Central Location / Central Tendecny

- A measure of Central Tendency describes a set of data by identifying the central position in the data set as a single value (It indicates where the center or the most typical value of the variable lies in collected set of measurements).
- Measures of center are often referred to as averages, which use to compare between the dataset values by adopting single value lies between higher and lower date set values. Also, it considers a representative sample value for a specific population such as a single value (Average) represents the annual rainfall average for 50 years values of annual averages.
- There are three kinds of averages of a data set; Mean, Median and Mode. In different cases some measures are more appropriate to use than other.
- Median and Mean apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative data.

1) **Mean (Arithmetic Mean)** is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. It is the point in a distribution of measurements about which the summed deviations are equal to zero. When talking about an average, most often we referring to the mean. The symbol “μ” is used for the mean of a population, while the symbol “ $\bar{x}$ " or "M” used for the mean of a sample.

Case (1) Find $\sum X,\ \sum X^2\ and \sum (X-1)^2\ \ for\ the\ data\ set$: 1 3 4

Solution:

$$\Sigma x = 1+3+4 = 8$$
$$\Sigma x^2 = 1^2+3^2+4^2 = 1+9+16 = 26$$
$$\Sigma (x-1)^2 = (1-1)^2+(3-1)^2+(4-1)^2 = 0^2+2^2+3^2 = 13$$

The **sample mean** of a set of n sample data is the number $\bar{x}$ defined by the formula:

$$\bar{x} = \frac{\Sigma x}{n} \qquad \textbf{OR} \qquad \bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

It is used the lowercase *n* to denote the number of measurements in a sample, which is called the **sample size**.

Case (2) Find the mean of the sample data: 2 -1 0 2

Solution: $\bar{x} = \frac{\Sigma x}{n} = \frac{2+(-1)+0+2}{4} = \frac{3}{4}$ **= 0.75**

Case (3) Find the sample mean for a random sample of ten measurements (x).

1.90 3.00 2.53 3.71 2.12 1.76 2.71 1.39 4.00 3.33

Solution:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1.90 + 3.00 + 2.53 + 3.71 + 2.12 + 1.76 + 2.71 + 1.39 + 4}{10}$$

$$= \frac{26.45}{10} = 2.645$$

Case (4) A random sample of 19 students gave the following data, X is the number of failure to pass the course and *f* is the frequency (the number of times that occurred in the data set). Find the sample mean…

x	0	1	2	3	4
f	3	6	6	3	1

Solution:

In the case (4) above, the data are presented by the frequency table. Each number in the first line of the table is a number that appears in the data set; the number below it is how many times it occurs. This means that three students in the sample have had no failure of course pass, six have had exactly one time of failure, and so on. The explicit list of all the observations in this data set is therefore 0 0 0 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 4

The sample size can be read directly from the table, without first listing the entire data set, as the sum of the frequencies:
$n = 3 + 6 + 6 + 3 + 1 = 19$. The sample mean can be computed directly from the table as well:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{0 \times 3 + 1 \times 6 + 2 \times 6 + 3 \times 3 + 4 \times 1}{19} = \frac{31}{19} = 1.6316$$

In the examples above the data sets were described as samples. Therefore the means were sample means, denoted by $\bar{x}$. If the data come from a census, so that there is a measurement for every element of the population, then the mean is calculated by exactly the same process of summing all the measurements and dividing by how many of them there are, but it is now the *population mean* and is denoted by μ, the lower case Greek letter mu.

The **population mean** of a set of N population data is the number μ defined by the formula:

$$\mu = \frac{\Sigma x}{N}$$

The mean of two numbers is the number that is halfway between them. For example, the average of the numbers 5 and 17 is $(5 + 17)/2 = 11$, which is 6 units above 5 and 6 units below 17. In this sense the average 11 is the "center" of the data set {5,17}. For larger data sets the mean can similarly be regarded as the "center" of the data.

Mean for grouped data can be calculated used the midpoints of the classes.

Case (5) For the following frequency distribution find the mean?

- Find the midpoints of each class such as:

$X_m = \frac{5.5 + 10.5}{2} = 8$ $\frac{10.5 + 15.5}{2} = 13$ and so on ….

- Multiply each midpoint by the frequency for each class such as:

$1 \cdot 8 = 8$ $2 \cdot 13 = 26$ and so on ….

- Find the summation of f . X_m
- Finally divide the summation of f . X_m by the sample size or number of measurements (n or N).

Class boundaries	Frequency
5.5 – 10.5	1
10.5 – 15.5	2
15.5 – 20.5	3
20.5 – 25.5	5
25.5 – 30.5	4
30.5 – 35.5	3
35.5 – 40.5	2
	n = 20

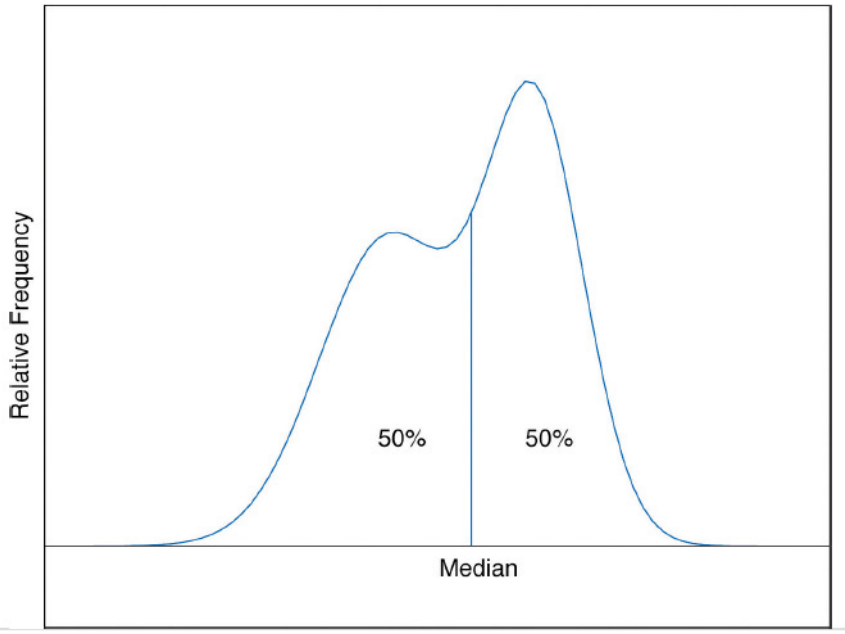
Midpoint (X_m)	f . X_m
8	8
13	26
18	54
23	115
28	112
33	90
38	76
	$\sum f \cdot X_m = 490$

2) Median

The median is another concept of average. It is the halfway point in a data set of sample, which can be adopted to find the Outlier data within data set. To find the median first; the data must be arranged in order and then find the value locates in the center of sample size. It has been denoted by $\tilde{x}$

24.8 22.8 24.6 192.5 25.2 18.5 23.7 → Raw Data set

18.5 22.8 23.7 24.6 24.8 25.2 192.5 → Outlier data (Anomalies)



Case (6) The number of precipitation times in Basra over last seven years is 10, 7, 9, 14, 12, 10, and 8. Find the median?

Solution:

1. Arrange the data in order:

7, 8, 9, 10, 10, 12, 14

2. Select the middle value

7, 8, 9, 10, 10, 12, 14 → $\tilde{x}$

Median

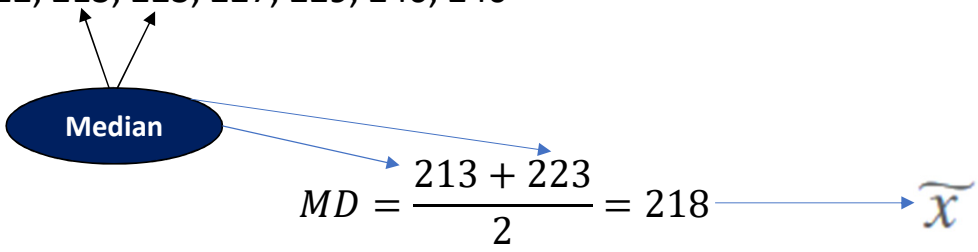
When there is an **even** number of values in the data set, the median will fall between two given values as illustrated in the following examples.

Case (7) The number of cloudy days for the top ten cloudiest cities is shown. Find the median?

209, 223, 211, 227, 213, 240, 240, 211, 229, 212

Solution:

209, 211, 211, 212, 213, 223, 227, 229, 240, 240

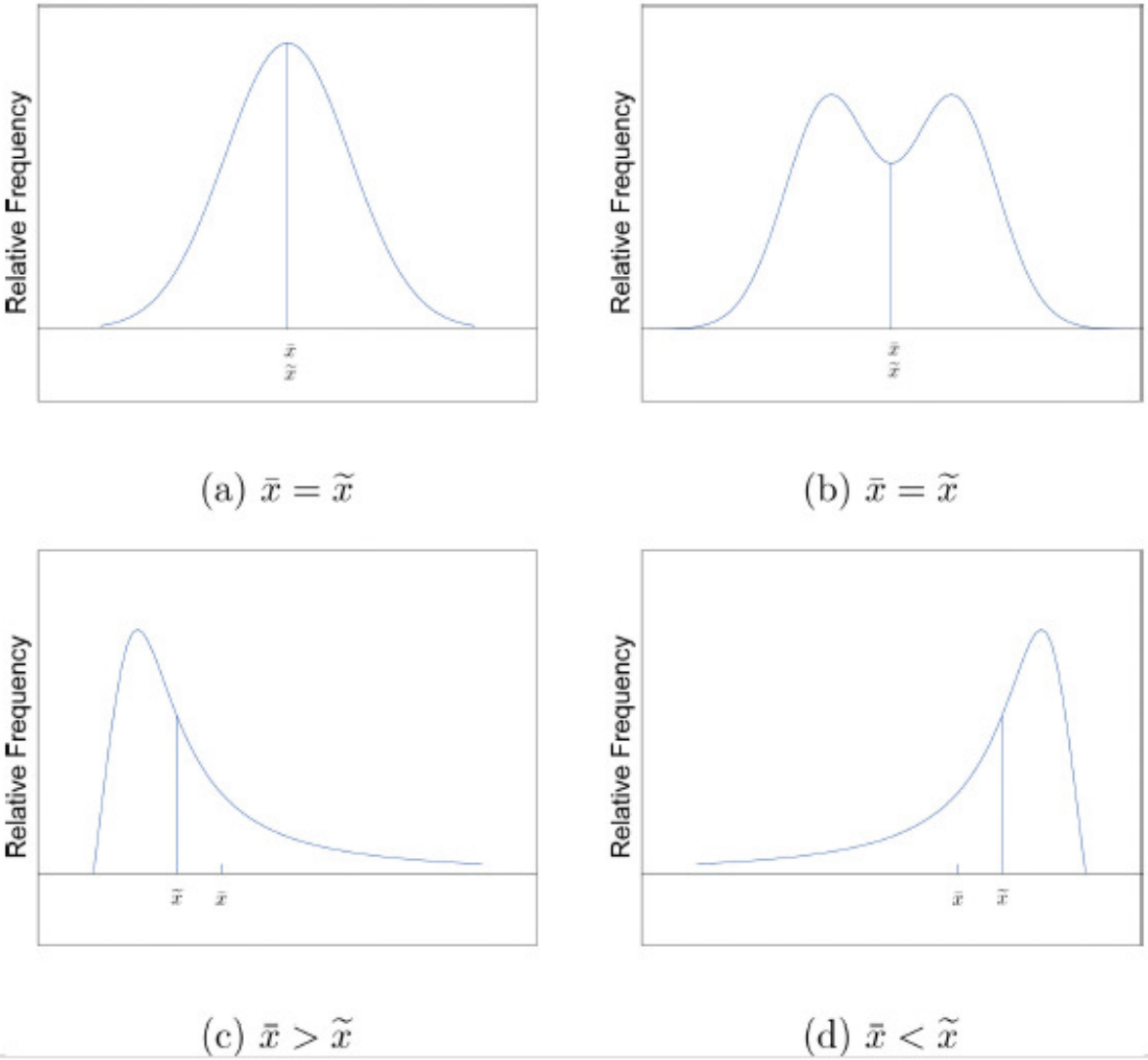


The **sample median**[7] $\tilde{x}$ *of a set of sample data for which there are an odd number of measurements is the middle measurement when the data are arranged in numerical order. The sample median* $\tilde{x}$ *of a set of sample data for which there are an even number of measurements is the mean of the two middle measurements when the data are arranged in numerical order.*

The population median is defined in similar way of sample median

The relation between the mean and the median for several common shapes of distributions is shown by the figures below:

- Figures (a and b) show symmetric distributions of Relative Frequency Histograms. Noted the mean and the median are equal.
- Figure (c) shows skewed right. The mean has been pulled to the right of the median by the long "right tail" of the distribution, the few relatively large data values.
- *Figure (d) shows Skewed left*. The mean has been pulled to the left of the median by the long "left tail" of the distribution, the few relatively small data values.



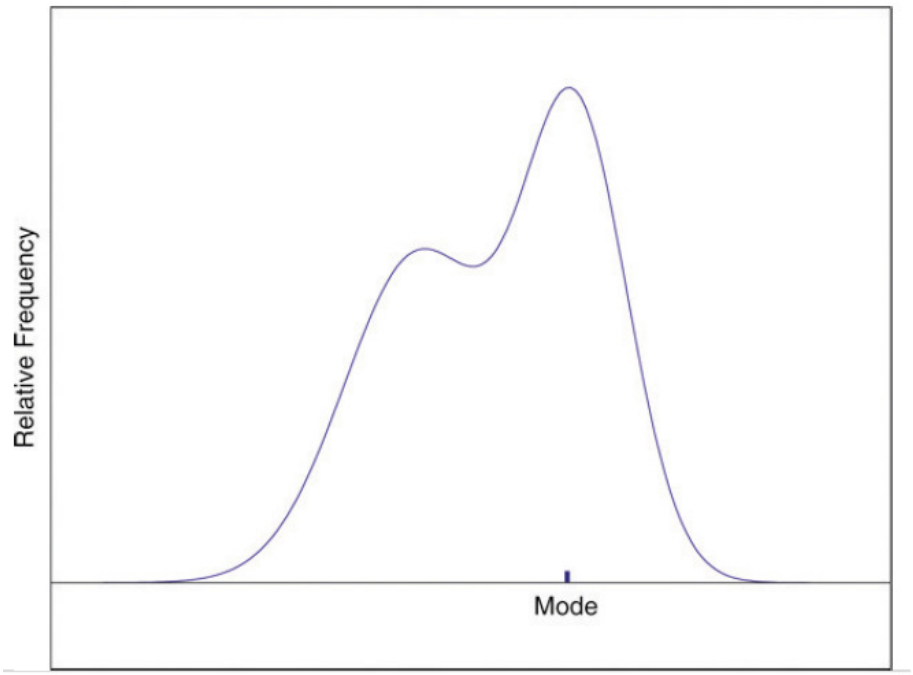
(a) $\bar{x} = \tilde{x}$ (b) $\bar{x} = \tilde{x}$

(c) $\bar{x} > \tilde{x}$ (d) $\bar{x} < \tilde{x}$

3) Mode

The sample mode of a set of sample data is the most frequently occurring value. The population mode is defined in a similar way.

On a relative frequency histogram, the highest point of the histogram corresponds to the mode of the data set. Figure below shows the Mode.



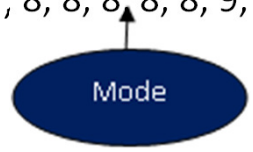
Case (8) The following data represent the rainy days for the last five years over Iraq. Find the mode?

8, 9, 9, 14, 8, 8, 10, 7, 6, 9, 7, 8, 10, 14, 11, 8, 14, 11

Solution

Arrange the data in order

6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 10, 10, 11, 11, 14, 14, 14



The mode equals to 8

For any data set there is always exactly one mean and exactly one median.

This need not be true of the mode; several different values could occur with the highest frequency. It could even happen that every value occurs with the same frequency, in which case the concept of the mode does not make much sense.

Case (9) Find the mode of evaporation data set over Basra along 12 months.

110, 713, 451, 740, 200, 118, 622, 977, 103, 752, 111, 634

Solution

Arrange the data in order

103, 110, 111, 118, 200, 451, 622, 634, 713, 740, 752, 977 ….. Since each value occurs only once, there is no mode.

Case (10) Find the mode for the following data: -1 0 2 0

Solution

Arrange the data in order …….. -1, 0, 0, 2 …. The mode is 0

Case (11) Find the mode for the following data:

X	0	1	2	3	4
f	3	6	6	3	1

Solution

The two most frequently observed values in the data set are 1 and 2.Therefore mode is a set of two values: {1,2}.

- ✓ The mode is a measure of central location since most real-life data sets have more observations near the center of the data range and fewer observations on the lower and upper ends. The value with the highest frequency is often in the middle of the data range.
- ✓ The mean, the median, and the mode each answer the question "Where is the center of the data set?" The nature of the data set, as indicated by a relative frequency histogram, determines which one gives the best answer.

The mode for grouped data can be shown by the following case. It is called Modal Class.

Case (12) Find the modal class for the following frequency distribution:

Class boundaries	Frequency
5.5 – 10.5	1
10.5 – 15.5	2
15.5 – 20.5	3
20.5 – 25.5	5
25.5 – 30.5	4
30.5 – 35.5	3
35.5 – 40.5	2

Modal Class

More frequently

4) **The midrange**

The midrange is a rough estimate of the middle. It is found by adding the lowest and highest values in the data set and dividing by 2. It is very rough estimate of the average and can be affected by one extremely high or low value.

$$MR = \frac{lowest\ value + highest\ value}{2}$$

EXERCISES

BASIC

1. For the sample data set {1,2,6} find
   a. Σx
   b. Σx^2
   c. $\Sigma (x-3)$
   d. $\Sigma (x-3)^2$
2. For the sample data set $\{-1,0,1,4\}$ find
   a. Σx
   b. Σx^2
   c. $\Sigma (x-1)$
   d. $\Sigma (x-1)^2$
3. Find the mean, the median, and the mode for the sample

   1 2 3 4
4. Find the mean, the median, and the mode for the sample

   3 3 4 4
5. Find the mean, the median, and the mode for the sample

   2 1 2 7
6. Find the mean, the median, and the mode for the sample

   −1 0 1 4 1 1
7. Find the mean, the median, and the mode for the sample data represented by the table

x	1	2	7
f	1	2	1

8. Find the mean, the median, and the mode for the sample data represented by the table

x	−1	0	1	4
f	1	1	3	1

9. Create a sample data set of size $n = 3$ for which the mean $\overline{x}$ is greater than the median $\widetilde{x}$.
10. Create a sample data set of size $n = 3$ for which the mean $\overline{x}$ is less than the median $\widetilde{x}$.
11. Create a sample data set of size $n = 4$ for which the mean $\overline{x}$, the median $\widetilde{x}$, and the mode are all identical.
12. Create a data set of size $n = 4$ for which the median $\widetilde{x}$ and the mode are identical but the mean $\overline{x}$ is different.

APPLICATIONS

13. Find the mean and the median for the LDL cholesterol level in a sample of ten heart patients.

    132 162 133 145 148

    139 147 160 150 153
14. Find the mean and the median, for the LDL cholesterol level in a sample of ten heart patients on a special diet.

    127 152 138 110 152

    113 131 148 135 158
15. Find the mean, the median, and the mode for the number of vehicles owned in a survey of 52 households.

x	0	1	2	3	4	5	6	7
f	2	12	15	11	6	3	1	2

16. The number of passengers in each of 120 randomly observed vehicles during morning rush hour was recorded, with the following results.

x	1	2	3	4	5
f	84	29	3	3	1

    Find the mean, the median, and the mode of this data set.
17. Twenty-five 1-lb boxes of 16d nails were randomly selected and the number of nails in each box was counted, with the following results.

x	47	48	49	50	51
f	1	3	18	2	1

    Find the mean, the median, and the mode of this data set.

ADDITIONAL EXERCISES

18. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 500 days, four mice have died but the fifth one survives. The recorded survival times for the five mice are

493 421 222 378 500*

where 500* indicates that the fifth mouse survived for at least 500 days but the survival time (i.e., the exact value of the observation) is unknown.

a. Can you find the sample mean for the data set? If so, find it. If not, why not?
b. Can you find the sample median for the data set? If so, find it. If not, why not?

19. Five laboratory mice with thymus leukemia are observed for a predetermined period of 500 days. After 450 days, three mice have died, and one of the remaining mice is sacrificed for analysis. By the end of the observational period, the last remaining mouse still survives. The recorded survival times for the five mice are

222 421 378 450* 500*

where * indicates that the mouse survived for at least the given number of days but the exact value of the observation is unknown.

a. Can you find the sample mean for the data set? If so, find it. If not, explain why not.
b. Can you find the sample median for the data set? If so, find it. If not, explain why not.

20. A player keeps track of all the rolls of a pair of dice when playing a board game and obtains the following data.

x	2	3	4	5	6	7
f	10	29	40	56	68	77

x	8	9	10	11	12
f	67	55	39	28	11

Find the mean, the median, and the mode.

21. Cordelia records her daily commute time to work each day, to the nearest minute, for two months, and obtains the following data.

x	26	27	28	29	30	31	32
f	3	4	16	12	6	2	1

a. Based on the frequencies, do you expect the mean and the median to be about the same or markedly different, and why?
b. Compute the mean, the median, and the mode.

22. An ordered stem and leaf diagram gives the scores of 71 students on an exam.

10	0 0
9	1 1 1 1 2 3
8	0 1 1 2 2 3 4 5 7 8 8 9
7	0 0 0 1 1 2 4 4 5 6 6 6 7 7 7 8 8 9
6	0 1 2 2 2 3 4 4 5 7 7 7 7 8 8
5	0 2 3 3 4 4 6 7 7 8 9
4	2 5 6 8 8
3	9 9

a. Based on the shape of the display, do you expect the mean and the median to be about the same or markedly different, and why?
b. Compute the mean, the median, and the mode.

23. A man tosses a coin repeatedly until it lands heads and records the number of tosses required. (For example, if it lands heads on the first toss he records a 1; if it lands tails on the first two tosses and heads on the third he records a 3.) The data are shown.

x	1	2	3	4	5	6	7	8	9	10
f	384	208	98	56	28	12	8	2	3	1

a. Find the mean of the data.
b. Find the median of the data.

24. a. Construct a data set consisting of ten numbers, all but one of which is above average, where the average is the mean.
b. Is it possible to construct a data set as in part (a) when the average is the median? Explain.

25. Show that no matter what kind of average is used (mean, median, or mode) it is impossible for all members of a data set to be above average.

26.
a. Twenty sacks of grain weigh a total of 1,003 lb. What is the mean weight per sack?
b. Can the median weight per sack be calculated based on the information given? If not, construct two data sets with the same total but different medians.

27. Begin with the following set of data, call it Data Set I.

5 −2 6 14 −3 0 1 4 3 2 5

a. Compute the mean, median, and mode.
b. Form a new data set, Data Set II, by adding 3 to each number in Data Set I. Calculate the mean, median, and mode of Data Set II.
c. Form a new data set, Data Set III, by subtracting 6 from each number in Data Set I. Calculate the mean, median, and mode of Data Set III.
d. Comparing the answers to parts (a), (b), and (c), can you guess the pattern? State the general principle that you expect to be true.

ANSWERS

1.
a. 9.
b. 41.
c. 0.
d. 14.

3. $\overline{x} = 2.5, \widetilde{x} = 2.5, \text{mode} = \{1,2,3,4\}$.

5. $\overline{x} = 3, \widetilde{x} = 2, \text{mode} = 2$.

7. $\overline{x} = 3, \widetilde{x} = 2, \text{mode} = 2$.

9. {0,0,3}.

11. {0,1,1,2}.

13. $\overline{x} = 146.9, \widetilde{x} = 147.5$

15. $\overline{x} = 2.6, \widetilde{x} = 2, \text{mode} = 2$

17. $\overline{x} = 48.96, \widetilde{x} = 49, \text{mode} = 49$

19.
a. No, the survival times of the fourth and fifth mice are unknown.
b. Yes, $\widetilde{x} = 421$.

21. $\overline{x} = 28.55, \widetilde{x} = 28, \text{mode} = 28$

23. $\overline{x} = 2.05, \widetilde{x} = 2, \text{mode} = 1$

25. Mean: $nx_{\min} \le \Sigma x$ so dividing by n yields $x_{\min} \le \overline{x}$, so the minimum value is not above average. Median: the middle measurement, or average of the two middle measurements, $\widetilde{x}$, is at least as large as $x_{\min}$, so the minimum value is not above average. Mode: the mode is one of the measurements, and is not greater than itself.

27.
a. $\overline{x} = 3.\overline{18}, \widetilde{x} = 3, \text{mode} = 5$.
b. $\overline{x} = 6.\overline{18}, \widetilde{x} = 6, \text{mode} = 8$.
c. $\overline{x} = -2.\overline{81}, \widetilde{x} = -3, \text{mode} = -1$.
d. If a number is added to every measurement in a data set, then the mean, median, and mode all change by that number.