Chapter 1 Review

MDM4U

David Chen

Section 1.2 – Displaying Categorical Data

1) Students were asked who their favourite EDM artist was. The results are shown in the table below.

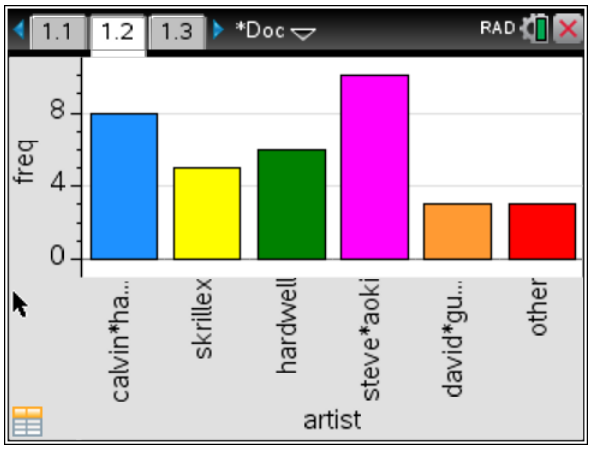
EDM Artist	Frequency
Calvin Harris	8
Skrillex	5
Hardwell	6
Steve Aoki	10
David Guetta	3
Other	3

a) What type of variable is 'favourite EDM artist'? *It is a nominal, categorical variable*

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

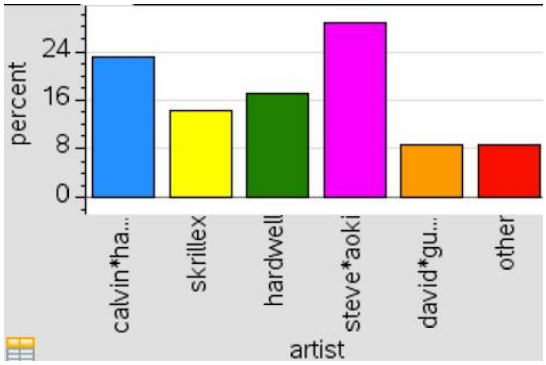
A bar graph is more appropriate to display categorical data.

c) Display the data using a bar graph or histogram.



d) Create a relative frequency table and corresponding relative frequency graph.

EDM Artist	Frequency	Relative Frequency
Calvin Harris	8	22.9%
Skrillex	5	14.3%
Hardwell	6	17.1%
Steve Aoki	10	28.6%
David Guetta	3	8.6%
Other	3	8.6%



2) A study among the Pima Indians of Arizona investigated the relationship between a mother's diabetic status and the appearance of birth defects in her children. The results appear in a two-way table below.

		Mother's Diabetic Status			
		Non-diabetic	**Pre-diabetic**	**Diabetic**	**Total**
Number of Birth Defects	**None**	754 =96.1%	362 =96.5%	38 =80.9%	1154
	One or More	31 =3.9%	13 =3.5%	9 =19.1%	53
	Total	785	375	47	1207

a) Copy the table and complete the totals

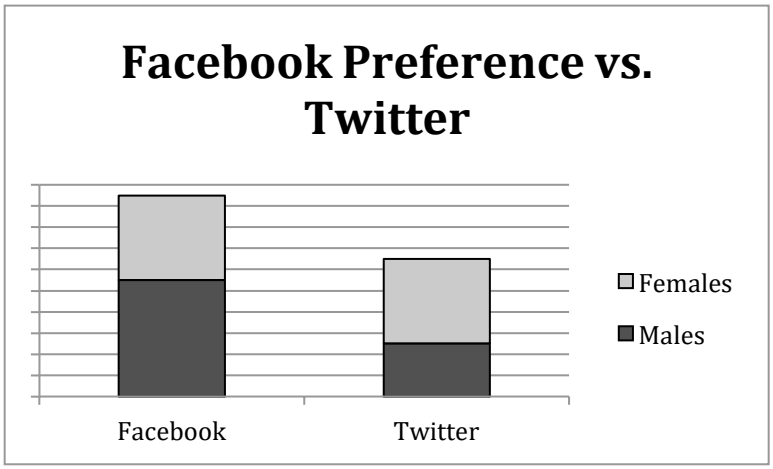
b) Calculate the conditional distribution of birth defects based on diabetic status.

The distribution of birth defects based on diabetic status is shown in the table (column percentages were calculated).

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

It seems that there is an association between a mother's diabetic status and the appearance of birth defects in her children. Non-diabetics and pre-diabetics appear to have babies with birth defects at about the same rate. However, those with diabetes have a much higher rate of babies with birth defects.

3) Lily wanted to find out if gender affects a person's preference of social networking sites. She asked her classmates in gr12 to answer the following question: "Would you rather have a profile on Facebook or Twitter?" She displayed her results in the graph below.



a) Lily forgot to add her vertical scale. If 5 males preferred Twitter, how many people did Lily interview altogether? **Indicate scale on graph**.

$n(interviewed) = 13 + 19 = 32$

b) Lily claims that social network preference depends on gender. Using her data, give one argument to support or refute Lily's claim. Creating a two-way contingency table and calculating a conditional distribution may be helpful.

		Social Network Preference		
		Facebook	**Twitter**	**Total**
Gender	**Male**	11 =68.75%	5 =31.25%	16
	Female	8 =50%	8 =50%	16
	Total	19	13	32

The row percentages were calculated to see if social network preference depends on gender.

Possible argument supporting Lily: The data shows that males are more likely to prefer facebook and less likely to prefer twitter compared to females.

Possible argument against Lily: Females show no preference between the facebook and twitter.

Section 1.3 – Displaying Quantitative Data

4) The number of points scored by a basketball team this season are recorded below

45, 71, 55, 62, 57, 68, 62, 48, 52, 60, 59, 75, 51, 49, 57, 56, 54, 63, 55, 67, 61, 58

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

Stem	Leaf
4	5 8 9
5	1 2 4 5 5 6 7 7 8 9
6	0 1 2 2 3 7 8
7	1 5

b) Determine the number of games where fewer than 55 points were scored

6

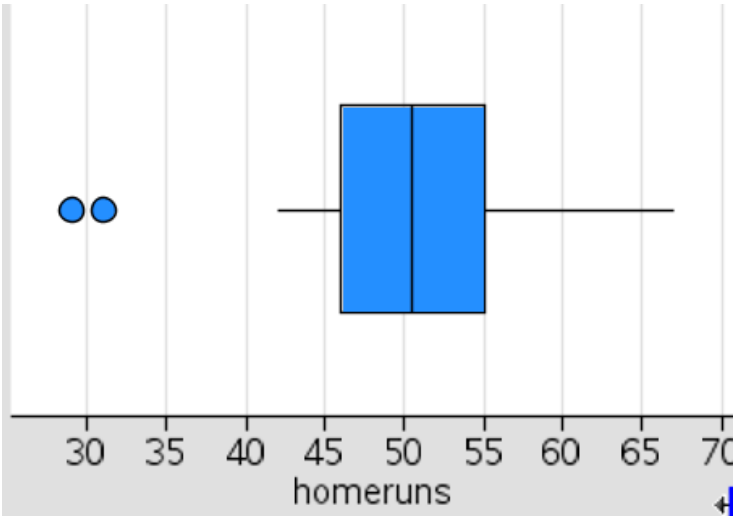
c) In what percent of games were more than 65 points scored

$\% > 65 = 100\left(\frac{4}{22}\right) = 18.2\%$

5) During the early part of the 1994 baseball season, many fans and players noticed that the number of home runs being hit seemed unusually large. Here are the data on the number of home runs hit by National League teams in the early part of the 1994 season. Calculate the five number summary for the data and then display it using a boxplot. Make sure to check for outliers.

29	31	42	46	47	48	48
53	55	55	55	63	63	67

Min = 29
$Q_1 = 46$
$Q_2 = 50.5$
$Q_3 = 55$
Max = 67
$IQR = 9$
Lower Threshold $= 46 - 1.5(9) = 32.5$
Upper Threshold $= 55 + 1.5(9) = 68.5$
Therefore 29 and 31 are outliers



6) The students in two high school physical education classes measured their maximum vertical jump. The results (in centimetres) are given in the following table.

32	26	20	42	61	12	34	39	40	21
18	24	45	52	26	13	28	33	41	50
38	28	31	17	43	29	30	16	35	45
20	22	29	14	38	30	53	60	14	26
29	30	40	34	18	41	33	38	27	15

a) Determine the range of the data.

$Range = 61 - 12 = 49$

b) Determine an appropriate bin width that will divide the data into 5 intervals.

$$Bin\ width = \frac{rounded\ range}{\#\ of\ intervals} = \frac{50}{5} = 10$$

c) Create a frequency table for the data

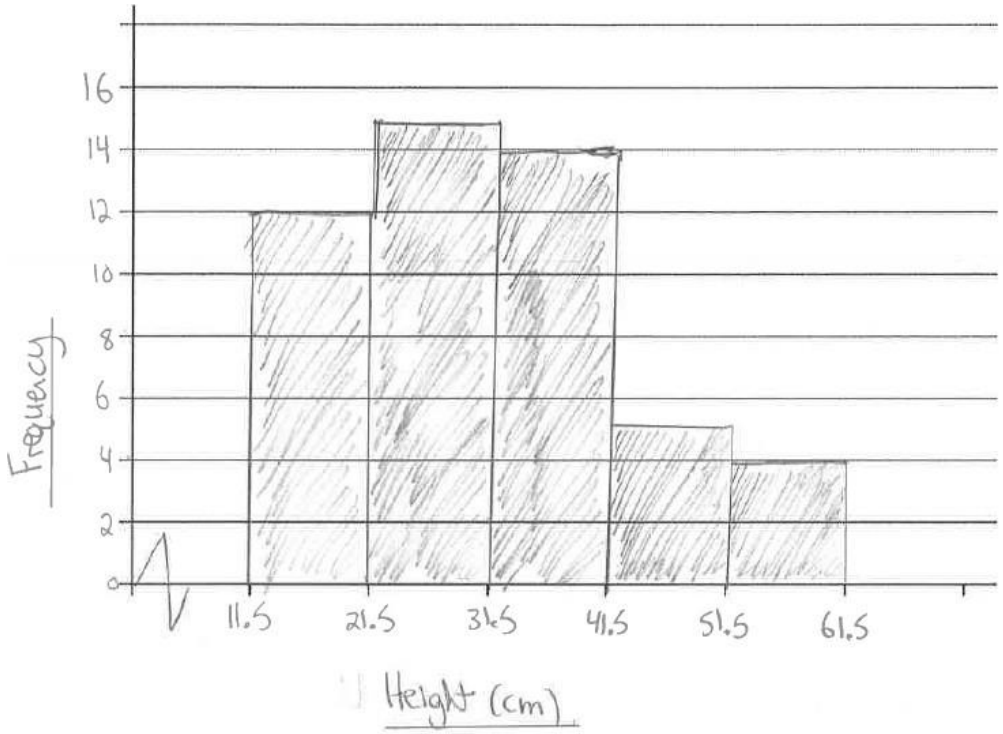
$$Starting\ point = 12 - \frac{50 - 49}{2}$$

$$= 12 - 0.5$$

$$= 11.5$$

Jump Interval	Frequency
11.5 – 21.5	12
21.5 – 31.5	15
31.5 – 41.5	14
41.5 – 51.5	5
51.5 – 61.5	4

d) Create a histogram of the data



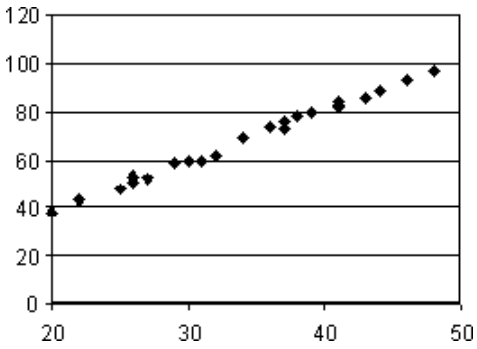
Section 1.4,1.5,1.6 – Linear Regression

7) The amount of time spent studying for a math test and the mark achieved would most likely show

a) A positive correlation
b) A negative correlation
c) No correlation
d) What is studying?

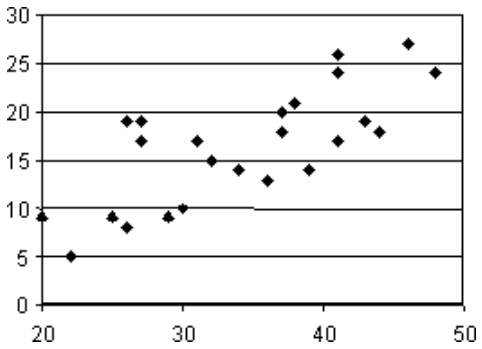
8) Describe the strength and direction of correlation represented by each of the following scatter plots.

a)



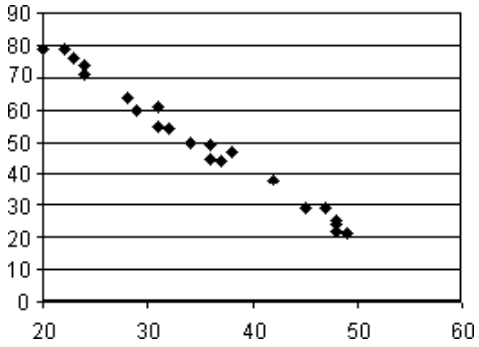
Strong, positive

b)



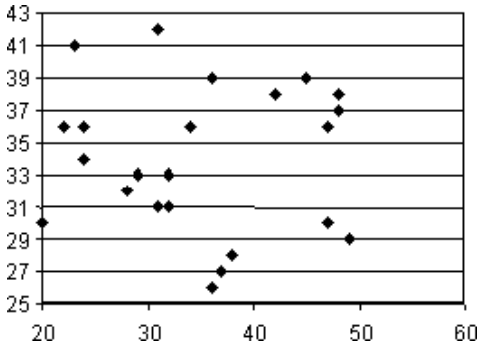
Moderate or weak, positive

c)



Strong, negative

d)



No correlation

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

a) A strong positive correlation
b) A weak positive correlation
c) A strong negative correlation
d) A weak negative correlation

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

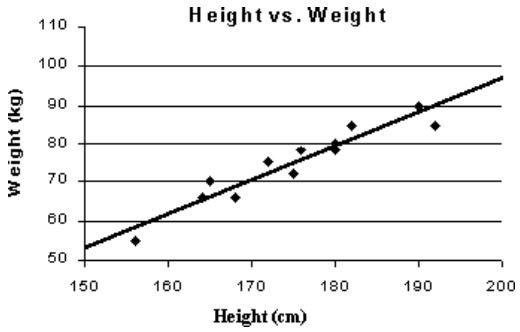
a) +1
b) -1
c) 100
d) 0

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

a) -0.64
b) 0.41
c) -0.8
d) 0.36

12) For the regression line shown, the coefficient of determination would be closest to

a) +1
b) 0
c) -1
d) 0.25



13) The residuals for a set of data represent the

a) Differences between consecutive *x*-values
b) Vertical differences between data point 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

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

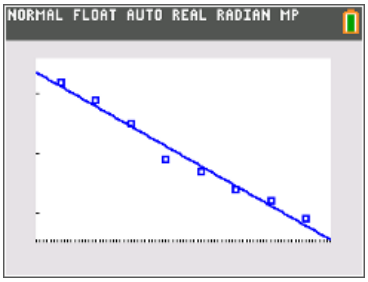
a) Very large
b) Positive
c) Negative
d) Very Small

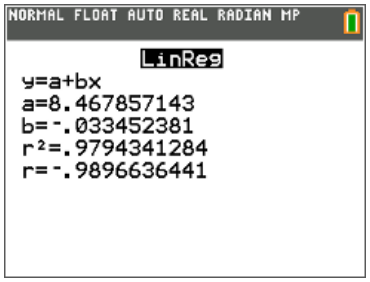
15) A coefficient of determination, $r^2 = 0.75$, indicates that

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 the approximate linear relationship with *x*
d) The data have a strong positive correlation

16) The amount of fuel consumed by a car travelling a distance of 100 km was measured at various speeds. The data recorded appears below:

Speed (km/h)	Gas Consumed (L)	Residuals
10	8.2	0.06667
20	7.9	0.10119
30	7.5	0.03571
40	6.9	-0.2298
50	6.7	-0.0952
60	6.4	-0.0607
70	6.2	0.07381
80	5.9	0.10833





a) Construct a scatterplot using your calculator

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

$\widehat{gas\ consumed} = 8.468 - 0.0335(speed)$

The slope of -0.0335 tells us that for every 1 km/h increase in speed, our model predicts a decrease of 0.0335 L of gas consumed for the 100 km trip.

The y-intercept of 8.468 tells us that at a speed of 0 km/h, our model predicts a usage of 8.468 L of gas for a 100 km trip.

c) Find and interpret correlation coefficient, r.

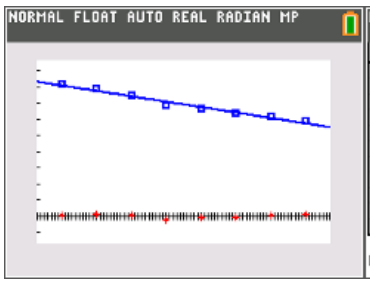
$r = -0.9897$*, this tells us there is a strong, negative, linear correlation between speed and gas consumed.*

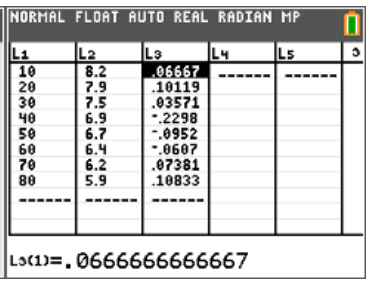
d) Find the coefficient of determination, r^2. Interpret it in the context of this data.

$r^2 = 0.9794$*; this tells us that about 97.94% of the variation in gas consumed can be explained by the approximate linear relationship with speed.*

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.





17) Consider the following data:

Grade 12 Average	85	90	76	78	88	84	76	96	86
First Year Average	74	83	68	70	75	72	64	91	78
Residuals	-1.767	1.4771	2.5925	2.2903	-4.221	-2.616	-1.408	2.5705	1.0815

Program the above data into a graphing calculator and performing a linear regression to complete the following analysis:

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

$\widehat{first\ year\ average} = -22.076 + 1.151(grade\ 12\ average)$

The slope of 1.151 tells us that for every 1% increase in grade 12 average, the model predicts a 1.151% increase in first year university average.

The y-intercept of -22.076 tells us that with a grade 12 average of 0%, the model predicts a first year university average of -22.076%.

b) Using the linear regression equations, what would you predict the first year marks to be of students who achieved the following grade 12 marks: (2 marks)

i) 100%; 93% ii) 67%; 55%

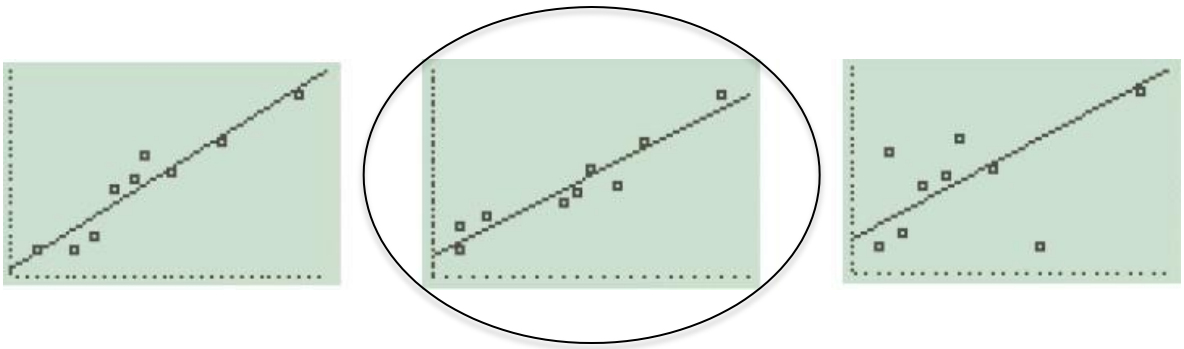
c) Find and interpret correlation coefficient, *r*.

$r = 0.9504$*; this tells us there is a strong, positive, linear correlation between grade 12 mark and first year university mark.*

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

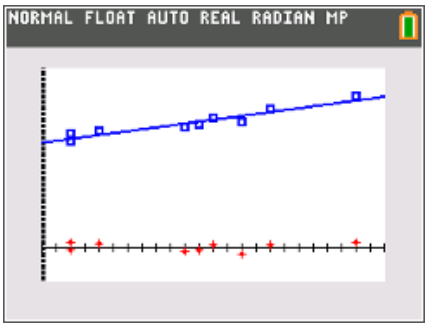
$r^2 = 0.9032$*; this tells us that about 90.32% of the variation in first year university average can be explained by the approximate linear relationship with grade 12 average.*

e) Which of the following screen shots best represents this data (CIRCLE IT):



f) 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.



18) Last year, five randomly selected students took a math aptitude test before they began their statistics course. In the table below, the x column shows scores on the aptitude test. The y column shows final statistics grades.

a) Complete the chart

	Math aptitude test score, x	**Statistics Grade, y**	x^2	y^2	xy
	95	85	9025	7225	8075
	85	95	7225	9025	8075
	80	70	6400	4900	5600
	70	65	4900	4225	4550
	60	70	3600	4900	4200
$\sum$	390	385	31150	30275	30500

b) Determine the equation of the least squares regression line ($\hat{y} = a + bx$). Interpret the slope and y-intercept in context.

Slope = $b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$

y-intercept = $a = \bar{y} - b\bar{x}$

Slope:

Slope = $b = \frac{(n\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{5(30500) - (39\#)(385)}{5(3115\$) - (39\#)^2} = \frac{235\$}{365\$} = 0.6438$

This indicates that for every 1% increase in the aptitude test score, the model predicts a 0.6438% increase in statistics grade.

y-intercept:

$\bar{x} = \frac{\sum x}{n} = \frac{390}{5} = 78$

$\bar{y} = \frac{\sum y}{n} = \frac{385}{5} = 77$

$y - intercept = a = \bar{y} - b\bar{x} = 77 - 0.6438(78) = 26.78$

This tells us that with an aptitude score of 0%, the model predicts a statistics grade of 26.78%.

$\hat{y} = a + bx \rightarrow predicted\ statistics\ score = 26.78 + 0.6438(aptitude\ score)$

c) Compute the correlation coefficient using the formula. Interpret r and r^2 in context.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} = \frac{5(30500) - (390)(385)}{\sqrt{[5(31150) - (390)^2][5(30275) - (385)^2]}} = \frac{2350}{3390.796367} = 0.693$$

$r = 0.693$; this indicates that there is a moderate, positive, linear correlation between aptitude test score and statistics grade.

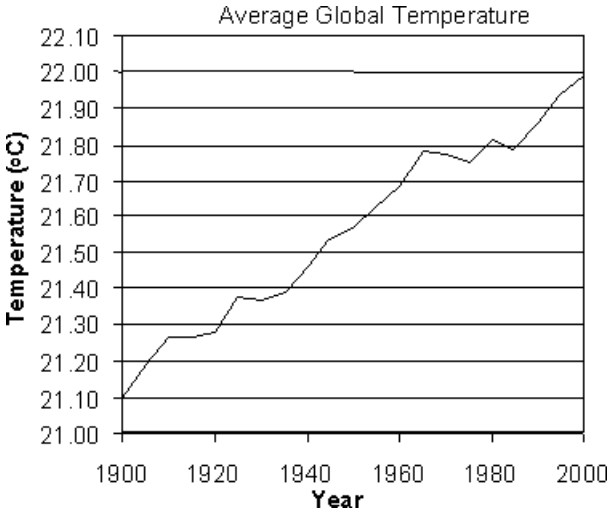
$r^2 = 0.48$; this indicates that approximately 48% of the variation in statistics score can be explained by the approximate linear correlation with aptitude test score.

Section 1.7 – Misrepresentations of Data

19) Graphs can be misleading because

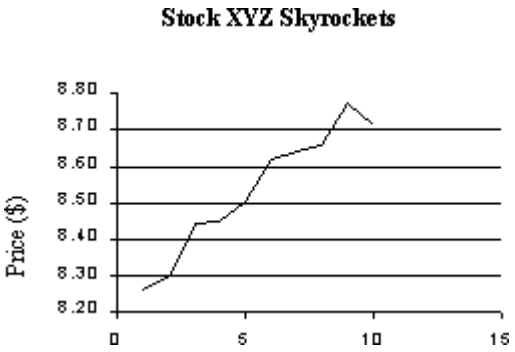
- **a)** They may use varying scales
- **b)** They may have suggestive captions
- **c)** They may be based on small samples
- **d) All of the above**

20) Which statement is most accurate about the following graph?



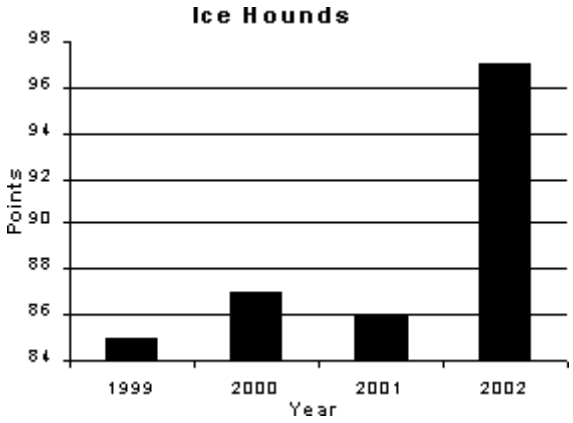
- **a)** The temperature increased dramatically in the last 100 years.
- **b)** The graph makes the temperature increase seem less than it really is
- **c) The temperature has increased slightly in the last 100 years**
- **d)** The graph gives very little information

21) Give reasons why the following graph is misleading.



- *The y-axis has been truncated which exaggerates the changes in price.*
- *Also the scale goes up by very small increments which adds to the exaggeration of change in price.*
- *The x-axis is missing at title, which makes the numbers along the x-axis meaningless.*

22) Based on the following graph, it appears as though the Ice Hounds had far more points in 2002 than in previous seasons. Explain how the graph could be changed to prevent this misrepresentation.



- The problem with the graph is that the y-axis has been truncated which exaggerates the differences in points between the years.
- This could be fixed by changing the scale on the y-axis to start at 0 and increase at equal increments.