Review

One-Variable Data Analysis

Page 315 #s 9 - 13

- 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.

- A measure of spread helps you understand how closely a set of data is clustered around its centre.
- The range is the difference between the maximum value and minimum value.
- A percentile is the percent of all the data that are less than or equal to the specific data point.
- Quartiles divide the data set into four equal parts. Q1 is the 25th percentile, Q2 is the median (or 50th) percentile, and Q3 is the 75th percentile.
- The interquartile range (IQR) is the distance between the first and third quartiles. To calculate, subtract the value for Q1 from the value for Q3. The interquartile range contains the middle 50% of the data.
- A box and whisker plot uses a rectangle to visually demonstrate the spread of the distribution along a number line by displaying the median, quartiles, and upper and lower extremes.
- An outlier exists if it is less than Q1 – 1.5 × IQR or greater than Q3 + 1.5 × IQR.

- The variance and standard deviation are measures of spread. The standard deviation is the square root of the variance.

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

Population standard deviation: $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$

Sample variance: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$

Sample standard deviation: $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

- You can use the following computational formulas to calculate standard deviation more easily.

Population standard deviation: $\sigma = \sqrt{\frac{\sum x^2 - N \cdot \mu^2}{N}}$

Sample standard deviation: $s = \sqrt{\frac{\sum x^2 - n \cdot \bar{x}^2}{n-1}}$

- The standard deviation of a set of data determines the average distance of the measurements from the mean. The larger the value, the greater the spread of the data. The units of the standard deviation are the same as for the mean.
- The z-score tells you the number of standard deviations that an observation in a data set is from the mean.

Population z-score: $z = \frac{x-\mu}{\sigma}$

Sample z-score: $z = \frac{x-\bar{x}}{s}$

- When you compare data values, it is possible to draw conclusions based on the data set results.
- There may or may not be a relationship between compared values.
- In some instances, graphs provide a stronger visual of the conclusion.
- You can use multiple bar graphs, split bar graphs, and relative split bar graphs to compare two similar data sets.
- Statistics are often used to represent certain points of view by manipulating graph axes, by citing only one measure of central tendency, or through measurement or sampling bias.
- It is key to perform a critical analysis of any statistical report.

- Statistics Canada collects statistics on Canadians and Canadian issues through the national census and through regular surveys.
- Statistics Canada holds a national census every five years.
- Data are published online at CANSIM. The data are available in table form, and often in graphical form.
- Statistical reports are available online on the Statistics Canada website, and are summarized in their bulletin, *The Daily*.
- When reading a statistical report, it is important to perform a critical analysis.

Solutions

9. Calculate

a) the mean, median, and mode

b) the range, standard deviation, and variance

c) the quartiles and interquartile range

The chart shows the temperature, in degrees Celsius, of coffee in 30 recently tested coffee makers.

76	68	72	73	70	69	68	73	81	72
66	85	72	72	69	72	67	74	73	69
75	65	70	71	71	71	68	74	79	73

Datum	Temperature	x - mean	(x - mean)^2
1	65	-6.9333	48.0711
2	66	-5.9333	35.2044
3	67	-4.9333	24.3378
4	68	-3.9333	15.4711
5	68	-3.9333	15.4711
6	68	-3.9333	15.4711
7	69	-2.9333	8.6044
8	69	-2.9333	8.6044
9	69	-2.9333	8.6044
10	70	-1.9333	3.7378
11	70	-1.9333	3.7378
12	71	-0.9333	0.8711
13	71	-0.9333	0.8711
14	71	-0.9333	0.8711
15	72	0.0667	0.0044
16	72	0.0667	0.0044
17	72	0.0667	0.0044
18	72	0.0667	0.0044
19	72	0.0667	0.0044
20	73	1.0667	1.1378
21	73	1.0667	1.1378
22	73	1.0667	1.1378
23	73	1.0667	1.1378
24	74	2.0667	4.2711
25	74	2.0667	4.2711
26	75	3.0667	9.4044
27	76	4.0667	16.5378
28	79	7.0667	49.9378
29	81	9.0667	82.2044
30	85	13.0667	170.7378
Total	2158	Total	531.8667

a) Mean = Sum ÷ # of Data

= 2158 ÷ 30

= 71.93°C

Median = Middle value(s)

= 72°C

Mode = Occurs most often

= 73°C

b) Range = Highest - Lowest

= 85 - 65

= 20°C

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

SD = √(531.8667 ÷ 29)

SD = 4.282549°C

Variance = SD^2

Variance = 18.34023°C^2

c) Upper Quartile = ¾ of the data

= 73°C

Lower Quartile = ¼ of the data

= 69°C

IQR = UQ - LQ

= 73 - 69

= 4

10. **a)** Is there any value(s) in the data set that could potentially change the outcome of the measures of central tendency? Explain.

b) Remove the value(s) identified in part a) and recalculate the mean, median, and mode. Which measure of central tendency is most appropriate to describe the distribution of the temperatures? Explain why.

c) What makes the other two measures less appropriate? Explain why.

The chart shows the temperature, in degrees Celsius, of coffee in 30 recently tested coffee makers.

76	68	72	73	70	69	68	73	81	72
66	85	72	72	69	72	67	74	73	69
75	65	70	71	71	71	68	74	79	73

a) We need to check for ouliers. An outlier is any value that is more than 1.5 times the IQR above/below the UQ/LQ.

Upper limit = UQ + 1.5(IQR)
= 73 + 1.5(4)
= 79

Lower limit = LQ - 1.5(IQR)
= 69 - 1.5(4)
= 63

There are two data points above 79 and none below 63. Therefore there are two outliers in this data set.

b) Recalculating gives a new mean of 1992 ÷ 28 = 71.14℃, a new median of 71.5℃ and the mode is still 72℃.

The median is the best measure to represent the average temperature as it is in the middle of the data set.

c) Mean - affected by outliers.

Mode - It is now the biggest of the three measures of central tendency

11. **a)** Create a frequency table by grouping the data into intervals.

b) Create a histogram and a box and whisker plot of the data.

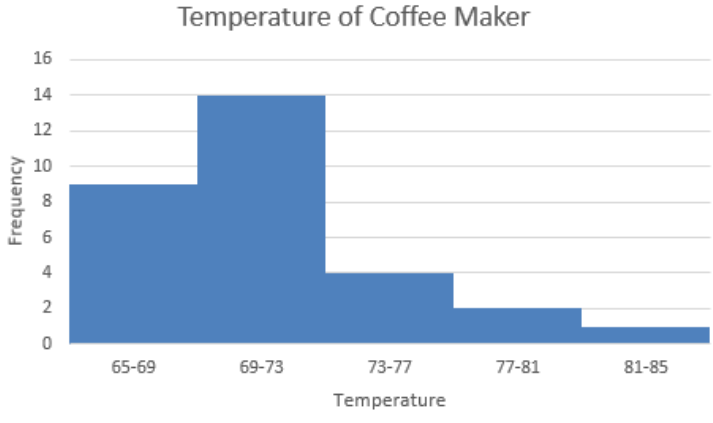
The chart shows the temperature, in degrees Celsius, of coffee in 30 recently tested coffee makers.

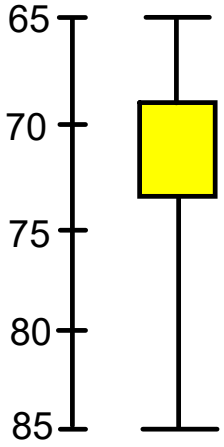
76	68	72	73	70	69	68	73	81	72
66	85	72	72	69	72	67	74	73	69
75	65	70	71	71	71	68	74	79	73

a)

Temperature	Frequency
65-69	9
69-73	14
73-77	4
77-81	2
81-85	1

b)





12. Coffee makers below the 5th and above the 95th percentiles are not recommended.

a) How many of these coffee makers will this include?

b) What are the temperatures of the coffee in the non-approved coffee makers?

The chart shows the temperature, in degrees Celsius, of coffee in 30 recently tested coffee makers.

76	68	72	73	70	69	68	73	81	72
66	85	72	72	69	72	67	74	73	69
75	65	70	71	71	71	68	74	79	73

a) Find the temperatures of the 5th and 95th percentiles.

$R = \frac{p}{100}(n+1)$

R = 0.05(30 + 1)

R = 1.55

$R = \frac{p}{100}(n+1)$

R = 0.95(30 + 1)

R = 29.45

Round down to nearest whole number.

R = 1

Find midpoint of 1st & 2nd values

5th percentile = (65 + 66)/2

= 65.5℃

R = 29

Find midpoint of 29th & 30th values

95th percentile = (81 + 85)/2

= 83℃

There is one coffee maker below the 5th percentile and one above the 95th percentile, so both of these are not recommended.

b) The non-approved coffee makers have temperatures of 65℃ and 85℃.

13. You want to make a generalization about variability of coffee temperatures in coffee makers. Do you have enough information to make this claim? If not, explain what other pieces of information you would need.

The chart shows the temperature, in degrees Celsius, of coffee in 30 recently tested coffee makers.

76	68	72	73	70	69	68	73	81	72
66	85	72	72	69	72	67	74	73	69
75	65	70	71	71	71	68	74	79	73

The sample given is of 30 coffee makers which is probably enough to draw a conclusion.

In making a generalization about the variability of the temperature, we can use a box plot to identify the interquartile range (middle 50% of the data). This strips out coffee makers that are hotter or colder than this 50%.

From previous questions we know that the IQR is from 69℃ to 73℃.