3.5 Critical Analysis

Newspapers and radio and television news programs often run stories involving statistics. Indeed, the news media often commission election polls or surveys on major issues. Although the networks and major newspapers are reasonably careful about how they present statistics, their reporters and editors often face tight deadlines and lack the time and mathematical knowledge to thoroughly critique statistical material. You should be particularly careful about accepting statistical evidence from sources that could be biased. Lobby groups and advertisers like to use statistics because they appear scientific and objective. Unfortunately, statistics from such sources are sometimes flawed by unintentional or, occasionally, entirely deliberate bias. To judge the conclusions of a study properly, you need information about its sampling and analytical methods.

INVESTIGATE & INQUIRE: Statistics in the Media

1. Find as many instances as you can of statistical claims made in the media or on the Internet, including news stories, features, and advertisements. Collect newspaper and magazine clippings, point-form notes of radio and television stories, and printouts of web pages.
2. Compare the items you have collected with those found by your classmates. What proportion of the items provide enough information to show that they used valid statistical methods?
3. Select several of the items. For each one, discuss
   a) the motivation for the statistical study
   b) whether the statistical evidence justifies the claim being made

The examples in this section illustrate how you can apply analytical tools to assess the results of statistical studies.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about how statistics can be misused. Describe two examples of the misuse of statistics.

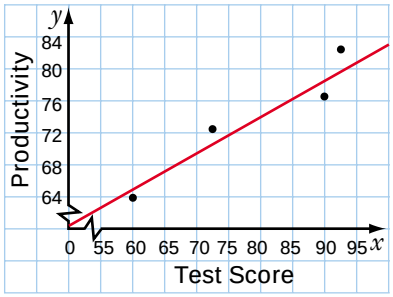
Example 1 Sample Size and Technique

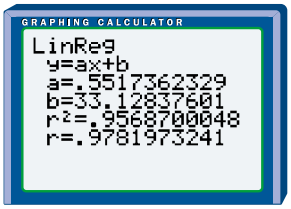
A manager wants to know if a new aptitude test accurately predicts employee productivity. The manager has all 30 current employees write the test and then compares their scores to their productivities as measured in the most recent performance reviews. The data is ordered alphabetically by employee surname. 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 applicants who do well on the aptitude test. Determine whether the manager's analysis is valid.

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

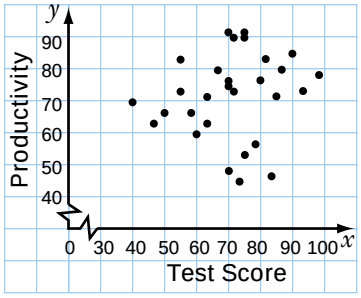
Solution

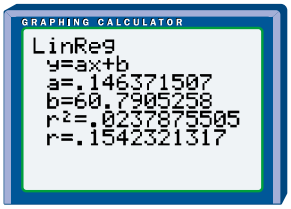
A linear regression of the systematic sample produces a line of best fit with the equation $y = 0.55x + 33$ and a correlation coefficient of $r = 0.98$, showing a strong linear correlation between productivity and scores on the aptitude test. Thus, these calculations seem to support the manager's conclusion. However, the manager has made the questionable assumption that a systematic sample will be representative of the population. The sample is so small that statistical fluctuations could seriously affect the results.





Examine the raw data. A scatter plot with all 30 data points does not show any clear correlation at all. A linear regression yields a line of best fit with the equation $y = 0.15x + 60$ and a correlation coefficient of only 0.15.





Thus, the new aptitude test will probably be useless for predicting employee productivity. Clearly, the sample was far from representative. The manager's choice of an inappropriate sampling technique has resulted in a sample size too small to make any valid conclusions.

In Example 1, the manager should have done an analysis using all of the data available. Even then the data set is still somewhat small to use as a basis for a major decision such as changing the company's hiring procedures. Remember that small samples are also particularly vulnerable to the effects of outliers.

Example 2 Extraneous Variables and Sample Bias

An advertising blitz by SuperFast Computer Training Inc. features profiles of some of its young graduates. The number of months of training that these graduates took, their job titles, and their incomes appear prominently in the advertisements.

Graduate	Months of Training	Income ($000)
Sarah, software developer	9	85
Zack, programmer	6	63
Eli, systems analyst	8	72
Yvette, computer technician	5	52
Kulwinder, web-site designer	6	66
Lynn, network administrator	4	60

a) Analyse the company's data to determine the strength of the linear correlation between the amount of training the graduates took and their incomes. Classify the linear correlation and find the equation of the linear model for the data.

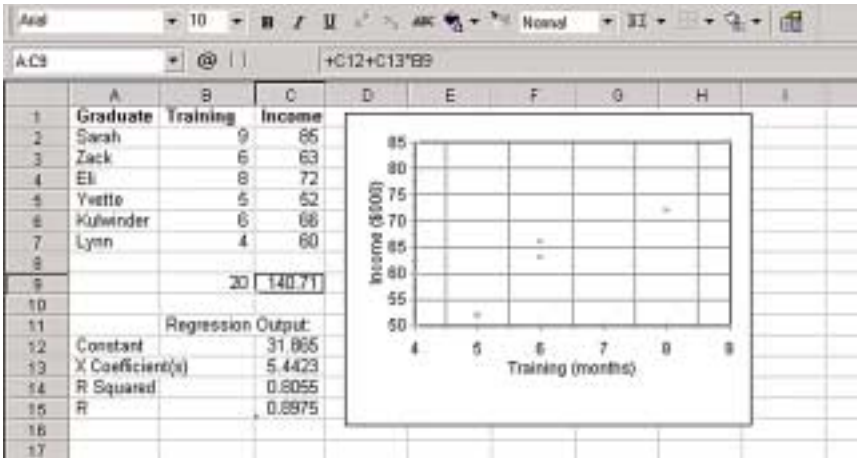
b) Use this model to predict the income of a student who graduates from the company's two-year diploma program after 20 months of training. Does this prediction seem reasonable?

c) Does the linear correlation show that SuperFast's training accounts for the graduates' high incomes? Identify possible extraneous variables.

d) Discuss any problems with the sampling technique and the data.

Solution

a) The scatter plot for income versus months of training shows a definite positive linear correlation. The regression line is $y = 5.44x + 31.9$, and the correlation coefficient is 0.90. There appears to be a strong positive correlation between the amount of training and income.



b) As shown in cell C9 in the screen above, substituting 20 months into the linear regression equation predicts an income of approximately

$y = 5.44(20) + 31.9$
$= 141$

Therefore, the linear model predicts that a graduate who has taken 20 months of training will make about $141 000 a year. This amount is extremely high for a person with a two-year diploma and little or no job experience. The prediction suggests that the linear model may not be accurate, especially when applied to the company's longer programs.

c) Although the correlation between SuperFast's training and the graduates' incomes appears to be quite strong, the correlation by itself does not prove that the training causes the graduates' high incomes. A number of extraneous variables could contribute to the graduates' success, including experience prior to taking the training, aptitude for working with computers, access to a high-end computer at home, family or social connections in the industry, and the physical stamina to work very long hours.

d) The sample is small and could have intentional bias. There is no indication that the individuals in the advertisements were randomly chosen from the population of SuperFast's students. Quite likely, the company carefully selected the best success stories in order to give potential customers inflated expectations of future earnings. Also, the company shows youthful graduates, but does not actually state that the graduates earned their high incomes immediately after graduation. It may well have taken the graduates years of hard work to reach the income levels listed in the advertisements. Further, the amounts given are incomes, not salaries. The income of a graduate working for a small start-up company might include stock options that could turn out to be worthless. In short, the advertisements do not give you enough information to properly evaluate the data.

Example 2 had several fairly obvious extraneous variables. However, extraneous variables are sometimes difficult to recognize. Such **hidden** or **lurking variables** can also invalidate conclusions drawn from statistical results.

Example 3 Detecting a Hidden Variable

An arts council is considering whether to fund the start-up of a local youth orchestra. The council has a limited budget and knows that the number of youth orchestras in the province has been increasing. The council needs to know whether starting another youth orchestra will help the development of young musicians. One measure of the success of such programs is the number of youth-orchestra players who go on to professional orchestras. The council has collected the following data.

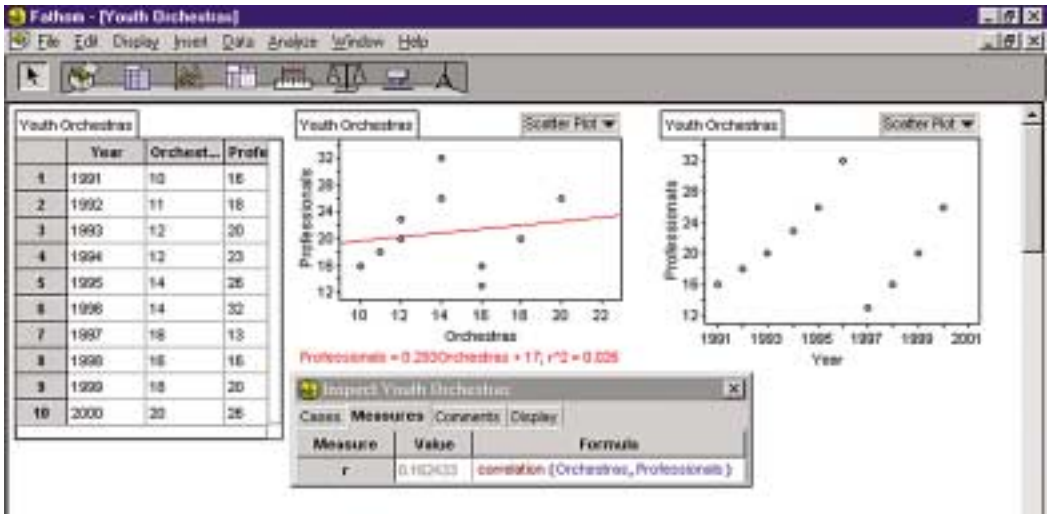
Year	Number of Youth Orchestras	Number of Players Becoming Professionals
1991	10	16
1992	11	18
1993	12	20
1994	12	23
1995	14	26
1996	14	32
1997	16	13
1998	16	16
1999	18	20
2000	20	26

a) Does a linear regression allow you to determine whether the council should fund a new youth orchestra? Can you draw any conclusions from other analysis?

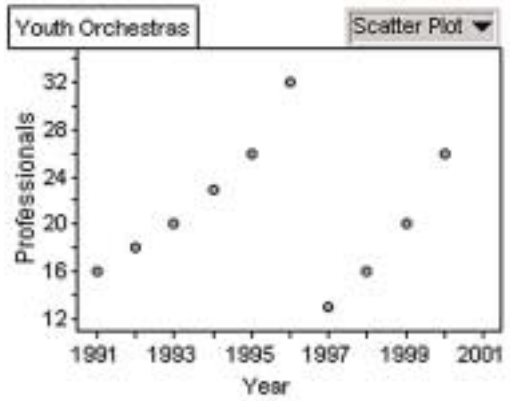
b) Suppose you discover that one of the country's professional orchestras went bankrupt in 1997. How does this information affect your analysis?

Solution

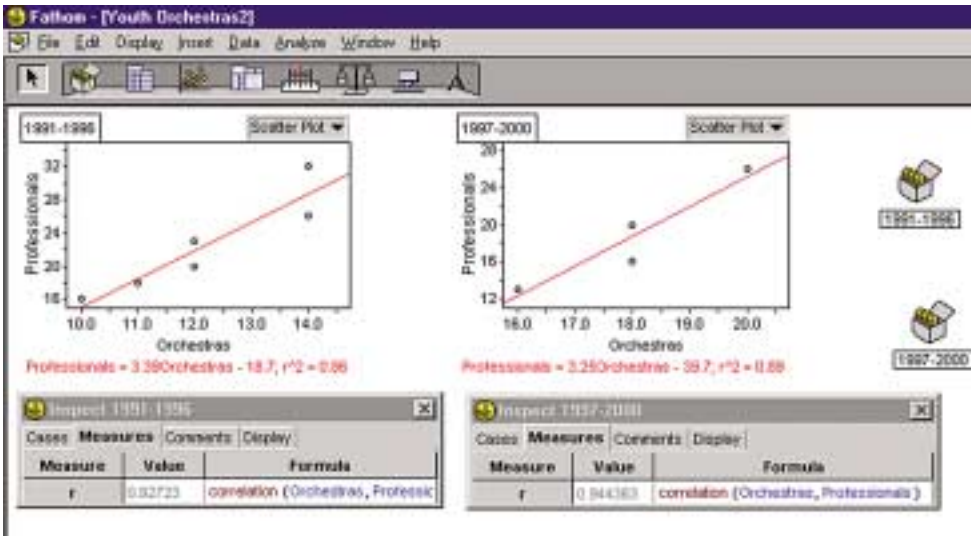
a) A scatter plot of the number of youth-orchestra members who go on to play professionally versus the number of youth orchestras shows that there may be a weak positive linear correlation. The correlation coefficient is 0.16, indicating that the linear correlation is very weak. Therefore, you might conclude that starting another youth orchestra will not help the development of young musicians. However, notice that the data points seem to form two clusters in the scatter plot, one on the left side and the other on the right. This unusual pattern suggests the presence of a hidden variable, which could affect your analysis. You will need more information to determine the nature and effect of the possible hidden variable.



You have enough data to produce a time-series graph of the numbers of young musicians who go on to professional orchestras. This graph also has two clusters of data points. The numbers rise from 1991 to 1996, drop substantially in 1997, and then rise again. This pattern suggests that something unusual happened in 1997.



b) The collapse of a major orchestra means both that there is one less orchestra hiring young musicians and that about a hundred experienced players are suddenly available for work with the remaining professional orchestras. The resulting drop in the number of young musicians hired by professional orchestras could account for the clustering of data points you observed in part a). Because of the change in the number of jobs available for young musicians, it makes sense to analyse the clusters separately.



Observe that the two sets of data both exhibit a strong linear correlation. The correlation coefficients are 0.93 for the data prior to 1997 and 0.94 for the data from 1997 on. The number of players who go on to professional orchestras is strongly correlated to the number of youth orchestras. So, funding the new orchestra may be a worthwhile project for the arts council.

The presence of a hidden variable, the collapse of a major orchestra, distorted the data and masked the underlying pattern. However, splitting the data into two sets results in smaller sample sizes, so you still have to be cautious about drawing conclusions.

When evaluating claims based on statistical studies, you must assess the methods used for collecting and analysing the data. Some critical questions are:

- Is the sampling process free from intentional and unintentional bias?
- Could any outliers or extraneous variables influence the results?
- Are there any unusual patterns that suggest the presence of a hidden variable?
- Has causality been inferred with only correlational evidence?

Project Prep

When collecting and analysing data for your statistics project, you can apply the concepts in this section to ensure that your conclusions are valid.

Key Concepts

- Although the major media are usually responsible in how they present statistics, you should be cautious about accepting any claim that does not include information about the sampling technique and analytical methods used.
- Intentional or unintentional bias can invalidate statistical claims.
- Small sample sizes and inappropriate sampling techniques can distort the data and lead to erroneous conclusions.
- Extraneous variables must be eliminated or accounted for.
- A hidden variable can skew statistical results and yet still be hard to detect.

Communicate Your Understanding

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

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.

3. Give an example of a hidden variable not mentioned in this section, and explain why this variable would be hard to detect.

Apply, Solve, Communicate

A

1. An educational researcher discovers that levels of mathematics anxiety are negatively correlated with attendance in mathematics class. The researcher theorizes that poor attendance causes mathematics anxiety. Suggest an alternate interpretation of the evidence.

2. A survey finds a correlation between the proportion of high school students who own a car and the students' ages. What hidden variable could affect this study?

B

3. A student compares height and grade average with four friends and collects the following data.

Height (cm)	Grade Average (%)
171	73
145	91
162	70
159	81
178	68

From this table, the student concludes that taller students tend to get lower marks.

a) Does a regression analysis support the student's conclusion?

b) Why are the results of this analysis invalid?

c) How can the student get more accurate results?

4. **Inquiry/Problem Solving** A restaurant chain randomly surveys its customers several times a year. Since the surveys show that the level of customer satisfaction is rising over time, the company concludes that its customer service is improving. Discuss the validity of the surveys and the conclusion based on these surveys.

5. **Application** A teacher offers the following data to show that good attendance is important.

Days Absent	Final Grade
8	72
2	75
0	82
11	68
15	66
20	30

A student with a graphing calculator points out that the data indicate that anyone who misses 17 days or more is in danger of failing the course.

a) Show how the student arrived at this conclusion.

b) Identify and explain the problems that make this conclusion invalid.

c) Outline statistical methods to avoid these problems.

6. Using a graphing calculator, Gina found the cubic curve of best fit for the salary data in the table on page 157. This curve has a coefficient of determination of 0.98, indicating an almost perfect fit to the data. The equation of the cubic curve is

$$\text{starting salary} = 0.0518y^3 - 310y^2 + 618\,412y - 411\,344\,091$$

where the salary is given in thousands of dollars and y is the year of graduation.

a) What mean starting salary does this model predict for Gina's class when they graduate in 2005?

b) Is this prediction realistic? Explain.

c) Explain why this model generated such an inaccurate prediction despite having a high value for the coefficient of determination.

d) Suggest methods Gina could use to make a more accurate prediction.

7. **Communication** Find a newspaper or magazine article, television commercial, or web page that misuses statistics of two variables. Perform a critical analysis using the techniques in this chapter. Present your findings in a brief report.

8. **Application** A manufacturing company keeps records of its overall annual production and its number of employees. Data for a ten-year period are shown below.

Year	Number of Employees	Production (000)
1992	158	75
1993	165	81
1994	172	84
1995	148	68
1996	130	58
1997	120	51
1998	98	50
1999	105	57
2000	110	62
2001	120	70

a) Create a scatter plot to see if there is a linear correlation between annual production and number of employees. Classify the correlation.

b) At some point, the company began to lay off workers. When did these layoffs begin?

c) Does the scatter plot suggest the presence of a hidden variable? Could the layoffs account for the pattern you see? Explain why or why not.

d) The company's productivity is its annual production divided by the number of

employees. Create a time-series graph for the company's productivity.

e) Find the line of best fit for the graph in part d).

f) The company has adopted a better management system. When do you think the new system was implemented? Explain your reasoning.

C

9. Search E-STAT, CANSIM II, or other sources for time-series data for the price of a commodity such as gasoline, coffee, or computer memory. Analyse the data and comment on any evidence of a hidden variable. Conduct further research to determine if there are any hidden variables. Write a brief report outlining your analysis and conclusions.

10. **Inquiry/Problem Solving** A study conducted by Stanford University found that behavioural counselling for people who had suffered a heart attack reduced the risk of a further heart attack by 45%. Outline how you would design such a study. List the independent and dependent variables you would use and describe how you would account for any extraneous variables.

Career Connection

Economist

Economists apply statistical methods to develop mathematical models of the production and distribution of wealth. Governments, large businesses, and consulting firms are employers of economists. Some of the functions performed by an economist include

- recognizing and interpreting domestic and international market trends
- using supply and demand analysis to assess market potential and set prices
- identifying factors that affect economic growth, such as inflation and unemployment
- advising governments on fiscal and monetary policies
- optimizing the economic activity of financial institutions and large businesses

Typically, a bachelor's degree in economics is necessary to enter this field. However, many positions require a master's or doctorate degree or specialized training. Since economists often deal with large amounts of data, a strong background in statistics and an ability to work with computers are definite assets.

An economist can expect to earn a comfortable living. Most employment opportunities for economists are in large cities. The current demand for economists is reasonably strong and likely to remain so for the foreseeable future, as governments and large businesses will continue to need the information and analysis that economists provide.

WEB CONNECTION

www.mcgrawhill.ca/links/MDM12

Visit the above web site and follow the links to learn more about a career as an economist and other related careers.