Measures of Central Tendency

MHR Page 263 #s 1 - 5, 7, 8, 10 & 13

Defintions

Mean

- The sum of the data entries divided by the number of entries

Median

- The middle value of all the data points when the data values are listed in numerical order
- If there is an even number of data points, then the median is the average between the two middle values

Mode

- The data value that occurs most often in the list of data points
- It is possible to have no mode, one mode, or more than one mode

In statistics, you can find the mean of a population and the mean of a sample of that population. A sample mean will approximate the actual mean of the population.

Population Mean

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

Sample Mean

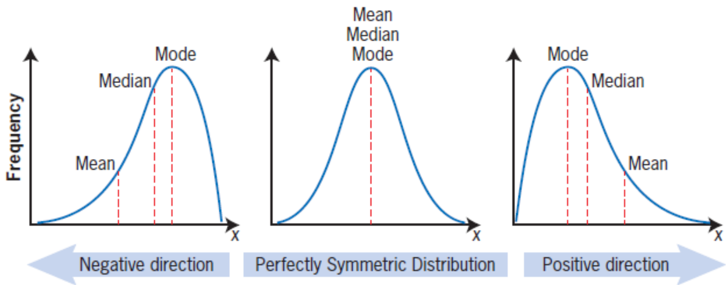
$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

where N is the size of the population and n is the sample size.

Although the calculations are the same, different symbols are used to indicate whether it represents a population or a sample.

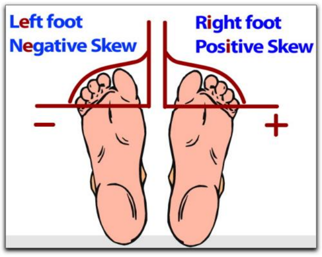
The measures of central tendency of a data set can be affected by the presence of outliers.

- In a symmetric distribution such as the uniform distribution, the mean, median, and mode will all be equal.
- In a non-symmetric or skewed distribution, the mean, median, and mode will differ.
- In a positively skewed distribution, the mode will be the lowest of the three values and the mean will be the highest.
- In a negatively skewed distribution, the mode will be the highest of the three values and the mean will be the lowest.



μ = "mu"

$\overline{x}$ = "x bar"



Imagine being on a slide and you are going to slide towards those numbers

Example 1

Evaluating Measures of Central Tendency

You are interviewing for an internship at a risk assessment firm to gain experience for your post-secondary program. The interviewer tells you that the average annual income of the 15 employees at the company is $73 518.27. The chart shows the actual incomes of the 15 employees.

$34 983	$18 980	$12 500	$48 980	$478 320
$17 305	$36 540	$12 500	$250 921	$32 654
$45 678	$33 855	$25 676	$33 450	$20 432

a) Determine the mean, median, and mode of the incomes.

b) Use the measures of central tendency to decide whether the interviewer's statement is accurate.

c) What is the effect of the outliers on the measures of central tendency?

d) Which measure of central tendency best represents the "average" income of the employees?

12,500 12,500 17,305
18,980 20,432 25,676
32,654 33,450 33,855
34,983 36,540 45,678
48,980 250,921 478,320

Mode is $12,500

a) Mean $\mu = \frac{\Sigma x}{N}$

$\mu = \frac{1,102,774}{15}$

$\mu \approx 73518.27$

The mean wage is about $73,518.27

Median position $= \frac{n+1}{2}$

$= \frac{15+1}{2} = 8^{th} \Rightarrow \$33,450$

b) Whilst the interviewer is correct in their statement, the "average" (mean) is affected by two outliers. The mode is lower than the mean (much!) so the median would be the most appropriate value to use.

c) The mean has been inflated by the outliers. If we recalculate it without them, it is then about $28733.31

d) Again, as the mean is greatly affected by the two outliers, the median is the best measure of central tendency to use.

Your Turn

Before heading on vacation to Mexico, you observe the actual high temperatures for seven days. The table shows the temperatures.

Day	Temperature (°C)
1	27
2	29
3	32
4	29
5	45
6	29
7	31

a) Determine the mean, median, and mode of the temperatures.

b) The weather report predicts that based on the previous seven-day forecast, the temperature on the day of your arrival should be 36 °C. Use the measures of central tendency in part a) to determine whether the weather report is accurate.

c) Is there an outlier in the data? How does it affect the measures of central tendency?

d) Which measure of central tendency would best represent the temperatures in this Mexican location? Explain.

27, 29, 29, 29, 31, 32, 45

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

$= \frac{222}{7}$

$\approx 31.7°C$

Median position $= \frac{n+1}{2} = \frac{7+1}{2} = 4$

4th value $\Rightarrow$ 29°C

Modal value $\Rightarrow$ 29°C

b) None of the measures of central tendency support this prediction.

c) The value 45°C is an outlier. This will result in an inflated mean. Recalculating without the outlier gives a mean of $177 \div 6 = 29.5°C$

d) As the outlier greatly affects the mean, the median or mode is the most appropriate measure of central tendency.

When the quantity of data is large, you can group the data into intervals to make them easier to analyse. When data are grouped into intervals, you can only approximate the centre of the data. To do this, assume that the data are evenly spaced in each interval, and use the midpoint to represent the values in each interval. Multiply the data values by their respective frequencies. Then, add these products and divide by the total frequency. You can use the following formula to approximate the mean for grouped data.

Mean for Grouped Data

$$\overline{x} = \frac{\sum f_i m_i}{\sum f_i}$$

where m_i is the midpoint of each interval and f_i is the frequency of each interval.

You can use a frequency distribution table to help organize your data.

Example 2

Using Grouped Data

The time taken to complete a chess game was recorded, to the nearest minute. The frequency table shows the data.

Time (min)	10–15	15–20	20–25	25–30	30–35
Frequency	2	20	18	10	5

a) Calculate the estimated mean, median, and mode times, in minutes, to complete a chess game.

b) Describe potential issues with finding the measures of central tendency of grouped data.

c) Graph the data using a histogram. Mark the measures of central tendency on the graph.

d) Discuss any skewing of the data with respect to the measures of central tendency.

Number of Minutes	Midpoint, m_i	Number of Games, f_i	$m_i f_i$	Cumulative Frequency
10–15	12.5	2	25	2
15–20	17.5	20	350	22
20–25	22.5	18	405	40
25–30	27.5	10	275	50
30–35	32.5	5	162.5	55

Frequency

Midpoint x Freq

Running total of frequencies

a) $\bar{x} = \dfrac{\sum m_i f_i}{\sum f_i}$

$= 1217.5 \div 55$

≈ 22.14

Mean is about 22 minutes

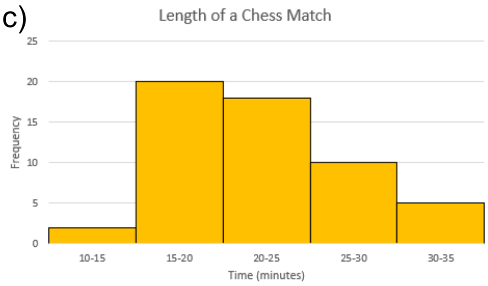
Median position = (55+1)÷2 = 28

28th value is in the 20-25 interval, ≈ 22.5 minutes

Modal interval is 15-20 (highest freq)

b) Because the data is grouped, we don't know what the actual values are. The larger the class interval, the more likely your answers are to be inaccurate. Therefore the mean and median can only be estimated.

c)



d) Since the data is positively skewed the modal interval is the least appropriate of the measures of central tendency. The median and mean are close in value so either could be used.

Your Turn

A group of children were asked how many hours a day they spend playing video games. The table shows the data.

a) Determine the estimated mean, median number of hours, and modal interval for the above distribution.

b) Discuss any skewing of the data with respect to the measures of central tendency.

Number of Hours	Midpoint, m_i	Number of Children, f_i	$m_i f_i$	Cumulative Frequency
0–2	1	3	3	3
2–4	3	11	33	14
4–6	5	7	35	21
6–8	7	2	14	23
8–10	9	1	9	24

a) $\bar{x} = \dfrac{\sum m_i f_i}{\sum f_i}$

$= 94 \div 24$

≈ 3.92

Mean is about 4 hours

Median position is (24+1)÷2 = 12.5

12th and 13th values are in interval 2-4 hours, ≈ 3 hours

Modal interval is 2-4 hours (highest freq)

b) Again, as the data is positively skewed, the modal interval is the least appropriate. The mean is the largest of the three measures of central tendency, so we should use the median for this example.

Certain values in a data set are sometimes of greater relative importance than others. In these cases, it is useful to calculate a weighted mean. To do this, multiply the weighting by the corresponding data value, find the sum of these products, and then divide by the total weighting.

Weighted Mean

$$\mu = \frac{\sum x_i w_i}{\sum w_i}$$

Where x_i represents each data value in the data set and w_i represents its weight or frequency.

Example 3

Using a Weighted Mean

A teacher is calculating the marks for the students in her Data Management class. She assigns the following values to each category:

Knowledge: 25%	Thinking: 10%
Application: 20%	Culminating Project: 15%
Communication: 15%	Final Exam: 15%

Kyle has not yet written his final exam, but his marks in the first five categories are 90, 79, 82, 70, and 85.

a) Determine the weighted mean for Kyle before the final exam.

b) How does this weighted mean differ from the unweighted mean?

c) What mark must Kyle receive on the final exam to finish the course with 84%?

a) $\mu_w = \frac{90(0.25) + 79(0.2) + 82(0.15) + 70(0.1) + 85(0.15)}{0.85}$ ← Hasn't done the exam

$= \frac{70.35}{0.85}$

≈ 82.76

b) $\mu = \frac{90 + 79 + 82 + 70 + 85}{5}$

$\mu = \frac{406}{5}$

$\mu = 81.2$

Slightly lower due to Kyle's highest mark carrying more weight

c) $\mu_w = \frac{\text{Total of weights}}{1}$

$84 = 70.35 + E(0.15)$

(-70.35) $\frac{13.65}{0.15} = \frac{0.15E}{0.15}$

$91 = E$

Kyle would a 91 on the exam to finish with a mark of 84.

Your Turn

A math department assigns the following weights for each category in its Advanced Functions course:

Knowledge: 25%
Application: 15%
Communication: 10%
Thinking: 10%
Culminating Project: 10%
Final Exam: 30%

Catherine's marks in the course so far are 87, 90, 76, 78, and 84 in each of the first five categories. She still needs to write the final exam.

a) Determine the weighted mean for Catherine before writing her final exam.

b) Is it possible for Catherine to receive a final mark of 90% in the course? Justify your answer.

a) $\mu_w = \dfrac{87(0.25) + 90(0.15) + 76(0.1) + 78(0.1) + 84(0.1)}{0.7}$

$= \dfrac{59.05}{0.7}$

≈ 84.4

b) $90 = \dfrac{\text{Total weights}}{1}$

$90 = 59.05 + E(0.3)$

$(-59.05) \quad \dfrac{30.95}{0.3} = \dfrac{0.3E}{0.3}$

$103.1\overline{6} = E$

Catherine would need an exam mark of $103.1\overline{6}$ to get a final mark of 90, so NO, it is not possible.

Key Concepts

- Three measures of central tendency are mean, median, and mode.
- The mean represents the average of a set of data.
- The median is the middle number when the numbers are arranged in numerical order.
- The mode is the number that occurs most often; it is possible to have one, more than one, or no mode.
- Outliers have a greater effect on the mean than other measures and either pull the mean up or drag the mean down.
- A weighted mean accounts for the relative importance of each value in the average.
- Grouped data are organized into intervals. Use the interval midpoints and frequencies to estimate the measures of central tendency.

R1. Which measure of central tendency is most affected by extreme values? Explain using specific examples to justify your answer.

The most affected by extreme values is the mean. Since outliers are extreme values they will push the mean away from the centre of a distribution. A very tall person in a class will increase the mean height, where as a shorter person would reduce the mean.

R2. Describe a situation in which it would be necessary for you to use

a) the mean

b) the weighted mean

c) grouped data

a) The mean is what is known as the "average" of a data set. An example would be when calculating your term average for a high school credit.

b) The weighted mean is when certain elements have more relative importance than others. An example would be your culminating activities being worth 30% of your final mark.

c) Grouped data is used when dealing with large amounts of data. Putting it into class intervals makes it easier to analyse. An example would be when looking at results of surveys.

R3. Which measure of central tendency is being used in each situation? Explain.

a) The average person has two hands, two eyes, two ears, and two legs.

b) The average time it takes to get to school is 38 min.

c) Johnny is an above average student.

a) Mode: this is what people have in most cases.

b) Mean: the sum of the travel times divided by the number of trips taken.

c) Median: the class average on a report card is the median.