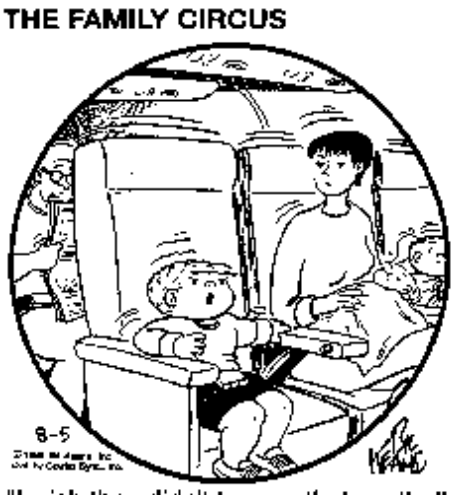
Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis



"I wish they didn't turn on that seatbelt sign so much! Every time they do, it gets bumpy."

Think, Pairs, Share:	What type of causal relationship does Billy believe exists in the Family Circus cartoon above? **Explain** your answer.

Billy believes there is a cause and effect relationship. That is, the seatbelt sign *causes* the turbulence.

Think, Pairs, Share:	What type of causal relationship really exists in the Family Circus cartoon above? **Explain** your answer.

This is a reverse cause and effect relationship. As turbulence is detected ahead, the pilot will put the seatbelt light on in anticipation of the bumpy ride that will follow.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

Think, Pairs, Share: Statistics in the Media

Consider the claim, **"4 out 5 dentists recommend Crest."**

1 Is this a **valid claim**? What statistical methods would be necessary to make this a **valid claim**?

We do not have enough information to determine if this is a valid claim. We need to know

- **sample size**
- **data collection method ... how the sample was chosen and what questions were asked**

2 What is the motivation for the statistical study?

The company wants to increase the sales of Crest toothpaste. However, they are not concerned with whether or not their survey is biased if it gives them favourable results.

Activity 1: Sample Size and Technique Page 203

A manager wants to know if a new aptitude test accurately predicts employee productivity. The manager has **all 30 current employees** write the test and then compares their **scores** to their **productivities** as measured in the most recent performance reviews. The data is ordered alphabetically by employee surname. In order to simplify the calculations, the manager selects a **systematic sample** using **every seventh employee**. Based on this **sample**, the manager concludes that the company should hire only applicant's who do well on the aptitude test. **Determine whether the manager's analysis is valid.**

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

Test Score	Productivity		Test Score	Productivity
98	78		95	72
57	81		56	72
82	83		71	90
76	44		68	74
65	62		77	51
72	89		**59**	**65**
91	**85**		83	47
87	71		75	91
81	76		66	77
39	71		48	63
50	66		61	58
75	90		78	55
71	48		**70**	**73**
89	**80**		68	75
82	83		64	69

- Split class into 5 groups.
- Have students complete 5 regressions on the **data**, that is **every 7th employee**. (Ie. Those in red in the chart.)
  - Linear
  - Quadratic
  - Cubic
  - Power
  - Exponential

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

Regression	Equation	r or r^2
Linear	$y \doteq 0.552x + 33.128$	$r \doteq 0.978\ 197$
Quadratic	$y \doteq -0.004x^2 + 1.231\ x + 8.220$	$r^2 \doteq 0.959\ 936$
Cubic	$y \doteq 0.004x^3 - 0.785x^2 + 57.624\ x - 1330.639$	$r^2 \doteq 0.999\ 999\ 999$
Power	$y \doteq 6.810x^{0.555}$	$r^2 \doteq 0.\ 964\ 852$
Exponential	$y \doteq 42.494\ (1.007)^x$	$r^2 \doteq 0.957\ 796$

Discussion

1 Which regression best matches the data? **Justify** your choice. Does it support the managers hypothesis?

A cubic regression of the systematic sample produces the following equation of best fit

$$y \doteq 0.0036x^3 - 0.785x^2 + 57.62\ x - 1330.6$$

and a coefficient of determination of

$$r^2 \doteq 0.999\ 999\ 999$$

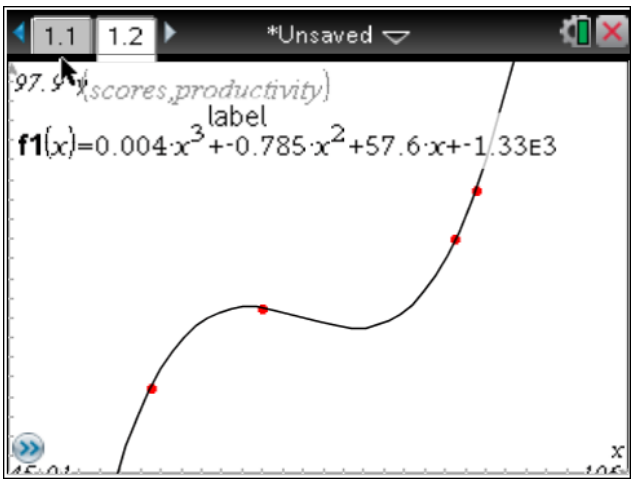
The cubic regression represents the curve of best fit since its coefficient of determination / correlation coefficient is the closest to perfect.

This shows a strong (almost perfect) cubic correlation between productivity and scores on the aptitude test which appears to support the manager's conclusion.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

2	Sketch the function that represents the cubic regression of your data. Is the correlation reasonable? **Explain** your answer.



It is important to view the graph of the regression because the cubic regression has intervals where it changes from increasing to decreasing and back to increasing. According to the manager hypothesis the data must be increasing and there is no reasonable scenario that could cause the productivity to decrease for an interval of test scores.

If the cubic that represented the curve of best fit for the regression was *always* increasing. Therefore, it could appear that a cubic correlation could represent the data well.

However, in either case, the domain and range must both be restricted to positive values for this function since negative scores are not valid pieces of data.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

3	What problem exists with this study?

The manager has assumed that a systematic sample will be representative of the population.

The sample is very small and statistical fluctuations could seriously affect the results.

4	How could the study be improved?

Complete the regression(s) on ALL of the data since it does not have a large population.

5	Complete the regressions using ALL of the data in the chart. • Linear • Quadratic • Cubic • Power • Exponential

Regression	Equation	r or r^2
Linear	$y \doteq 0.146x + 60.791$	$r \doteq 0.154\ 232$
Quadratic	$y \doteq 0.003x^2 - 0.308\ x + 76.073$	$r^2 \doteq 0.027\ 485$
Cubic	$y \doteq 0.000\ 097x^3 - 0.017x^2 + 1.030x + 47.584$	$r^2 \doteq 0.028\ 369$
Power	$y \doteq 45.397x^{0.102}$	$r^2 \doteq 0.\ 011\ 262$
Exponential	$y \doteq 61.892\ (1.002)^x$	$r^2 \doteq 0.014\ 226$

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

5	Does the cubic regression still match the data? **Explain**.

A cubic regression of the population produces a curve of best fit with the equation

$$y \doteq 0.000\,097x^3 - 0.016\,8x^2 + 1.029\,06x + 47.58$$

and a correlation coefficient of

$$r^2 \doteq 0.028$$

Therefore, since the coefficient of determination is only 0.028, it is not a good match to the data.

6	Which regression best matches the data? **Justify** your answer. Does it support the managers hypothesis?

A linear regression of the population produces the following equation of best fit

$$y \doteq 0.15x + 60$$

and a coefficient of determination of

$$r \doteq 0.15$$

The linear regression represents the equation of best fit since its coefficient of determination / correlation coefficient is the closest to perfect.

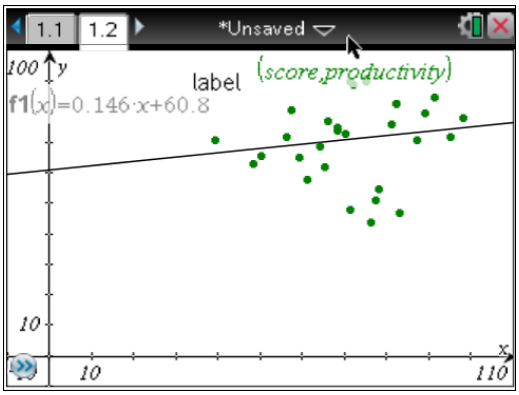
However, there is no curve of best fit that matches the data since NONE of the regressions results in a strong correlation between productivity and scores on the aptitude test.

Therefore, the aptitude test appears to be useless for predicting employee productivity and the manager's hypothesis is NOT supported.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

7	Sketch the function that represents the linear regression of your data. Are there any pieces of data that are outliers and could be affecting your results? **Explain** your answer.



Since the data appears to be in clusters, there are not any points that are significantly different than the other pieces of data. Therefore, we can conclude that there are not outliers and there are no data points that are reasonable to remove.

8	What conclusions can be made about the original study?

The sample was NOT representative because the manager chose a sample size that was too small to draw conclusions.

All of the data should have been used. However, it is still too small a population to make a conclusion about the company's hiring practices.

There does NOT appear to be any correlation between the data.

9	What conclusions can be made about outliers?

Small samples are vulnerable to outliers.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

10	If the employer had simply done a linear regression on the systematic sample, what assumptions would he have made?

A linear regression of the systematic sample produces the following line of best fit

$y \doteq 0.55x + 33$

with a correlation coefficient of

$r \doteq 0.98$

Thus, indicating a strong, positive linear correlation which supports the managers hypothesis that the aptitude test is a good measure of the employee's productivity and it has a reasonable fit to the type of the data that exists since it is not logical for the data to increase, decrease and increase as it appeared to in the cubic curve of best fit.

11	Is this conclusion valid?

No, this conclusion is not valid since a linear regression of the population produces the following line of best fit

$y \doteq 0.15x + 60$

with a correlation coefficient of

$r \doteq 0.15$

Thus, indicating no linear correlation which does NOT support the managers hypothesis that the aptitude test is a good measure of the employee's productivity. Therefore, the aptitude test appears to be useless for predicting employee productivity.

Since his sample was so small his results were significantly altered.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

12	What difference is there between the **correlation coefficients** and the **coefficients of determination** for the **systematic sample** and the **population**.

The correlation coefficients and the coefficients of determination for the systematic sample all indicate a strong correlation which leads to inaccurate conclusions. The correlation coefficients and the coefficients of determination for the population all indicate no correlation.

13	What **extraneous variables** exist that could affect the results?

- **reading is not required for the job, but is for the aptitude test**
- **book smart is different from ability**
- **lack of sleep before the test**
- **lack of breakfast**
- **previous exposure to the test (they applied and went through the process before)**
- **general well being (ie. sick today)**
- **emotional statement (ie. an upset at home leading to lack of focus)**
- **nerves**

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

Hidden (Lurking) Variables

- can also invalidate conclusions drawn from statistical results.
- **extraneous variables** that are difficult to recognize.

Examples of Hidden (Lurking) Variables

1 A survey finds a **correlation** between the proportion of high school students who own a car and the student's ages.

The **hidden variable** is the income of the students which is not obvious to the people completing the study.

2 A huge decrease in plane travel occurred in 2001.

The **hidden variable** is the World Trade Centre attacks with highjacked planes.

3 A huge increase in flu deaths in 1957.

The **hidden variable** is that this was the last occurrence of swine flu.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

Evaluating Claims from Statistical Studies

1. Is it free from intentional and unintentional bias.
2. Could outliers and extraneous variables affect your data?
3. Are there any unusual patterns that suggest the presence of a hidden variable?
4. Has causality been inferred with only correlational evidence?

For example, in the systematic sample, it would appear that an increased score on the aptitude test causes an increase in productivity. However, this has not been shown with anything but correlational evidence which we have since discovered is invalid.

Think, Pairs, Share

1 **Explain** how a small sample size can lead to invalid conclusions?

A small sample size may fail to reflect the characteristics of the population, due to statistical fluctuation. Fluctuations in larger sample sizes, by contrast, tend to cancel out.

Unit 2: Two Variable Statistics

Lesson 8: Critical Analysis

2	A city councillor states that there are problems with the management of the police department because the number of reported crimes in the city has risen despite increased spending on law enforcement. Comment on the validity of this argument.

The city councillor may not have considered that it is possible that

- **the number of reported crimes has increased because the increase in funding has resulted in more officers patrolling the city.**
- **since the officers are more visible people may be approaching them to report more crimes.**

An increase in population could also result in an increase in crimes and an increase in reported crimes.

The city councillor may not have considered hidden variables such as
- **hosting the G20 summit**
- **blackout in 2003**

3	Give an example of a hidden variable and **explain** why this variable would be hard to detect.

An increase in the draft age for junior hockey players (ie. OHL) into the NHL would be a hidden variable that can affect the number of players drafted, measured against the number of junior teams.

Homework: MHR Mathematics of Data Management
Pages 209 - 210 # 1, 3 - 5