One-Variable Data Analysis

Extra Practice

MHR Page 311 #s 1 - 8, 10 & 11

Solutions

1. The graph shows histograms of men's and women's heights in centimetres on the same set of axes.

If the data sets were combined, the distribution of heights would have

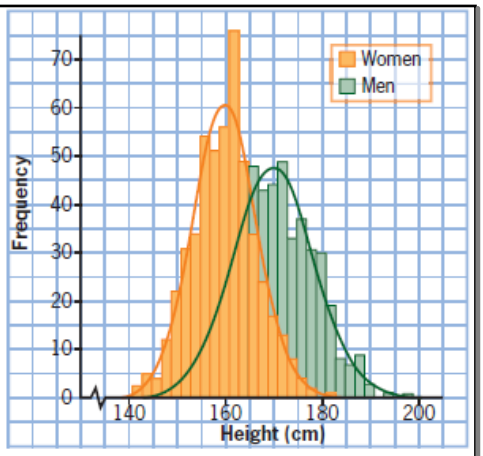
A no measures of central tendency

B two modes

C only one set of measures of central tendency

D none of the above

C



Once the data sets are combined, you then have one "new" data set. It will have only one set of central tendencies.

2. Final marks in Maria's Data Management course are based on 70% for term work, 15% for the exam, and 15% for the final course project. What term mark did Maria receive if her final mark was 87 and she received 84 on the exam and 95 on her final project?

A 87% B 83%

C 85% D 86%

D

70% of Term Work + 15% of Exam + 15% of Project = Final Mark

$0.7(x) + 0.15(84) + 0.15(95) = 87$

$0.7x + 12.6 + 14.25 = 87$

$0.7x + 26.85 = 87$

$0.7x = 60.15$

$x = 85.928...$

3. Find Q3 for the following masses of students in kilograms:

70 74 78 80 81 84 90 92 94

A 90
B 87
C 91
D 92

C

The 9 data points are already ordered.

Q3 is the 75th percentile. Using the percentile rank formula $R = \frac{p}{100}(n+1)$

R = 0.75(9 + 1)

R = 7.5

We need the midpoint of the 7th and 8th values.

Q3 = (90 + 92) / 2 = 91

4. What measure of central tendency is most appropriate to announce the most used bridge, on a daily basis, in Canada?

A mean
B mode
C median
D weighted mean

B

To find the most used bridge we need the MODE which measures the bridge with the highest total usage.

5. A set of nine different masses of pet cats are arranged in numerical order. The fifth mass is then increased by one. Which measure of spread for the data set could this change?

A the range
B the standard deviation
C the interquartile range
D all of the above

B

A - Range is highest minus lowest, so this is unaffected.

C - IQR is Q3 minus Q1, again this is unaffected.

6. If you are given the data listed below and are asked to use the interquartile range, could you successfully determine which baseball player's home run season totals are more consistent? Explain why or why not.

Ron: 20 21 23 25 18 19
Joshua: 20 20 23 24 19 22

Yes. The IQR is the middle 50% of a data set and therefore has a smaller range than the full set of data. The IQR will ignore very good, as well as very bad, seasons so that you get a more consistent set to compare.

Find Q3 and Q1 for both players using $R = \frac{p}{100}(n+1)$

Q3 = 0.75(6 + 1)
Q3 = 5.25

Q1 = 0.25(6 + 1)
Q1 = 1.75

Order the data

Ron: 18, 19, 20, 21, 23, 25 Q3 = 23.5, Q1 = 18.75, IQR = 4.75

Joshua: 19, 20, 20, 22, 23, 24 Q3 = 23.25, Q1 = 19.75, IQR = 3.5

Joshua has a smaller IQR, so he is the more consistent player.

7. Explain why sampling bias is not a major concern for the national census conducted by Statistics Canada.

Sampling bias is not a concern for the national census conducted by Statistics Canada because every household has to complete it by law. Therefore, they know that the information gathered covers the whole population.

8. The mean daily temperature during January was –12.1 °C, with a standard deviation of 5.6 °C. Use z-scores to indicate which of the following daily mean temperatures is closest to the monthly mean.

a) –17.4 °C
b) –3.6 °C
c) 0 °C
d) –6.4 °C

a) $z = \frac{x - \bar{x}}{s}$

z = (-17.4 - (-12.1)) / 5.6

z = -0.9464...

b) $z = \frac{x - \bar{x}}{s}$

z = (-3.6 - (-12.1)) / 5.6

z = 1.5178...

c) $z = \frac{x - \bar{x}}{s}$

z = (0 - (-12.1)) / 5.6

z = 2.1607...

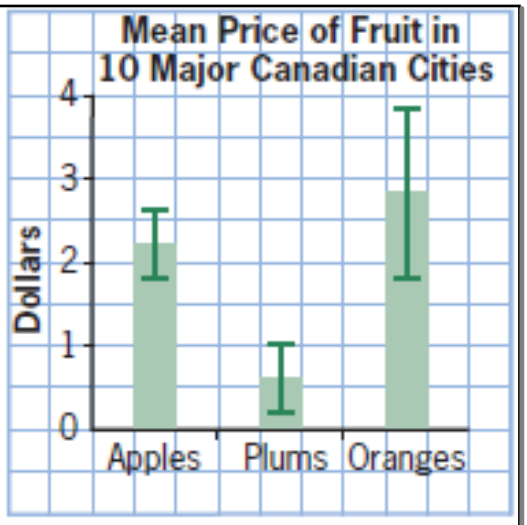
d) $z = \frac{x - \bar{x}}{s}$

z = (-6.4 - (-12.1)) / 5.6

z = 1.0178...

The temperature of -17.4°C is closest to the mean, because it has the closest z-score to zero.

10. The graph illustrates price fluctuations for three types of fruit. Each bar shows the mean price, with plus and minus one standard deviation superimposed. State the mean and interpret the standard deviation for each type of fruit.



Apples:

Mean = $2.25

SD = (2.60 - 1.80) / 2 = $0.40

Plums:

Mean = $0.60

SD = (1.00 - 0.20) / 2 = $0.40

Oranges:

Mean = $2.80

SD = (3.80 - 1.80) / 2 = $1.00

Mean = Height of column

SD = The line shows ± 1 SD. Subtract the two values and halve it to find the SD.

11. For a data management project, Ryan sent a survey to the teachers in his school, asking them how many years they have taught. Thirty teachers responded. Here are their responses:

3, 12, 2, 2, 18, 27, 19, 0, 14, 15, 3, 17, 12, 37, 25, 17, 22, 1, 5, 5, 18, 13, 18, 6, 1, 10, 10, 4, 9, 28

a) Calculate the mean, interquartile range, and standard deviation.

b) Organize the data into a frequency distribution with five intervals.

c) Estimate the mean, interquartile range, and standard deviation using the frequency distribution in part b). How do they compare to the true values?

d) Illustrate all the calculations on appropriate graphs.

e) What percentile rank is associated with 10 years of teaching?

f) How many years of teaching are represented by the 90th percentile?

g) Determine whether there are outliers. Identify any that are present.

h) Analyse the validity of Ryan's sampling method.

a)

Datum	Years Taught	x - mean	(x - mean)^2
1	0	-12.4333	154.5878
2	1	-11.4333	130.7211
3	1	-11.4333	130.7211
4	2	-10.4333	108.8544
5	2	-10.4333	108.8544
6	3	-9.4333	88.9878
7	3	-9.4333	88.9878
8	4	-8.4333	71.1211
9	5	-7.4333	55.2544
10	5	-7.4333	55.2544
11	6	-6.4333	41.3878
12	9	-3.4333	11.7878
13	10	-2.4333	5.9211
14	10	-2.4333	5.9211
15	12	-0.4333	0.1878
16	12	-0.4333	0.1878
17	13	0.5667	0.3211
18	14	1.5667	2.4544
19	15	2.5667	6.5878
20	17	4.5667	20.8544
21	17	4.5667	20.8544
22	18	5.5667	30.9878
23	18	5.5667	30.9878
24	18	5.5667	30.9878
25	19	6.5667	43.1211
26	22	9.5667	91.5211
27	25	12.5667	157.9211
28	27	14.5667	212.1878
29	28	15.5667	242.3211
30	37	24.5667	603.5211
Total	373	Total	2553.3667

The mean is 12.4333

The IQR is 14

The SD is 9.3833 (sample, not population)

Mean	12.4333
IQR	14
SD	9.3833

b)

Years Taught	Frequency	Midpoint	F x M	F x (x - mean)^2
0 - 8	11	4	44	958.2222
8 - 16	8	12	96	14.2222
16 - 24	7	20	140	311.1111
24 - 32	3	28	84	645.3333
32 - 40	1	36	36	513.7778
Total =	30	Total =	400	
Mean =	13.3333		SD =	9.1777

c) The mean is 13.3333

The IQR is 16 (24th value in 16-24, 8th value in 0-8, using midpoints IQR = 20 - 4)

The SD is 9.1777 (sample, not popultion)

The mean and IQR have increased slightly and the SD has decreased slightly when comparing the grouped data to the raw data.

11. For a data management project, Ryan sent a survey to the teachers in his school, asking them how many years they have taught. Thirty teachers responded. Here are their responses:

3, 12, 2, 2, 18, 27, 19, 0, 14, 15, 3, 17, 12, 37, 25, 17, 22, 1, 5, 5, 18, 13, 18, 6, 1, 10, 10, 4, 9, 28

a) Calculate the mean, interquartile range, and standard deviation.

b) Organize the data into a frequency distribution with five intervals.

c) Estimate the mean, interquartile range, and standard deviation using the frequency distribution in part b). How do they compare to the true values?

d) Illustrate all the calculations on appropriate graphs.

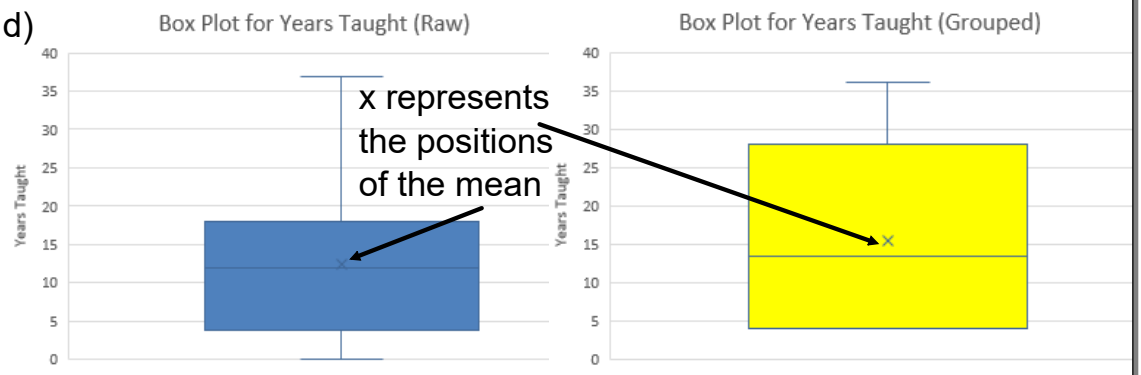
e) What percentile rank is associated with 10 years of teaching?

f) How many years of teaching are represented by the 90th percentile?

g) Determine whether there are outliers. Identify any that are present.

h) Analyse the validity of Ryan's sampling method.

d)



e) Using the percentile formula $p = 100\dfrac{(L + 0.5E)}{n}$

L = 12 (12 values < 10), E = 2 (2 values = 10), n = 30 (30 values total)

p = 100(12 + 0.5(2)) / 30

p = 43.333...

10 years of teaching experience is the 43rd percentile.

11. For a data management project, Ryan sent a survey to the teachers in his school, asking them how many years they have taught. Thirty teachers responded. Here are their responses:

3, 12, 2, 2, 18, 27, 19, 0, 14, 15, 3, 17, 12, 37, 25, 17, 22, 1, 5, 5, 18, 13, 18, 6, 1, 10, 10, 4, 9, 28

a) Calculate the mean, interquartile range, and standard deviation.

b) Organize the data into a frequency distribution with five intervals.

c) Estimate the mean, interquartile range, and standard deviation using the frequency distribution in part b). How do they compare to the true values?

d) Illustrate all the calculations on appropriate graphs.

e) What percentile rank is associated with 10 years of teaching?

f) How many years of teaching are represented by the 90th percentile?

g) Determine whether there are outliers. Identify any that are present.

h) Analyse the validity of Ryan's sampling method.

f) Using the percentile rank formula $R = \dfrac{p}{100}(n + 1)$

R = 0.90(30 + 1)

R = 27.9

The 90th percentile is 26 years of teaching.

Round down to 27. Find midpoint between 27th and 28th values.

90th percentile = (25 + 27) / 2 = 26.

g) Outliers are values outside of the quartiles by more than 1.5(IQR)

Lower limit = Q1 - 1.5(IQR)

= 4 - 1.5(14)

= -17

Upper limit = Q3 + 1.5(IQR)

= 18 + 1.5(14)

= 39

There are no data points less than -17 or greater than 39, so there are no outliers in this data set.

h) There are some issues with Ryan's sampling method. As we don't know how many teachers are in his school, we don't know if the sample is representative of the whole population. Since the survey was sent to the teachers, they got to choose whether to respond or not. So this is a voluntary response survey.