Chapter 1 Exam Review – Graphical Displays of Data

MDM4U

David Chen

Section 1.2 – Displaying Categorical Data

1) A researcher asked 150 high school students what their favourite fast food restaurant was. The results are in the table below:

Restaurant	Number of Students	Relative Frequency
McDonald's	22	**14.7%**
Wendy's	38	**25.3%**
Subway	22	**14.7%**
Harvey's	11	**7.3%**
Pizza Pizza	29	**19.3%**
A&W	6	**4%**
KFC	9	**6%**
Other	13	**8.7%**

a) What type of variable is 'favourite fast food restaurant?'

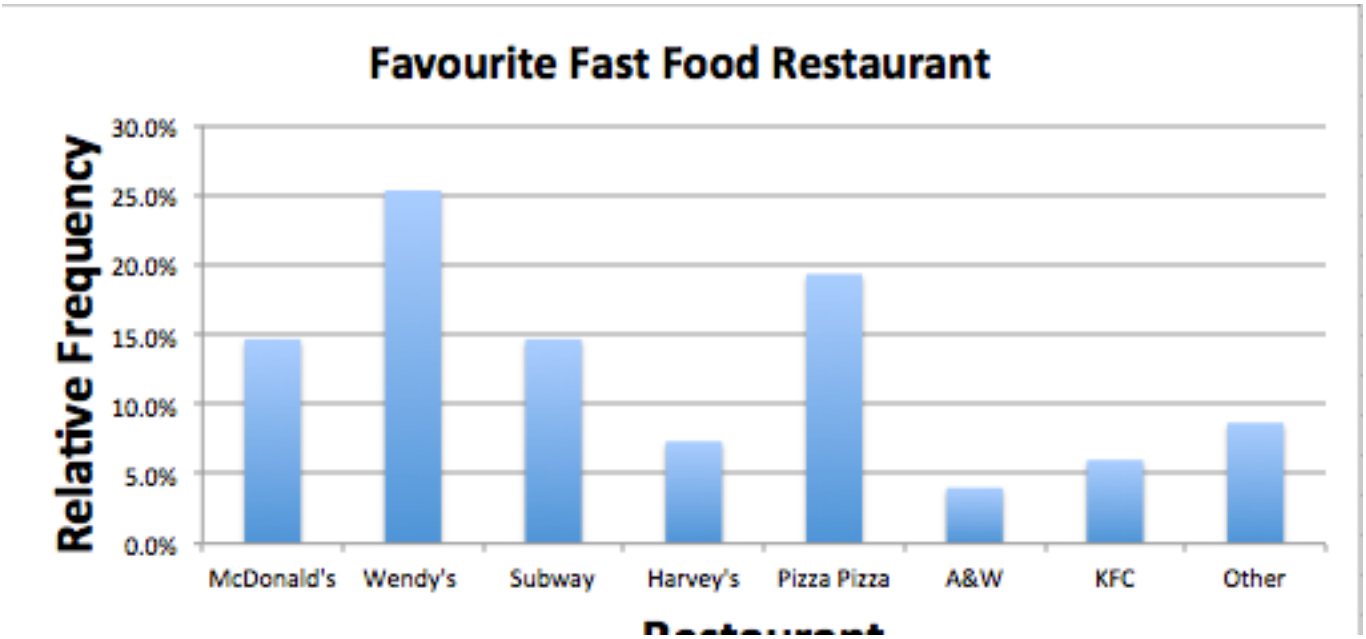
Categorical, nominal

b) Would it be more appropriate to make a histogram or bar graph to display this data?

Bar graph's are more appropriate for categorical data

c) Complete the relative frequency column

d) Display the relative frequencies using a bar graph or histogram.



2) A student is interested in whether there is a relationship between gender and major at her college. She randomly sampled some men and women on campus and asked them if their major was part of the natural sciences (NS), social sciences (SS), or humanities (H). Her results appear in the table below.

		Major			
		NS	**SS**	**H**	**Total**
Gender	**Women**	15 =27.3%	22 =40%	18 =32.7%	55
	Men	13 =52%	8 =32%	4 =16%	25
	Total	28	30	22	80

a) Complete the totals

b) Determine if major depends on gender by calculating the conditional distribution of major based on gender (row percentages).

c) Use your conditional distribution to describe the relationship between the variables.

Based on the conditional distribution of major based on gender, it appears that major does depend on gender. A higher percentage of males are enrolled in natural sciences, while there is a higher percentage of females in social sciences and humanities.

Section 1.3 – Displaying Quantitative Data

3) The number of goals by Jaromir Jagr in each of his 21 NHL seasons is recorded below

27, 32, 34, 32, 32, 72, 47, 35, 44, 42, 52, 31, 36, 31, 54, 30, 25, 19, 16, 24, 17

a) Construct a stem-and-leaf plot to display the data

Stem	Leaf
1	6 7 9
2	4 5 7
3	0 1 1 2 2 2 4 5 6
4	2 4 7
5	2 4
6	
7	2

b) Determine the percent of seasons where greater than 45 goals were scored.

$= 100\left(\frac{4}{21}\right) \cong 19\%$

c) Use the chart below to show the five number summary for Jagr's goals. Also compute the IQR. [3]

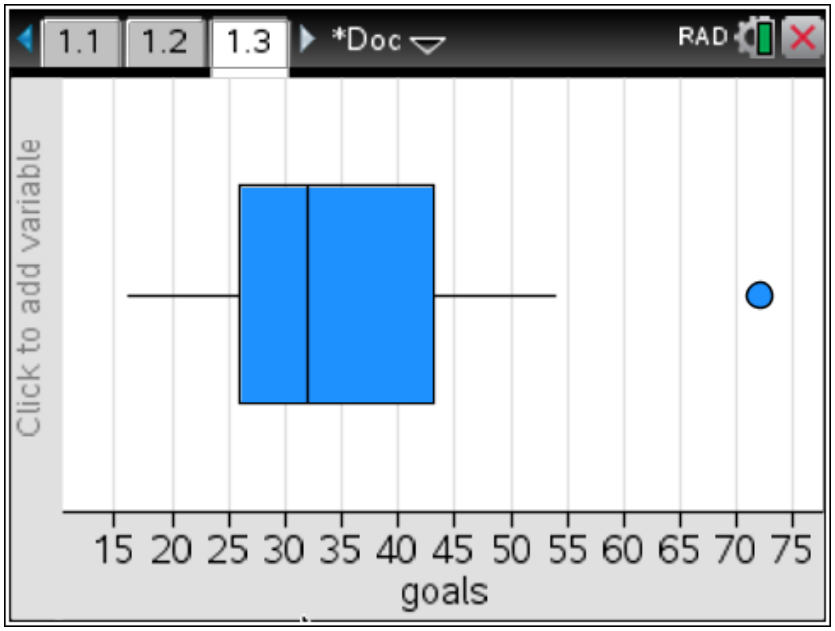
Max	72
Min	16
Q_1	26
Q_2	32
Q_3	43
IQR	17

d) Determine if there are any outliers in the data. Show your work including upper and lower threshold values.

$Lower\ threshold = Q_1 - 1.5(IQR) = 26 - 1.5(17) = 0.5$
$Upper\ threshold = Q_3 + 1.5(IQR) = 43 + 1.5(17) = 68.5$

Therefore, 72 is an outlier

e) Create a boxplot to display the data.



4) The heights of the 2013 Toronto Raptors (in centimeters) are listed below:

201, 183, 191, 211, 201, 201, 203, 213, 206, 206, 183, 208, 198, 198, 211

a) Determine the range of the data.

$Range = 213 - 183 = 30$

b) Determine an appropriate bin width that will divide the data into 7 intervals.

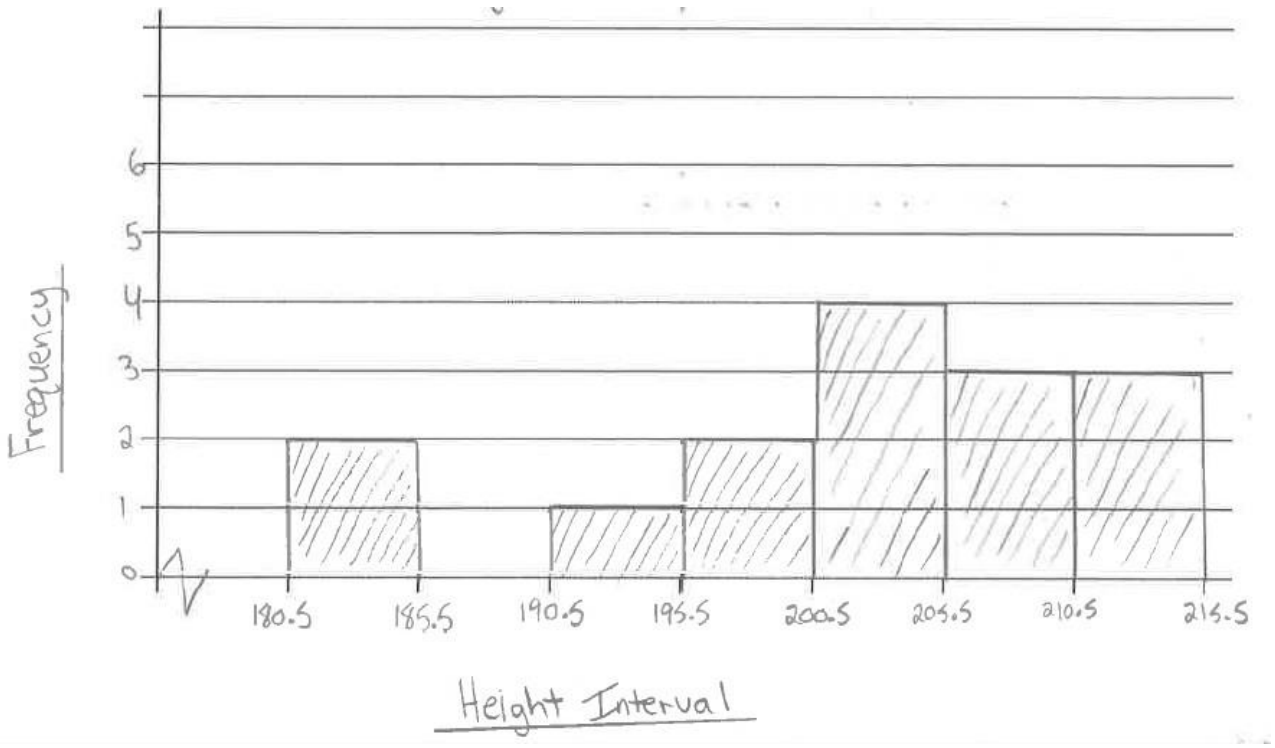
$$Bin\ Width = \frac{rounded\ range}{number\ of\ intervals} = \frac{35}{7} = 5$$

c) Create a frequency table for the data

$$Starting\ Point$$
$$= 183 - \frac{35 - 30}{2}$$
$$= 183 - 2.5$$
$$= 180.5$$

Height Interval	Frequency
180.5 – 185.5	2
185.5 – 190.5	0
190.5 – 195.5	1
195.5 – 200.5	2
200.5 – 205.5	4
205.5 – 210.5	3
210.5 – 215.5	3

d) Create a histogram of the data



Section 1.5 – Linear Regression Using Technology

5) Two variables have a correlation coefficient of $r = 0.9$. This indicates

a. a strong positive correlation
c. a strong negative correlation
b. a weak positive correlation
d. a weak negative correlation

6) If two variables have no correlation, their correlation coefficient would have a value of

a. +1
c. 100
b. −1
d. 0

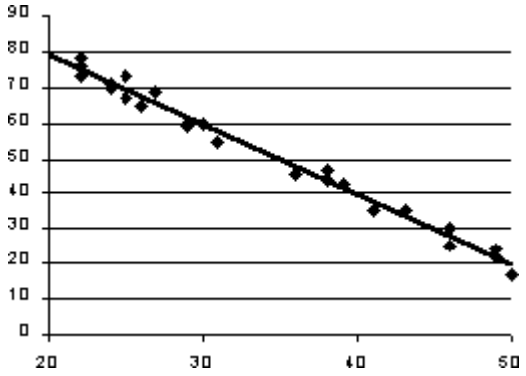
7) Two variables have a coefficient of determination of 0.64. The correlation coefficient could be

a. −0.64
c. -0.8
b. 0.41
d. 0.36

8) A relationship in which all data values lie on the regression line has a correlation coefficient of

a. 1
c. −1
b. 0
d. +1 or -1

9) The regression line shown would have a correlation coefficient closest to



a. +1
b. 0.5
c. -1
d. 0

10) The residuals for a set of data represent the

a. differences between consecutive *x*-values
b. vertical differences between data points and the line of best fit
c. data points that lie below the line of best fit
d. data points that do not lie on the line of best fit

11) If a set of data has a very strong correlation, the residual values will be

a. very large
b. positive
c. negative
d. very small

12) A coefficient of determination, $r^2 = 0.75$, indicates that

a. 75% of the data lie on the regression line
b. the slope of the regression line is 0.75
c. 75% of the variance in *y* can be explained by its approximate linear relationship with *x*
d. the data have a strong positive correlation

13) Which of the following is an example of a negative correlation?

a. amount of studying and mark on a test
b. temperature and number of kids at a pool
c. a person's arm length and leg length
d. number of people and slices of pizza per person

14) A set of data having small residual values means that

a. the correlation coefficient is close to 0
b. there is a positive correlation
c. there is a negative correlation
d. there is a strong correlation

15) A positive residual value means that the data point lies:

a. Close to the line of best fit
b. Above the line of best fit
c. On the line of best fit
d. Far away from the line of best fit

16) What type of linear correlation is represented when the correlation coefficient is ---0.7?

a. Strong negative
b. Moderate negative
c. Weak negative
d. No correlation

17) What type of correlation is represented when the correlation coefficient is 0.41?

a. Strong positive
b. Moderate positive
c. Weak positive
d. No correlation

18) This table shows the data for the full---time employees of a small company.

Age (year)	33	25	19	44	50	54	38	29
Annual Income (in thousands)	33	31	18	52	56	60	44	35
Residuals	---4.099	3.1044	---2.993	2.2472	---0.6551	---1.257	1.1494	2.5028

a) Construct a scatterplot using your calculator

b) Find the equation of the regression line and interpret the slope and y---intercept in context.

Regression Equation: predicted income $= -0.864 + 1.150(age)$

Slope $= 1.15$; *for every* 1 *year increase in age, the model predicts a* \$1150 *increase in annual income*

$y -$ *intercept* $= -0.864$; *at age* 0, *the model predicts an annual income of* $-$ \$864

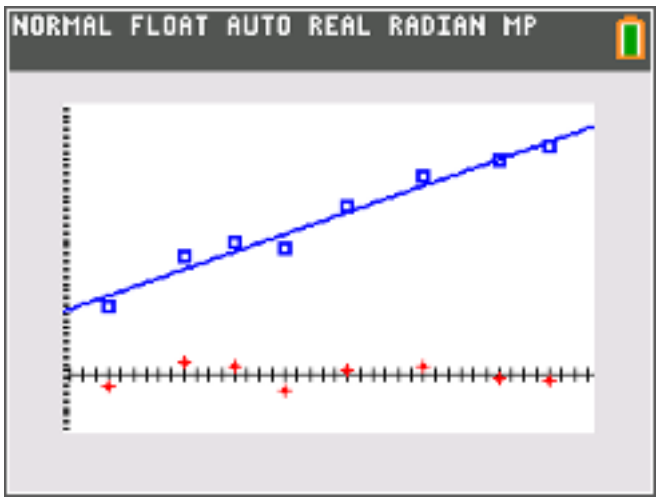
c) Find and interpret correlation coefficient, *r*.

$r = 0.98$; *this tells us there is a strong, positive, linear correlation between age and annual income*

d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

$r^2 = 0.965$; *this tells us that about* 96.5% *percent of the variation in annual income can be explained by its approximate linear relationship with age.*

e) Calculate the residual values, record them and analyze them using the residual plot to help. Is a linear model a good fit?



There is no distinguishable pattern in the residual plot and the residual values are relatively small. This tells us that the linear regression is a good model for the data.

f) Using the linear regression equation, what would you predict the annual income of a 40 year old to be?

$predicted\ income = -0.864 + 1.150(40) = 45.136$

The predicted annual income of a 40 year old is $45 136