One-Variable Data Analysis Review

Learning Goals

Section	After this section, I can
6.1	• interpret the mean, median, and mode of a set of data • choose the measure of central tendency that best describes the data
6.2	• describe the variability in a sample or population using measures of spread • calculate the range • understand how to use quartiles and percentiles to analyse data
6.3	• use technology to calculate the variance and standard deviation of a data set • calculate and understand the significance of a *z*-score • relate the positive or negative scores to their locations in a histogram • develop significant conclusions about a data set
6.4	• interpret statistical summaries to describe a one-variable data set and to compare two related one-variable data sets • understand whether the data presented are valid and reliable • describe how statistical summaries can misrepresent one-variable data • make inferences and make and justify conclusions from statistical summaries of one-variable data • interpret statistics in the media, assess the validity of conclusions made, and explain how statistics are used to promote a certain point of view
6.5	• collect data through secondary sources • generate, using technology, the relevant graphical summaries of one-variable data • interpret statistical summaries • assess the validity of conclusions presented in the media • draw conclusions from the analysis of data and evaluate the strength of the evidence

MHR Page 308 #s 1 - 11

Solutions

1. a) Define the three measures of central tendency.
   b) Explain how each measure is determined.
   c) Provide a real-life example of where each measure is most appropriate.

a) Mean = the arithmetic average of the data.

Median = middle value once the data is ordered.

Mode = the value that occurs most often.

b) Mean = sum of data divided by the number of items of data.

Median = middle value once the data is ordered. If there are two middle values, take the mean of them.

Mode = identify value that occurs most often.

c) Mean = average time taken by an athlete to run 100 metres.

Median = used to represent an employees salary.

Mode = the most frequently sold shoe size.

2. Calculate the mean, median, and mode of the data sets. Express your answers to one decimal place.

a) 75 989 54 76 675 45 242 54
85 342 12 931 2 37 675

Rank	Data	Ordered
1	75	2
2	85	12
3	989	37
4	342	45
5	54	54
6	12	54
7	76	75
8	931	76
9	675	85
10	2	242
11	45	342
12	37	675
13	242	675
14	675	931
15	54	989

Mean $= \frac{\sum x}{n}$

$= \frac{75+989+54+76+675+45+242+54+85+342+12+931+2+37+675}{15}$

$= \frac{4294}{15}$

≈ 286.3

Median = 8th position once ordered

= 76

Mode = 54 and 675 (both appear twice)

2. Calculate the mean, median, and mode of the data sets. Express your answers to one decimal place.

b) 7 19 21 5 17 31 62 7 50 10 7 34

Rank	Data	Ordered
1	7	5
2	19	7
3	21	7
4	5	7
5	17	10
6	31	17
7	62	19
8	7	21
9	50	31
10	10	34
11	7	50
12	34	62

Mean $= \frac{\sum x}{n}$

$= \frac{7+19+21+5+17+31+62+7+50+10+7+34}{12}$

$= \frac{270}{12}$

$= 22.5$

Mode = 7 (appears thrice)

Median = Midway between 6th and 7th values once ordered

= (17 + 19) / 2

= 36 / 2

= 18

2. Calculate the mean, median, and mode of the data sets. Express your answers to one decimal place.

c) 1856 6754 2346 5200
6754 9564 2346 1880

Rank	Data	Ordered
1	1856	1856
2	6754	1880
3	6754	2346
4	9564	2346
5	2346	5200
6	2346	6754
7	5200	6754
8	1880	9564

Mean $= \frac{\sum x}{n}$

$= \frac{1856+6754+2346+5200+6754+9564+2346+1880}{8}$

$= \frac{36\,700}{8}$

$= 4587.5$

Median = Midway between 4th and 5th values once ordered

= (2346 + 5200) / 2

= 7546 / 2

= 3773

Mode = 2346 and 6754 (both appear twice)

3. A softball player's slugging average is calculated using the formula

$$SLG = \frac{S + 2D + 3T + 4H}{B},$$

where S is the number of singles, D is the number of doubles, T is the number of triples, H is the number of home runs, and B is the number of times batting. Calculate each baseball player's slugging average.

a) Jane, with 85 singles, 15 doubles, 1 triple, and 20 home runs, in 308 times at bat.

b) Tonya, with 56 singles, 25 doubles, 0 triples, and 38 home runs, in 294 times at bat.

c) Monique, with 112 singles, 10 doubles, 9 triples, and 6 home runs, in 315 times at bat.

a) For Jane, $S = 85$, $D = 15$, $T = 1$, $H = 20$, and $B = 308$.

$$SLG = \frac{S + 2D + 3T + 4H}{B}$$
$$= \frac{85 + 2(15) + 3(1) + 4(20)}{308}$$
$$= \frac{198}{308}$$
$$\approx 0.643$$

Jane's slugging average is 0.643

b) For Tonya, $S = 56$, $D = 25$, $T = 0$, $H = 38$, and $B = 294$.

$$SLG = \frac{S + 2D + 3T + 4H}{B}$$
$$= \frac{56 + 2(25) + 3(0) + 4(38)}{294}$$
$$= \frac{258}{294}$$
$$\approx 0.878$$

Tonya's slugging average is 0.878

c) For Monique, $S = 112$, $D = 10$, $T = 9$, $H = 6$, and $B = 315$.

$$SLG = \frac{S + 2D + 3T + 4H}{B}$$
$$= \frac{112 + 2(10) + 3(9) + 4(6)}{315}$$
$$= \frac{183}{315}$$
$$\approx 0.581$$

Monique's slugging average is 0.581

4. a) Determine the mean, median, and modal interval of the data set.

b) Graph the data with a histogram and mark the measures of central tendency on the graph.

Salary Range ($ thousands)	Number of Employees
30–40	10
40–50	18
50–60	31
60–70	14
70–80	5
80–90	0
90–100	2
100–110	5

a)

Salary Range $1000s	x_1	y_1	x_1y_1	Cumulative Frequency
30-40	35	10	350	10
40-50	45	18	810	28
50-60	55	31	1705	59
60-70	65	14	910	73
70-80	75	5	375	78
80-90	85	0	0	78
90-100	95	2	190	80
100-110	105	5	525	85

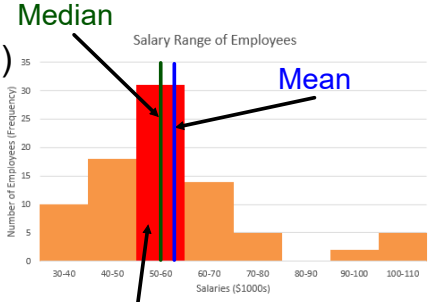
Mean

$$\text{Mean} = \frac{\text{total}(x_1y_1)}{\text{total}(y_1)}$$

= 57.2352941176

$$= \frac{4865}{85} = \$57{,}235$$

b)



Median = 43rd piece of data (85+1) / 2

= this occurs in the 50 - 60 interval

= $55,000 (use the midpoint of the interval)

Mode = Interval with the biggest frequency

= 50 - 60 (frequency of 31)

= $50,000 - $60,000

4. a) Determine the mean, median, and modal interval of the data set.

b) Graph the data with a histogram and mark the measures of central tendency on the graph.

Salary Range ($ thousands)	Number of Employees
30–40	10
40–50	18
50–60	31
60–70	14
70–80	5
80–90	0
90–100	2
100–110	5

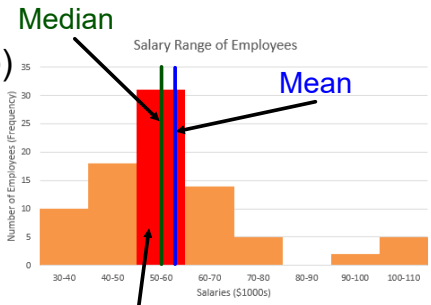
a)

Salary Range $1000s	Midpoint	Number of Employees	Midpoint x Frequency	Cumulative Frequency
30-40	35	10	350	10
40-50	45	18	810	28
50-60	55	31	1705	59
60-70	65	14	910	73
70-80	75	5	375	78
80-90	85	0	0	78
90-100	95	2	190	80
100-110	105	5	525	85
		Grouped Mean =		57.235294

Mean = Sum of Midpoint x Frequency / Cumulative Frequency Total

$= \frac{4865}{85} = \$57{,}235$

b)



Median = 43rd piece of data (85+1) / 2

= this occurs in the 50 - 60 interval

= $55,000 (use the midpoint of the interval)

Mode = Interval with the biggest frequency

= 50 - 60 (frequency of 31)

= $50,000 - $60,000

5. a) Describe what is meant by percentiles and quartiles.

b) Explain how quartiles would be useful for a store ordering shoe sizes.

a) A percentile is the percentage of the data that is equal or below a specific piece of data. Quartiles are when the data is divided into four equal parts. Q1 is the same as the 25th percentile, Q2 is the 50th percentile (median), and Q3 is equal to the 75th percentile.

b) A shoe store could use the interquartile range (IQR = Q3 - Q1). This shows the middle 50% of the shoe sizes sold, so the store would order more of these sizes as these would be the most likely to be bought.

6. The table provides the number of Facebook friends for a sample of 178 people aged 18 to 25.

a) Determine the percentiles for each of the Number of Friends intervals.

b) Determine the quartiles and the interquartile range.

c) Make a box and whisker plot.

d) Determine whether there are any outliers.

$$p = 100\frac{(L + 0.5E)}{n}$$

Number of Friends	Frequency	Cumulative Frequency	Percentile
0-25	3	3	0.8
25-50	18	21	6.7
50-75	16	37	16.3
75-100	35	72	30.6
100-125	62	134	57.9
125-150	23	157	81.7
150-175	14	171	92.1
175-200	0	171	96.1
200-225	5	176	97.5
225-250	2	178	99.4

a) You could use a spreadsheet to set this up. First add a column for cumulative frequency. Then add a percentile column. Alternatively, just crunch them out!

b) 178 data points.

Q1 = 0.25(178+1) = 44.75th value = 75-100 friends ⟶ 87.5 friends.

Q3 = 0.75(178+1) = 134.25th value = 100-125 friends ⟶ 112.5 friends.

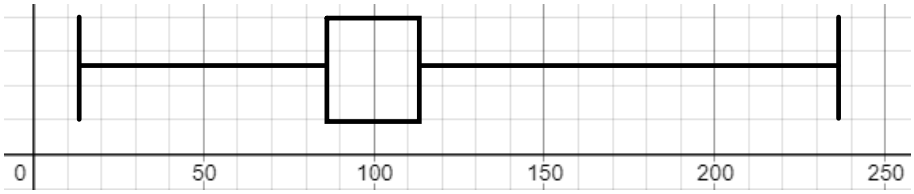
Interquartile Range = Q3 - Q1

= 112.5 - 87.5

= 25 friends

c) To create the box plot you need your **FIVE** values:

Lowest = 12.5, Q1 = 87.5, Median = 112.5, Q3 = 112.5, Highest = 237.5



d) An outlier is any value that 1.5 times the IQR below the Q1 or above the Q3.

Lower Limit:

= Q1 - 1.5(IQR)

= 87.5 - 1.5(25)

= 50

Upper Limit:

= Q3 + 1.5(IQR)

= 112.5 + 1.5(25)

= 150

There are 21 (3+18) below Q1 and 21 (14+0+5+2) above Q3 making a total of 42 outliers for this set of data.

7. The table provides the full-time enrollments of Ontario universities for 2012–2013.

a) Calculate the mean, variance, and standard deviation.

b) What is the z-score for York University?

c) Which universities have a z-score of –2 or less?

a) Mean = Total Enrolled / # of Universities

= 400,329 / 20

= 20016.45

All Ontario Universities are included so use the POPULATION formula.

Variance $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$

= 5053034377 / 20

= 252,651,719

Standard Deviation = √Variance

= √252,651,719

= 15,895.02

$a = [1427, 15678, 21988, 20692, 7046, 6635, 2$

a = 20 element list

Mean

$\text{mean}(a)$

= 20016.45

Sum of the deviations squared (x - mean)^2

$\text{total}\left((a - \text{mean}(a))^{(2)}\right)$

$= 5.053034377 \times 10^9$

Variance

$\frac{\text{total}\left((a - \text{mean}(a))^{(2)}\right)}{20}$

= 252651718.848

Standard deviation

$\sqrt{\frac{\text{total}\left((a - \text{mean}(a))^{(2)}\right)}{20}}$

= 15895.021826

7. The table provides the full-time enrollments of Ontario universities for 2012–2013.

a) Calculate the mean, variance, and standard deviation.

b) What is the z-score for York University?

c) Which universities have a z-score of –2 or less?

a) Mean = Total Enrolled / # of Universities

= 400,329 / 20

= 20016.45

All Ontario Universities are included so use the POPULATION formula.

Variance $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$

= 5053034377 / 20

= 252,651,719

Standard Deviation = √Variance

= √252,651,719

= 15,895.02

University	Enrollment (x)	Mean	x - Mean	(x - Mean)^2
1	1427	20016.45	-18589.5	345567651.3
2	15678	20016.45	-4338.45	18822148.4
3	21988	20016.45	1971.55	3887009.403
4	20692	20016.45	675.55	456367.8025
5	7046	20016.45	-12970.5	168232573.2
6	6635	20016.45	-13381.5	179063204.1
7	24798	20016.45	4781.55	22863220.4
8	3757	20016.45	-16259.5	264369714.3
9	3570	20016.45	-16446.5	270485717.6
10	8469	20016.45	-11547.5	133343601.5
11	33581	20016.45	13564.55	183997016.7
12	19901	20016.45	-115.45	13328.7025
13	22194	20016.45	2177.55	4741724.003
14	69081	20016.45	49064.55	2407330067
15	6760	20016.45	-13256.5	175733466.6
16	31611	20016.45	11594.55	134433589.7
17	29108	20016.45	9091.55	82656281.4
18	15984	20016.45	-4032.45	16260653
19	13557	20016.45	-6459.45	41724494.3
20	44492	20016.45	24475.55	599052547.8
Mean =	20016.45		Total =	5053034377

b) York University = 44,492

$$z = \frac{x - \mu}{\sigma}$$

= (44,492 - 20016.45) / 15895.02

= 1.5398

York University has a z-score of 1.5398

c) Use the formula to solve for x, then compare.

$$z = \frac{x - \mu}{\sigma}$$

-2 = (x - 20016.45) / 15895.02

-31790.04 = x - 20016.45

-11,773.59 = x

No universities have x= -11,773.59, so there are no universities with a z-score of -2.

Selected Universities	Total Full-Time Enrollment
Algoma University	1 427
Brock University	15 678
Carleton University	21 988
University of Guelph	20 692
Lakehead University	7 046
Laurentian University	6 635
McMaster University	24 798
Nipissing University	3 757
OCAD University	3 570
University of Ontario Institute of Technology	8 469
University of Ottawa	33 581
Queen's University	19 901
Ryerson University	22 194
University of Toronto	69 081
Trent University	6 760
University of Waterloo	31 611
University of Western Ontario	29 108
Wilfrid Laurier University	15 984
University of Windsor	13 557
York University	44 492

8. What does it mean to have a z-score of 1.5?

This means that the piece of data is 1.5 standard deviations above the mean of the data set.

9. The quality control department of Cool Cola tested the bottle fillers and found them to fill 500 mL bottles to a mean of 501.1 mL, with a standard deviation of 0.48 mL. The company's standard is set at test results being within two standard deviations of the mean.

a) What is the acceptable range of fills?

b) Why would the company want to overfill the bottles?

c) Three bottles of Cool Cola were tested for fill volume. Which are acceptable?

i) 501.0 mL ii) 502.1 mL

iii) 500 mL

a) Lower limit = Mean - 2(SD)

= 501.1 - 2(0.48)

= 501.1 - 0.96

= 500.14 mL

Upper limit = Mean + 2(SD)

= 501.1 + 2(0.48)

= 501.1 + 0.96

= 502.06 mL

The acceptable range is from 500.14 mL to 502.06 mL

b) They may want to overfill to account for the CO_2 that is in the drink.

c) (i) 501.0 mL is within the range

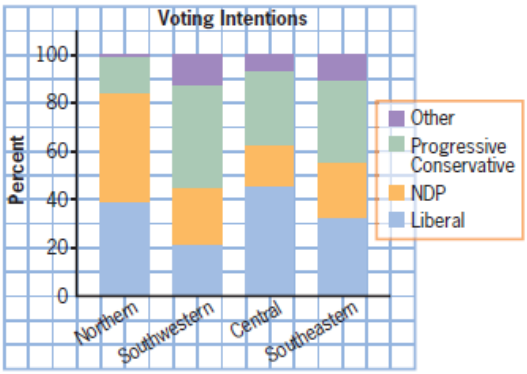
(ii) 502.1 mL is **NOT** within the range (too high)

(iii) 500 mL is **NOT** within the range (too low)

10. The graph shows the voting intentions in four regions of Ontario, taken from a poll of 2000 voters three days before an election.

a) Identify three pieces of information that you can read from the graph.

b) Would you consider the graph to be a valid predictor of the outcome of the election? Explain.



a) You could have:

(1) Voting Intentions,

(2) Voting info for the given region,

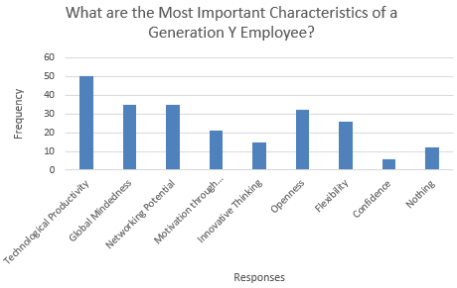
(3) Percentage of sample in a region that voted for that party

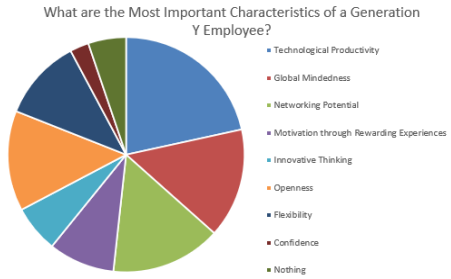
b) This is a very small sample to represent the whole of Ontario, so no, it is not likely to be a good predictor of the outcome of the election.

11. The following information was collected by a large marketing firm interested in attracting individuals from Generation Y to apply for a position at the company. The firm surveyed its managers to develop a list of qualities and benefits that Generation Y employees bring to the workplace. The table shows the number of responses to the question, "What are the most important characteristics of a Generation Y employee?" Visually represent the data and provide a conclusion based on the inferences that can be made.

Could have either a bar chart or pie chart

Characteristic	Frequency
Technological productivity	50
Global mindedness	35
Networking potential	35
Motivation through rewarding experiences	21
Innovative thinking	15
Openness	32
Flexibility	26
Confidence	6
Nothing	12





Based upon the responses, we can infer that the managers are looking for technologically productive employees.

Generation Y are known for wanting to save the planet (global mindedness), working collaboratively (networking) and speaking what's on their mind (openness). These are not as important to some managers.