

# **INFERENCE:**

## *Connecting the Question and the Solution*

### **Resources for AP Statistics Teachers from the NCSSM Statistics Leadership Institute 2001**

This packet contains

- Sample problems and solutions
- Notes for teachers
- Comments on the AP exam problems

**Bonnie Montgomery**  
Morris Knolls High School  
Rockaway, New Jersey  
[mididim@msn.com](mailto:mididim@msn.com)

**Mary Ellen Noyes**  
Mary Institute and Saint Louis Country Day School  
Saint Louis, Missouri  
[mnoyes@micds.org](mailto:mnoyes@micds.org)

**Margaret Wirth**  
J. H. Rose High School  
Greenville, North Carolina  
[piwirth@greenvillenc.com](mailto:piwirth@greenvillenc.com)

## Table of Contents

### Part 1: Problems for Inference

Introduction.....	page 3
Worksheet.....	page 4
Solutions and Notes.....	page 14

### Part 2: Scenarios with Multiple Inference Problems

Introduction.....	page 37
<u>Scenario 1</u>	
Worksheet.....	page 38
Solutions and Notes.....	page 39
<u>Scenario 2</u>	
Worksheet.....	page 41
Solutions and Notes.....	page 43

### Part 3: Inference Set-up Problems

Introduction.....	page 53
Worksheet.....	page 54
Solutions and Notes.....	page 56

### Part 4: Notes on Inference on the 2001 AP Exam.....page 63

## **PART 1**

### **PROBLEMS FOR INFERENCE**

#### Introduction

The impetus for this project was the large number of students who failed to correctly recognize, classify, and carry out the inference problems on the AP exam. Many times in textbook problems, students are told specifically which test is needed or students know which test to use simply by which chapter they are currently studying. AP test problems do not tell directly either which test to perform or if a significance test is even required. The group of problems that follow is an assortment of examples in inference which cover all the inference procedures on the AP syllabus. All of the problems were either taken from or adapted from various textbooks. We have selected problems which through the wording or the data presentation may not be as straightforward as those to which students are accustomed. Following the worksheet you will find, for each problem, the name of the test to be used, the null and alternative hypotheses and a set of notes about the problem. Teachers may find the worksheet most useful as a review tool prior to the AP exam.

#### Note on Solutions

We have included checks on the required conditions for each inference procedure. In the case of two sided tests, if students choose to use confidence intervals they must still include the hypotheses.

**PART 1**  
**PROBLEMS FOR INFERENCE**

**Worksheet**

1. From a sample of 13,912 households in Washtenaw County, Michigan, 2,993 persons were identified as being 60 years or older. Of these, 1,956 agreed to participate in a study. The following table is based on data from the married women in the sample who answered the questions about whether they were sexually active and whether they drank coffee. (Data from A. C. Dionkno et al., "Sexual Function in the Elderly," *Archives of Internal Medicine* 150 (1990): 197-200.)

Sexually Active		
	Yes	No
Coffee Drinker	15	25
Not coffee drinker	115	70

Is there a relationship between coffee drinking and sexual activity for married women 60 or older?

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

2. In biology laboratory students test the Mendelian theory of inheritance using corn. The Mendelian theory of inheritance claims that frequencies of the four categories smooth and yellow, wrinkled and yellow, smooth and purple, and wrinkled and purple will occur in the ratio 9:3:3:1. If a student counted 124, 30, 43, and 11, respectively, for these four categories, are these data compatible with the Mendelian theory?

(Problem adapted from *Probability And Statistical Inference*, 5<sup>th</sup> ed., R. V. Hogg and E. A. Tanis, Saddle River, NJ: Prentice Hall, 1997.)

3. One of the most dangerous contaminants deposited over European countries following the Chernobyl accident of April 1987 was radioactive cesium. To study cesium transfer from contaminated soil to plants, researchers collected soil samples and samples of mushroom mycelia from 17 wooded locations in Umbria, Central Italy, from August 1986 to November 1989. Measured concentrations (Bq/kg) of cesium in the soil and in the mushrooms are listed below. (Data from R. Borio et al., "Uptake of Radiocesium by Mushrooms," Science of the Total Environment 106 (1991): 183-90.) Although the soil concentration of cesium is of interest, it is expensive and difficult to measure. Thus the cesium in the mushroom samples will be used to determine the cesium content in the soil.
- Construct a scatter plot of soil concentration versus mushroom concentration.
  - Fit a simple linear regression model relating soil concentration of cesium to mushroom concentration of cesium.
  - Do the data suggest there is a linear relationship between contaminated mushroom concentration of cesium and soil concentration of cesium?
  - Is the model useful in predicting the amount of soil contamination based on the uptake of radiocesium by mushrooms?

Sample	Cesium in Mushrooms (L2)	Cesium in Soil (L3)
1	1	33
2	9	55
3	14	138
4	17	319
5	20	415
6	17	425
7	14	442
8	15	475
9	34	279
10	41	329
11	46	82
12	49	86
13	53	55
14	60	60
15	79	144
16	99	292
17	190	1310

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

4. Students in a statistics class measured the lengths of their right and left feet to the nearest millimeter. The right and left foot measurements were equal for 103 of the 215 students, but the two foot lengths were different for 112 students. Assuming this class is representative of all college students, is there evidence that the proportion of college students whose feet are the same length differs from one-half?

(Problem from *Mind on Statistics*, J. M. Utts and R. F. Heckard, Pacific Grove, CA: Duxbury, 2002.)

5. Cyclozocine was an alternative to methadone for treating heroin addiction. The following data came from 14 males who were chronic heroin addicts. After cyclozocine had removed the addicts' physical dependence on heroin, they were asked a battery of questions designed to assess their psychological dependence. The test scores are called Q-scores, and high values represent less psychological dependence. The Q-scores were as follows (Resnick et al., 1970):

51	53	43	36	55	55	39
43	45	27	21	26	22	43

From past experience, the mean score for addicts who have not had a cyclozocine treatment is about 28. Is cyclozocine an effective treatment for psychological dependence?

(Problem adapted from *Chance Encounters*, C. J. Wild and G. A. F. Seber, New York: John Wiley & Sons, 2000.)

6. A Gallup Poll taken in May 2000 asked the question: "In general, do you feel that the laws covering the sale of firearms should be made: more strict, less strict, or kept as they are now?" Of the  $n = 493$  men who responded, 52% said "more strict," while of the  $n = 538$  women who responded, 72% said "more strict." Assuming these respondents constitute random samples of U.S. men and women, is there sufficient evidence to

conclude that a higher proportion of women than men in the population think these laws should be made more strict? Justify your answer.

(Problem taken from *Mind on Statistics*, J. M. Utts and R. F. Heckard, Pacific Grove, CA: Duxbury, 2002.)

7. It is believed that regular physical exercise leads to a lower resting pulse rate. Following are data for  $n=20$  individuals on resting pulse rate and whether the individuals regularly exercise or not. Assuming this is a random sample from a larger population, use this sample to determine whether the mean pulse is lower for those who exercise.

Person	Pulse	Regularly Exercises
1	72	No
2	62	Yes
3	72	Yes
4	84	No
5	60	Yes
6	63	Yes
7	66	No
8	72	No
9	75	Yes
10	64	Yes
11	62	No
12	84	No
13	76	No
14	60	Yes
15	52	Yes
16	60	No
17	64	Yes
18	80	Yes
19	68	Yes
20	64	Yes

(Problem taken from *Mind on Statistics*, J. M. Utts and R. F. Heckard, Pacific Grove, CA: Duxbury, 2002.)

8. The study “Evaluation of Prescribed Burning in Relation to Available Deer Browse” was undertaken at the Virginia Polytechnic Institute and State University in 1964 to determine if fire can be used as a viable management tool to increase the amount of forage available to deer during the crucial months in late winter and spring. Calcium is a required element for plants and animals. The amount taken up and stored in the plant is closely correlated to the amount present in the soil. It was hypothesized that a fire may change the calcium levels present in the soil and thus affect the amount available to the deer. A large tract of land in the Fishburn Forest was selected for a prescribed burn. Soil samples were taken from 12 plots of equal area just prior to the burn on May 20, 1964, and analyzed for calcium. On July 16, 1964, postburn calcium levels were analyzed from the same plots. These values, in kilograms per plot, are presented in the following table:

Calcium Level  
( kg/plot)

Plot	Preburn	Postburn
1	50	9
2	50	18
3	82	45
4	64	18
5	82	18
6	73	9
7	77	32
8	54	9
9	23	18
10	45	9
11	36	9
12	54	9

Do these data support the hypothesis?

(Problem adapted from *Probability And Statistics for Engineers and Scientists*, 4<sup>th</sup> ed., R. E. Walpole and R. H. Myers, New York: Macmillan Publishing Company, 1972.)

9. The article “Effect of Temperature on the pH of Skim Milk” (J. of Dairy Research (1988): 277-280) reported on a study involving temperature ( $^{\circ}C$ ) under specified experimental conditions and milk pH. The accompanying data set is a representative subset of that which appeared in the article:

Temp	pH
4	6.85
4	6.79
24	6.63
24	6.65
25	6.72
38	6.62
38	6.57
40	6.52
45	6.50
50	6.48
55	6.42
56	6.41
60	6.38
67	6.34
70	6.32
78	6.34

Do these data strongly suggest that there is a negative (inverse) linear relationship between temperature and pH? Provide statistical justification for your conclusions.

(Problem adapted from *Introduction to Statistics and Data Analysis*; R. Peck, C. Olsen, and J. Devore; Pacific Grove, CA: Duxbury, 2001.)

10. It is generally believed that red-cockaded woodpeckers require or preferentially select old-age pine trees and stands for constructing nest and roost cavities. This hypothesis is important since the need for suitable habitat for nesting and roosting is vital to the continued existence of the species. The information for this exercise came from part of a study conducted by DeLotelle and Epting [1988]. Trees were sampled from stands of trees occupied by colonies of woodpeckers. Trees with current or abandoned woodpecker holes were termed *cavity trees*. Untouched trees in a defined neighborhood of cavity trees were called *colony trees*. The ages (in years) of cavity trees and colony trees were compared, giving the following summary statistics:

	<b>Cavity Trees</b>	<b>Colony Trees</b>
<b>Sample Size</b>	54	143
<b>Sample Mean</b>	104.1	83.6
<b>Sample Standard Deviation</b>	24.1	38.3

Is there evidence that cavity trees are older than colony trees on average?

(Problem adapted from *Chance Encounters*, C. J. Wild and G. A. F. Seber, New York: John Wiley & Sons, Inc., 2000.)

11. Researchers examined 103 adolescent males who had been diagnosed as hyperactive as children and identified those with drug abuse problems and those with antisocial personalities. They then examined 100 adolescent males who had not been diagnosed as hyperactive as children, using the same procedure. The data are given below.

---

*Adolescent problems of males diagnosed as hyperactive (n = 103) or not hyperactive (n = 100) as children. Numbers are percents of samples.*

---

	HYPERACTIVE	NOT HYPERACTIVE
Drug abuse problem	16	4
Antisocial personality	27	8

---

- a. Comparing the males who were and were not hyperactive as children, is there a significant difference in the percentages of those with drug abuse problems?
- b. Comparing the males who were and were not hyperactive as children, is there a significant difference in the percentages of those with an antisocial personality?

(Problem adapted from *Statistics and Data Analysis; An Introduction*, 2<sup>nd</sup> ed., A. F. Siegel and C. J. Morgan, New York: John Wiley & Sons, Inc., 1996.)

12. On September 16, 1992, *USA Today* reported the results of a study which links female sexuality to genes. The study, conducted by researchers at Boston University, involved 115 homosexual or bisexual women who had twin sisters. Seventy-seven of the women had identical twins, who share all of the same genes, and 38 of the women had fraternal twins, who are related from a genetic standpoint just like other siblings. The researchers determined the sexual preference of the twin sisters. The results are as follows:

<b>Sexual Preference of Twin Sisters</b>		
<b>Type of Twin</b>	<b>Number Surveyed</b>	<b># Homosexual or Bisexual</b>
<b>Identical</b>	77	39
<b>Fraternal</b>	38	6

If the proportion of identical twins who are also homosexual is significantly larger than the proportion of fraternal twins who are also homosexual, the researchers will conclude that homosexual preference is linked to genes. Based on the data, is there evidence that homosexual preference is genetically linked?

(Problem adapted from *Discovering Statistics; An Adventure in Problem Solving*, J. S. Hawkes, Charleston: Quant Publishing, 1995.)

13. Charles Darwin carried out an experiment to study whether seedlings from cross-fertilized plants tend to be superior to those from self-fertilized plants. He covered a number of plants with fine netting so that insects would be unable to fertilize them. He fertilized a number of flowers on each plant with their own pollen and he fertilized an equal number of flowers on the same plant with pollen from a distant plant. (He did not say how he decided which flowers received which treatments.) The seeds from the flowers were allowed to ripen and were set in wet sand to germinate. He placed two seedlings of the same age in a pot, one from a seed from a self-fertilized flower and one from a seed from a cross-fertilized flower. (The fertilization experiments were described by Darwin in an 1878 book; these data were found in D. F. Andrews and A. M. Herzberg, *Data*, New York: Springer-Verlag, 1985, pp. 9-12.) Below are the heights of the plants at a particular age.

---

*Darwin's data: plant heights (inches) for 15 pairs of plants of the same age, one of which was grown from a seed from a cross-fertilized flower and the other of which was grown from a seed from a self-fertilized flower*

---

Plant Heights (inches)		
Pair	Cross-fertilized	Self-fertilized
1	23.5	17.375
2	12	20.375
3	21	20
4	22	20
5	19.125	18.375
6	21.5	18.625
7	22.125	18.625
8	20.375	15.25
9	18.25	16.5
10	21.625	18
11	23.25	16.25
12	21	18
13	22.125	12.75
14	23	15.5
15	12	18

---

If we consider taller plants “superior” to shorter plants, do these data provide evidence that cross-fertilization produces superior plants to self-fertilization? Support your answer with the appropriate statistical justification.

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

## PART 1

### PROBLEMS FOR INFERENCE

#### Solutions and Notes

1. From a sample of 13,912 households in Washtenaw County, Michigan, 2,993 persons were identified as being 60 years or older. Of these, 1,956 agreed to participate in a study. The following table is based on data from the married women in the sample who answered the questions about whether they were sexually active and whether they drank coffee. (Data from A. C. Dionkno et al., "Sexual Function in the Elderly," *Archives of Internal Medicine* 150 (1990): 197-200.)

	Sexually Active	
	Yes	No
Coffee Drinker	15	25
Not coffee drinker	115	70

Is there a relationship between coffee drinking and sexual activity for married women 60 or older?

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

#### Solution

**Procedure type:** *Chi-square Test for Independence*

$H_0$ : *Coffee drinking and sexual activity are not associated for married women 60 or older*

$H_a$ : *Coffee drinking and sexual activity are associated for married women 60 or older*

#### **Notes**

*The key to the solution is recognizing the question is about the relationship between two categorical variables. When we see large differences in conditional distributions within*

rows of a contingency table we suspect the variables are related and the relationship between categorical variables is real. The statistical significance of the association between the categorical variables, sexual activity and coffee drinking, can be examined using the chi-square statistic. The chi-square test for a two-way table is a large sample test requiring  $n$ , the total table count, to be sufficiently large. The conditions for performing the chi-square test address the largeness issue. The rules on the definition of “sufficiently large” vary according to the text. A commonly used rule is that the expected counts should be greater than 1 and 80% of the expected counts should be at least 5. We will also have to assume that the sample is representative of the population. The two hypotheses for the significance test should always be stated in the context of the two variables.

Notice that a single sample was taken and then each observation was categorized by two variables. This structure of the data is recognized in the chi-square test for independence. In contrast, a comparison of two proportions ( e.g. “Is the proportion of sexually active women 60 or older the same for coffee drinkers and non-coffee drinkers?”) would be a test for homogeneity. A test comparing two means would be inappropriate because the data are categorical.

2. In biology laboratory students test the Mendelian theory of inheritance using corn. The Mendelian theory of inheritance claims that frequencies of the four categories smooth and yellow, wrinkled and yellow, smooth and purple, and wrinkled and purple will occur in the ratio 9:3:3:1. If a student counted 124, 30, 43, and 11, respectively, for these four categories, are these data compatible with the Mendelian theory?

(Problem adapted from *Probability And Statistical Inference*, 5<sup>th</sup> ed., R. V. Hogg and E. A. Tanis, Saddle River, NJ: Prentice Hall, 1997.)

### Solution

**Procedure type:** Chi-square goodness of fit test

$H_o$  : The data are compatible with the Mendelian theory and occurs in ratio 9:3:3:1

$H_a$  : The data are not compatible with the Mendelian theory and do not occur in ratio 9:3:3:1

### Notes

The problem is asking if the counts of 124, 30, 43, and 11 are occurring in the ratio 9:3:3:1. Students need to recognize that 124, 30, 43 and 11 are count data. Count data are

*categorical and for categorical data we use either a test for proportions or a Chi-Square test. Because the problem is looking at more than two groups a Chi-Square test must be used. Suggest to students that they create a table to enter observed and expected values. Students must first add their counts:  $124+30+43+11=108$  to find  $n$  and then divide the total into the ratio 9:3:3:1 to find the expected values. Those values would be 117, 39, 39, 13.*

*Students should check the necessary condition of the Chi-square by checking the expected values to make sure they are at least 5. From what is written above, this is true.*

*Students sometimes incorrectly state the degrees of freedom. One mistake students may make is to state the degrees of freedom as  $n-1$  or 107 in this case. Another mistake students may make is to use the degree of freedom formula for a Chi-Square test of independence; that is,  $(\text{rows} - 1) \cdot (\text{columns} - 1)$ . The Chi-Square goodness-of-fit test has a degree of freedom of  $(\text{number of columns}) - 1$ . The degrees of freedom for this problem is 3.*

*Students may incorrectly use the number 9, 3, 3, 1 as additional count data and proceed by using a 2 by 4 table. Students should be reminded that 9:3:3:1 is simply an expected ratio for the original count data.*

3. One of the most dangerous contaminants deposited over European countries following the Chernobyl accident of April 1987 was radioactive cesium. To study cesium transfer from contaminated soil to plants, researchers collected soil samples and samples of mushroom mycelia from 17 wooded locations in Umbria, Central Italy, from August 1986 to November 1989. Measured concentrations (Bq/kg) of cesium in the soil and in the mushrooms are listed below. (Data from R. Borio et al., "Uptake of Radiocesium by Mushrooms," Science of the Total Environment 106 (1991): 183-90.) Although the soil concentration of cesium is of interest, it is expensive and difficult to measure. Thus the cesium in the mushroom samples will be used to determine the cesium content in the soil.
  - a. Construct a scatter plot of soil concentration versus mushroom concentration.
  - b. Fit a simple linear regression model relating soil concentration of cesium to mushroom concentration of cesium.
  - c. Do the data suggest there is a linear relationship between contaminated mushroom concentration of cesium and soil concentration of cesium?
  - d. Is the model useful in predicting the amount of soil contamination based on the uptake of radiocesium by mushrooms?

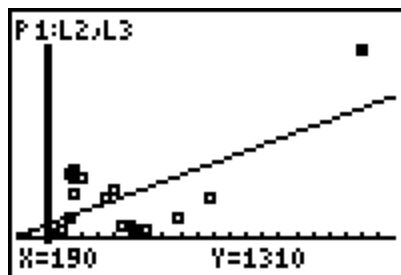
Sample	Cesium in Mushrooms (L2)	Cesium in Soil (L3)
1	1	33
2	9	55
3	14	138
4	17	319
5	20	415
6	17	425
7	14	442
8	15	475
9	34	279
10	41	329
11	46	82
12	49	86
13	53	55
14	60	60
15	79	144
16	99	292
17	190	1310

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

**Solution**

**Procedure types:** Scatterplot, Linear Regression, Linear Regression t-Test

- a) Scatter plot of mushroom concentrations of cesium (X), soil concentrations of cesium(Y)



b) Linear Regression

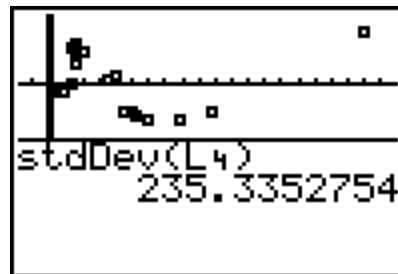
```
LinReg
y=a+bx
a=101.5874149
b=4.237485418
r2=.4063862044
r=.6374842778
```

```
Plot1 Plot2 Plot3
\Y1=4.2374854177
915X+101.5874149
0083
\Y2=
\Y3=
\Y4=
\Y5=
```

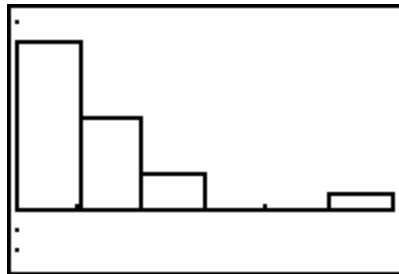
The residuals are in L4

Plot of the residuals

L2	L3	L4	4
1	33	-72.82	
9	55	-84.72	
14	138	-22.91	
17	319	145.38	
20	415	228.66	
17	425	251.38	
14	442	281.09	
L4 = LRESID			



Histogram of the residuals



**Notes**

L2 displays the amount of cesium in the mushroom samples, L3 displays the amount of cesium in the soil samples, and L4 is the list of residuals based on the linear regression model. The regression model is designed to use the amount of cesium in the mushrooms to determine the amount of cesium in the soil, since collecting the mushroom samples is assumed to be more economical and convenient than collecting soil samples. The standard deviation of the residuals (235), which appears large for the data set, is also displayed. A histogram of the residuals is skewed and indicates the point (190, 1310) may be an outlier.

c)

$H_0: b_1 = 0$  ( the population slope is 0, so cesium in soil samples and cesium in mushrooms are not linearly related)

$H_1: b_1 \neq 0$  (the population slope is not 0, so cesium in soil samples and cesium in mushrooms are linearly related)

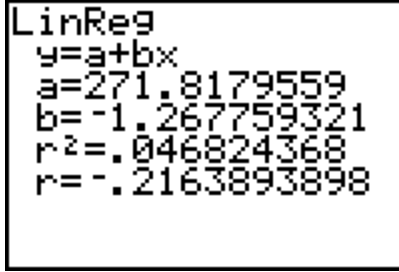
```
LinRegTTest
y=a+bx
b≠0 and ρ≠0
t=3.204520949
P=.0059088992
df=15
↓a=101.5874149
█
```

### Notes

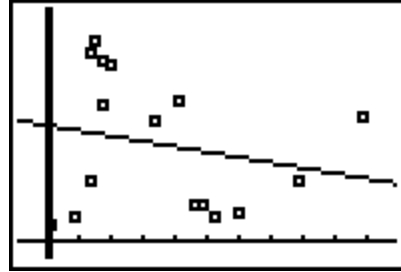
The significance test for the slope, the linear regression t-test, is used to determine if the linear model is indeed useful. The significance test of whether or not the population slope,  $b_1$ , is 0 tells us if the linear relationship between the soil samples and the mushrooms is statistically significant. The p-value of 0.0059 gives us statistical evidence that the two variables are related and the model could be used to conclude that a linear relationship exists in the population represented by this sample. The scatterplot and linear regression are easily accomplished on either a calculator or a computer. In observing the output from the regression, the  $r=0.637$ ,  $r^2 = 0.406$ , and the residual plot offer some evidence for using this linear model in the prediction process. The fact that only 41% of the variability in the soil concentration is explained by the least squares regression model, the large standard error for the residuals, and the evidence of an outlier in the scatter plot encourage further investigation.

d) To determine if the model would be a useful predictor of the amount of cesium in the soil based on the amount of cesium in the mushrooms involves looking at all of the results from the previous questions. The scatter plot, the regression line, the residuals, the residual plot,  $r$ ,  $r^2$ , the standard deviation of the residuals and the inference for slope t-test are all viewed together to make a decision on the appropriateness of the model for prediction purposes. The standard deviation of the residuals (235) shows that the prediction errors are likely to be large. Together these indicators would suggest that while there is evidence of a linear relationship, this linear model is not particularly useful for making predictions. One additional approach to determining the value of the linear model for prediction is to explore the effect of the potentially influential value (190, 1310) evidenced in the scatter plot. We may wish to eliminate the influential point (190, 1310) and investigate this further by performing the regression analysis again.

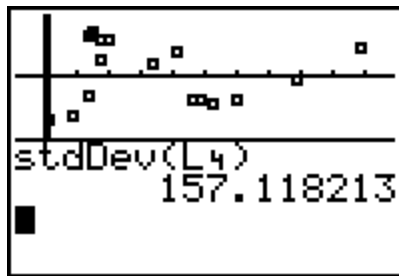
Regression w/o (190,1310)



Scatter plot w/o (190,1310)

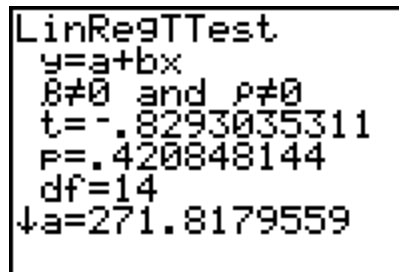


Residuals plot



$H_0 : b_1 = 0$  ( the population slope is 0, so cesium in soil samples and cesium in mushrooms are not linearly related)

$H_1 : b_1 \neq 0$  (the population slope is not 0, so cesium in soil samples and cesium in mushrooms are linearly related)



The change in  $r$  from 0.637 to -0.216, a major change in strength and direction of the correlation, and the change in the coefficient of determination,  $r^2$ , from 0.406 to 0.0468, a dramatic drop, both clearly demonstrate the strong effect this influential point has on the fitted model. The standard error of the residuals also drops from 235 to 157. And the  $p$ -value of 0.42 in the inference for slope  $t$ -test suggests that there is not a linear relationship between cesium in the soil samples and cesium in the mushrooms. Thus, the influence of this one point on the model is clearly demonstrated in the new statistical results. Using all of this comparative information with and without the point reinforces our suspicions that our

original linear model, which is so obviously influenced by one point, is not a wise choice for predictions.

4. Students in a statistics class measured the lengths of their right and left feet to the nearest millimeter. The right and left foot measurements were equal for 103 of the 215 students, but the two foot lengths were different for 112 students. Assuming this class is representative of all college students, is there evidence that the proportion of college students whose feet are the same length differs from one half?

(Problem from *Mind on Statistics*, J. M. Utts and R. F. Heckard, Pacific Grove, CA: Duxbury, 2002.)

### Solution

**Procedure type:** Large sample hypothesis test for proportions

$$\begin{aligned} H_o : p &= .5 \\ H_a : p &\neq .5 \end{aligned} \quad \begin{array}{l} \text{where } p \text{ represents the proportion of college students whose feet are the} \\ \text{same length} \end{array}$$

### **Notes**

This problem is solved by using a large sample hypothesis test for proportion. Students should note that the type of data in the problem is categorical data. We are looking to see if the proportion of college students whose feet are the same length differs from  $\frac{1}{2}$ . This statement tells you that it is a two-tailed test.

Students could solve this problem using a confidence interval and look for 0 in the interval.

Students should remember to check the necessary conditions, which are that we have a simple random sample from the population and that  $n \cdot p \geq 10$  and  $n \cdot (1 - p) \geq 10$ . The value of  $p$  should be the value in the null hypothesis, which in this case is 0.5. Many students mistakenly use the value of  $\hat{p}$ , which is  $\frac{103}{215} = 0.479$ . Note that some texts use a less conservative assumption, which is to replace 10 with 5. One can see that the necessary condition of large sample size is satisfied.

Clearly this problem is not a simple random sample, but we will treat it as if it is since the problem states, "Assuming this class is representative of all college students."

Students need to realize that they do not need to use all the numbers given. If 103 students out of 215 students have feet equal in length, then it follows that  $215-103=112$  students have left and right feet of two different measurements. Remind students to clearly identify the population parameter in the null and alternative hypotheses so that the reader knows what is being tested. Caution students that more care is needed if the test is one-tailed. Teachers may want to try the problem again using a one-tailed scenario.

5. Cyclozocine was an alternative to methadone for treating heroin addiction. The following data came from 14 males who were chronic heroin addicts. After cyclozocine had removed the addicts' physical dependence on heroin, they were asked a battery of questions designed to assess their psychological dependence. The test scores are called Q-scores, and high values represent less psychological dependence. The Q-scores were as follows (Resnick et al. (1970)):

51	53	43	36	55	55	39
43	45	27	21	26	22	43

From past experience, the mean score for addicts who have not had a cyclozocine treatment is about 28. Is cyclozocine an effective treatment for psychological dependence?

(Problem adapted from *Chance Encounters*, C. J. Wild and G. A. F. Seber, New York: John Wiley & Sons, 2000.)

### Solution

**Procedure type:** One-sided, one-sample *t*-test

Let  $\mathbf{m}$  = the mean Q-score for chronic heroin addicts who have had a cyclozocine treatment.

$H_0 : \mathbf{m} = 28$       The mean Q-score for chronic heroin addicts who have had a cyclozocine treatment is less than or equal to 28.

$H_1 : \mathbf{m} > 28$       The mean Q-score for chronic heroin addicts who have had a cyclozocine treatment is greater than 28.

### Notes

The comparison of the effectiveness of the cyclozocine treatment on chronic heroin addicts is evaluated by comparing their mean  $Q$ -score to the standard  $Q$ -score of 28. An effective treatment would have a mean  $Q$ -score sufficiently greater than 28 to provide evidence against the null hypothesis and confirm the alternative hypothesis. To perform the  $t$ -test for significance we must meet certain conditions. To assess if the population is normal we use the small sample (14 participants) distribution as an indicator. It appears approximately normal when viewed as a boxplot, with no outliers. Although we are not told this is a simple random sample, we hope it is representative of the population of chronic heroine addicts. The standard deviation of the population is unknown. Thus the  $t$ -test is an appropriate choice for a test of significance.

6. A Gallup Poll taken in May 2000 asked the question: "In general, do you feel that the laws covering the sale of firearms should be made: more strict, less strict, or kept as they are now?" Of the  $n = 493$  men who responded, 52% said "more strict," while of the  $n = 538$  women who responded, 72% said "more strict." Assuming these respondents constitute random samples of U.S. men and women, is there sufficient evidence to conclude that a higher proportion of women than men in the population think these laws should be made more strict? Justify your answer.

(Problem taken from *Mind on Statistics*, J. M. Utts and R. F. Heckard, Pacific Grove, CA: Duxbury, 2002.)

### Solution

**Procedure type:** Difference of two proportions  $z$  test

$$\begin{array}{l} H_o : p_w - p_m = 0 \\ H_a : p_w - p_m > 0 \end{array} \quad \text{OR} \quad \begin{array}{l} H_o : p_w = p_m \\ H_a : p_w > p_m \end{array}$$

where  $p_m$  and  $p_w$  represent the proportion of men and women respectively who support "more strict" laws in the sale of firearms.

### Notes

This problem is solved by performing a large-sample difference of two proportions test. This is evident because there are two populations that we are studying, men and women. The problem also asks students to find evidence of a higher percentage or proportion of women than men who think that the sale of firearms should be more strict, indicating a one-tailed test. Students may need to be reminded that for the hypothesis test, in calculating the test

statistic, students must use a pooled estimate for the proportion. In a hypothesis test we are assuming the null hypothesis is true, and the null hypothesis assumes population proportions for men and women are equal.

Students need to check the conditions that the sample size is large enough. One rule for checking this would be:  $n_w \cdot p_w \geq 10$ ,  $n_w \cdot (1 - p_w) \geq 10$  and  $n_m \cdot p_m \geq 10$ ,  $n_m \cdot (1 - p_m) \geq 10$  where  $n_m$  and  $n_w$  represent the number of men and women respectively. Some texts use:  $n_w \cdot \hat{p} \geq 10$ ,  $n_w \cdot (1 - \hat{p}) \geq 10$  and  $n_m \cdot \hat{p} \geq 10$ ,  $n_m \cdot (1 - \hat{p}) \geq 10$  where  $\hat{p}$  represents the pooled estimate of  $p$ . Other texts use 5 in place of 10.

Possible incorrect solutions would include difference of two means or Chi-Square. Difference of means cannot be correct because we have proportions as opposed to average percentages. If students are confused on this issue, ask them what the original data must look like. Is it numerical or categorical? For answers of categorical, tests of proportion are correct. For answers of numerical,  $t$ -tests are appropriate. In this case the raw data must be in the form, "more strict," "less strict," or "kept as they are now." This is categorical data. If our original data were presented as a list of percentages for many different samples (which is numerical), a  $t$ -test would be used. Students could also attempt a solution using Chi-Square. Since Chi-Square is always two-tailed and we are doing a one-tailed test this would not be appropriate. Chi-Square could work as an alternative solution if this example was not one-tailed.

7. It is believed that regular physical exercise leads to a lower resting pulse rate. Following are data for  $n = 20$  individuals on resting pulse rate and whether the individuals regularly exercise or not. Assuming this is a random sample from a larger population, use this sample to determine whether the mean pulse is lower for those who exercise.

Person	Pulse	Regularly Exercises
1	72	No
2	62	Yes
3	72	Yes
4	84	No
5	60	Yes
6	63	Yes
7	66	No
8	72	No
9	75	Yes
10	64	Yes
11	62	No
12	84	No
13	76	No
14	60	Yes
15	52	Yes
16	60	No
17	64	Yes
18	80	Yes
19	68	Yes
20	64	Yes

(Problem taken from *Mind on Statistics*, J. M. Utts and R. F. Heckard, Pacific Grove, CA: Duxbury, 2002.)

**Solution**

*Procedure type: Two sample t-test*

$$\begin{array}{l}
 H_o : \mathbf{m}_{yes} - \mathbf{m}_{no} = 0 \\
 H_a : \mathbf{m}_{yes} - \mathbf{m}_{no} < 0
 \end{array}
 \quad
 \text{OR}
 \quad
 \begin{array}{l}
 H_o : \mathbf{m}_{yes} = \mathbf{m}_{no} \\
 H_a : \mathbf{m}_{yes} < \mathbf{m}_{no}
 \end{array}$$

where  $\mathbf{m}_{yes}$  represents the mean pulse rate of those who exercise and  $\mathbf{m}_{no}$  represents the mean pulse rate of those who do not exercise

**Notes**

The problem asks to test whether there is a difference between the **mean** pulse rate of subjects who exercise as opposed to those who do not. This indicates a **two sample t-test**. It is a two sample test because there are two separate populations, those who exercise and those who do not. It is the **difference** between means because we are looking at the

*difference between the mean pulse rate of those who exercise and the mean pulse rate of those who do not.*

*Students should remember to check the condition that the pulse distributions for the two populations (those who do exercise regularly and those who do not) should be approximately normal. Students can draw a boxplot or a normal quantile plot for **both** samples in order to check for normality and evidence of outliers, an indication that the condition of normality may not be satisfied. Because of the small sample size, using a histogram to check for normality would not be appropriate.*

*Students may try to do this problem as a difference of two proportions because they may be accustomed to yes or no indicating proportions. This would be incorrect as we are looking for the **mean pulse rate** difference. Students may also try to do this as a single-sample *t*. Again this would be incorrect because we are looking for **difference** between those who exercise and those who don't. Students may think this a paired sample *t*-test. This is not correct because the samples are independent.*

8. The study “ Evaluation of Prescribed Burning in Relation to Available Deer Browse” was undertaken at the Virginia Polytechnic Institute and State University in 1964 to determine if fire can be used as a viable management tool to increase the amount of forage available to deer during the crucial months in late winter and spring. Calcium is a required element for plants and animals. The amount taken up and stored in the plant is closely correlated to the amount present in the soil. It was hypothesized that a fire may change the calcium levels present in the soil and thus affect the amount available to the deer. A large tract of land in the Fishburn Forest was selected for a prescribed burn. Soil samples were taken from 12 plots of equal area just prior to the burn on May 20, 1964, and analyzed for calcium. On July 16, 1964, postburn calcium levels were analyzed from the same plots. These values, in kilograms per plot, are presented in the following table:

Calcium Level  
( kg/plot)

Plot	Preburn	Postburn
1	50	9
2	50	18
3	82	45
4	64	18
5	82	18
6	73	9
7	77	32
8	54	9
9	23	18
10	45	9
11	36	9
12	54	9

Do these data support the hypothesis?

(Problem adapted from *Probability And Statistics for Engineers and Scientists*, 4<sup>th</sup> ed., R. E. Walpole and R. H. Myers, New York: Macmillan Publishing Company, 1972.)

**Solution**

**Procedure type:** Matched Pairs, t-test

$$H_o : \mathbf{m}_d = 0$$

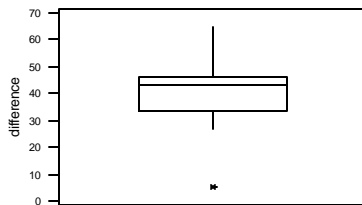
$$H_a : \mathbf{m}_d \neq 0$$

Where  $\mathbf{m}_d$  is the mean of the differences between pre- and postburn

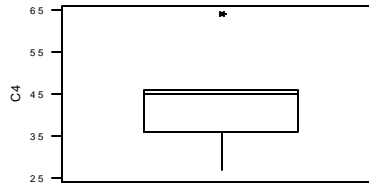
**Notes**

The soil samples are matched in pairs for calcium content prior to the burn and then again after the burn. In order to use the t procedure several conditions must be met: the samples must be simple random samples and the difference in calcium content between the preburn and the postburn measurements should be normally distributed. Although not specifically stated we do assume the plots within the forest are SRS based on the given information. In checking the boxplot for the distribution of the differences we discover the plot is skewed and there is one outlier at five. In looking at the data we found that all the differences are positive. Originally our thought was to remove the outlier. The removal of the outlier would make the test statistic less significant since the outlier is the closest value to 0. Another problem resulted because removal of the outlier created another outlier at 64! Now we need to remove that outlier because this is making the test more significant. In doing the test with both outliers, one outlier or neither outlier the results were significant. We feel confident in reporting that there was a change in calcium levels in the soil.

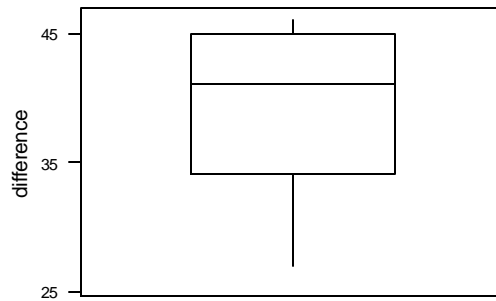
**Boxplot of the differences  
no outliers removed.**



**Boxplot of differences with outlier of 5  
removed**



**Boxplot of differences with outliers of 64 and 5 removed**



9. The article “Effect of Temperature on the pH of Skim Milk” (J. of Dairy Research (1988):277-280) reported on a study involving temperature ( $^{\circ}\text{C}$ ) under specified experimental conditions and milk pH. The accompanying data set is a representative subset of that which appeared in the article:

Temp	pH
4	6.85
4	6.79
24	6.63
24	6.65
25	6.72
38	6.62
38	6.57
40	6.52
45	6.50
50	6.48
55	6.42
56	6.41
60	6.38
67	6.34
70	6.32
78	6.34

Do these data strongly suggest that there is a negative (inverse) linear relationship between temperature and pH? Provide statistical justification for your conclusions.

(Problem adapted from *Introduction to Statistics and Data Analysis*; R. Peck, C. Olsen, and J. Devore; Pacific Grove, CA: Duxbury, 2001.)

### **Solution**

**Procedure type:** *t*-test for the slope

$$\begin{aligned} H_o : \mathbf{b}_1 &= 0 \\ H_a : \mathbf{b}_1 &< 0 \end{aligned} \quad \text{where } \mathbf{b}_1 \text{ represent the slope of the linear model}$$

### **Notes**

*In asking if there is a negative linear relationship between temperature and pH and by asking students to provide statistical justification for their conclusions, the question is asking students to test the significance of the slope. Students should realize that the problem suggests temperature explains the pH of milk, making temperature the independent variable (x) and pH the dependent variable (y).*

*Rejecting the null hypothesis would be evidence that a negative linear relationship did exist between temperature and pH. Since the p value is very close to 0 in this example, this would indicate a negative linear relationship between the two variables.*

*Students should begin by drawing a scatterplot to check for a linear relationship, fitting a least squares line, and creating a residual plot to look for patterns and evidence of unequal variance. By drawing a boxplot or a normal quantile plot of the residuals students can check to see if the error terms are approximately normal.*

*Because the data are presented as two columns, students may mistakenly do either a two-sample t or paired t-test. This error would be avoided if students took note of the words “linear relationship.” Also note that the question did not ask students to find a difference between two populations.*

10. It is generally believed that red-cockaded woodpeckers require or preferentially select old-age pine trees and stands for constructing nest and roost cavities. This hypothesis is important since the need for suitable habitat for nesting and roosting is vital to the continued existence of the species. The information for this exercise came from part of a study conducted by DeLotelle and Epting [1988]. Trees were sampled from stands of trees occupied by colonies of woodpeckers. Trees with current or abandoned woodpecker holes were termed *cavity trees*. Untouched trees in a defined neighborhood of cavity trees were called *colony trees*. The ages (in years) of cavity trees and colony trees were compared, giving the following summary statistics:

	Cavity Trees	Colony Trees
<b>Sample Size</b>	54	143
<b>Sample Mean</b>	104.1	83.6
<b>Sample Standard Deviation</b>	24.1	38.3

Is there evidence that cavity trees are older than colony trees on average?

(Problem adapted from *Chance Encounters*, C. J. Wild and G. A. F. Seber, New York: John Wiley & Sons, Inc., 2000)

**Solution**

**Procedure type:** *Two-sample t-test*

$H_0 : \mathbf{m}_{cav} = \mathbf{m}_{col}$ , where  $\mathbf{m}_{cav}$  is the mean age of cavity trees and  $\mathbf{m}_{col}$  is the mean age of colony trees

$H_0 : \mathbf{m}_{cav} > \mathbf{m}_{col}$

**Notes**

*The two populations are the cavity trees and the colony trees. The test is one-sided because the question asks if the cavity trees are **older** than colony trees.*

*Because this problem provides only summary statistics, students cannot check the distribution of the data using graphical analysis. With such large sample sizes, however, the distributions of the individual data values need not be normal, because the Central Limit Theorem ensures that the distributions of the  $\bar{x}$  's will be approximately normal. They also need to assume that the two samples are independent, which we cannot determine conclusively from the information given. A paired situation would arise if, for instance, the researchers identified pairs of trees that were similar in some way (such as location or species) and within each pair one was a cavity tree and the other a colony tree. In that case we would need information on the pairings and the data for each individual tree or the summary statistics for the differences in ages, not the summary statistics for the two separate samples.*

*Students who do not read carefully may assume the data constitute a two-way table and perform a chi-square test. But the non-integer entries will create problems if students attempt to treat them as counts.*

11. Researchers examined 103 adolescent males who had been diagnosed as hyperactive as children and identified those with drug abuse problems and those with antisocial personalities. They then examined 100 adolescent males who had not been diagnosed as hyperactive as children, using the same procedure. The data are given below.

<i>Adolescent problems of males diagnosed as hyperactive (n = 103) or not hyperactive (n = 100) as children. Numbers are percents of samples.</i>		
	HYPERACTIVE	NOT HYPERACTIVE
Drug abuse problem	16	4
Antisocial personality	27	8

- Comparing the males who were and were not hyperactive as children, is there a significant difference in the percentages of those with drug abuse problems?
- Comparing the males who were and were not hyperactive as children, is there a significant difference in the percentages of those with an antisocial personality?

(Problem adapted from *Statistics and Data Analysis; An Introduction*, 2<sup>nd</sup> ed., A. F. Siegel and C. J. Morgan, New York: John Wiley & Sons, Inc., 1996.)

**Solution**

**Procedure type:** Two-proportion z test

**Notes**

For part a):

$H_0: p_H = p_N$ , where  $p_H$  is the proportion of males who were hyperactive as children with drug abuse problems and  $p_N$  is the proportion of males who were not hyperactive as children with drug abuse problems

$H_a: p_H \neq p_N$

In this case the proportion of males who were hyperactive as children with drug abuse problems is 0.16 and the proportion of males who were not hyperactive as children with these problems is 0.04. The sample sizes are 103 for males who were hyperactive as children and 100 for males who were not hyperactive as children. In checking the required conditions for the two-proportion z test, different instructors favor different procedures, and in this case they yield different results. If we use the sample proportions separately, evaluating  $n_1 \hat{p}_1$ ,

$n_1\hat{q}_1$ ,  $n_2\hat{p}_2$ , and  $n_2\hat{q}_2$ , we find that  $100 * 0.04$  is less than 5, which causes us concern in carrying out the procedure. (Some instructors would compare those values to 10, which is a more conservative condition. In this case the data do not satisfy that condition.) But if we check the conditions using the pooled sample proportion,  $\frac{16+4}{103+100}$ , multiplying both  $\hat{p}$  and  $\hat{q}$  by either sample size yields a value greater than 5. This is the least conservative condition of the three commonly used in AP statistics courses. These data satisfy that condition, so students using this condition would proceed with the analysis. Instructors (and students) should recognize that blindly applying this condition may yield inaccurate analysis. For instance, if  $\hat{p}_1 = \frac{2}{100}$  and  $\hat{p}_2 = \frac{98}{100}$ , we can see that the two sample proportions are not likely to be normally distributed, because they are near 0 and 1, respectively. (Recall that sample proportions are limited to the values between 0 and 1, so their distributions cannot extend beyond those values.) But if we use the pooled sample proportion in this case, we obtain  $\frac{2+98}{100+100} = \frac{100}{200} = 0.5$ , and multiplying by 100 yields 50, greater than both 5 and 10. Students using this procedure would then wrongly proceed with the tests that are based on the assumption of normality of the distributions of the sample proportions.

In addition, we are given no information on the sampling procedure, so to proceed students would have to agree to treat the data as arising from a random sample.

The hypotheses for b) are

$H_0: p_H = p_N$ , where  $p_H$  is the proportion of males who were hyperactive as children with an antisocial personality and  $p_N$  is the proportion of males who were not hyperactive as children with an antisocial personality

$H_a: p_H \neq p_N$

In this case the proportion of hyperactive males with an antisocial personality is 0.27 and the proportion of non-hyperactive males with this problem is 0.08. Here the sample proportions and sample sizes do not cause us concern in checking the conditions, but the sampling procedure once again is an issue.

Students who do not read this problem carefully can easily make mistakes. One error would be to assume the data are in a two-way table and perform a chi-square analysis. But the numerical values are not counts, and a particular male may have both a drug abuse problem and an antisocial personality. Students may also use the row sums as their sample sizes for each of the questions.

Some students may decide to use a chi-square analysis by converting the percents to counts. But the problem does not seek the relationship between hyperactivity diagnosis and adolescent status, so they would then have to create a separate two-way table for each question. Their rows would be (for part a) Drug Abuse Problem and No Drug Abuse

*Problem, and all their entries would be counts. Since the test is two-sided, this is a reasonable procedure.*

12. On September 16, 1992, *USA Today* reported the results of a study which links female sexuality to genes. The study, conducted by researchers at Boston University, involved 115 homosexual or bisexual women who had twin sisters. Seventy-seven of the women had identical twins, who share all of the same genes, and 38 of the women had fraternal twins, who are related from a genetic standpoint just like other siblings. The researchers determined the sexual preference of the twin sisters. The results are as follows:

Sexual Preference of Twin Sisters		
Type of Twin	Number Surveyed	# Homosexual or Bisexual
Identical	77	39
Fraternal	38	6

If the proportion of identical twins who are also homosexual is significantly larger than the proportion of fraternal twins who are also homosexual, the researchers will conclude that homosexual preference is linked to genes. Based on the data, is there evidence that homosexual preference is genetically linked?

(Problem adapted from *Discovering Statistics; An Adventure in Problem Solving*, J. S. Hawkes, Charleston: Quant Publishing, 1995.)

**Solution**

**Procedure type:** Two-proportion  $z$  test

$H_0 : p_{id} = p_{fr}$ , where  $p_{id}$  is the proportion of identical twin sisters who are homosexual and  $p_{fr}$  is the proportion of fraternal twin sisters who are homosexual

$H_a : p_{id} > p_{fr}$

**Notes**

*In this problem we have two populations, identical twin sisters and fraternal twin sisters. In the sample, the proportion of identical twin sisters who are homosexual is  $\frac{39}{77}$  and the proportion of fraternal twin sisters who are homosexual is  $\frac{6}{38}$ . In checking the conditions*

for the two proportion procedure, we find that 39, 77-39, 6, and 38-6 are all greater than 5. But the problem description does not describe the sampling technique, so in order to proceed students must assume the researchers incorporated random sampling.

One error students may make is to immediately assume that this problem requires a chi-square analysis, because the data are presented in what appears to be a two-way table. But the entries in the first column of the table are actually the sample sizes, which would be the row totals of the correct two-way table. In addition, the test is one-sided, and chi-square tests are only appropriate for two-sided situations.

Another error would be to add the entries in each row and use the row totals as the sample sizes.

13. Charles Darwin carried out an experiment to study whether seedlings from cross-fertilized plants tend to be superior to those from self-fertilized plants. He covered a number of plants with fine netting so that insects would be unable to fertilize them. He fertilized a number of flowers on each plant with their own pollen and he fertilized an equal number of flowers on the same plant with pollen from a distant plant. (He did not say how he decided which flowers received which treatments.) The seeds from the flowers were allowed to ripen and were set in wet sand to germinate. He placed two seedlings of the same age in a pot, one from a seed from a self-fertilized flower and one from a seed from a cross-fertilized flower. (The fertilization experiments were described by Darwin in an 1878 book; these data were found in D. F. Andrews and A. M. Herzberg, *Data* (New York: Springer-Verlag, 1985), pp. 9-12.) Below are the heights of the plants at a particular age.

---

***Darwin's data: plant heights (inches) for 15 pairs of plants of the same age, one of which was grown from a seed from a cross-fertilized flower and the other of which was grown from a seed from a self-fertilized flower***

---

<b>Plant Heights (inches)</b>		
<b>Pair</b>	<b>Cross-fertilized</b>	<b>Self-fertilized</b>
<i>1</i>	23.5	17.375
<i>2</i>	12	20.375
<i>3</i>	21	20
<i>4</i>	22	20
<i>5</i>	19.125	18.375
<i>6</i>	21.5	18.625
<i>7</i>	22.125	18.625
<i>8</i>	20.375	15.25
<i>9</i>	18.25	16.5
<i>10</i>	21.625	18
<i>11</i>	23.25	16.25
<i>12</i>	21	18
<i>13</i>	22.125	12.75
<i>14</i>	23	15.5
<i>15</i>	12	18

---

If we consider taller plants “superior” to shorter plants, do these data provide evidence that cross-fertilization produces superior plants to self-fertilization? Support your answer with the appropriate statistical justification.

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

### Solution

**Procedure type:** Paired *t*-test

$H_0 : \mathbf{m} = 0$ , where  $\mathbf{m}$  is the mean difference in height (in inches), self-fertilized – cross-fertilized

$H_a : \mathbf{m} < 0$

or

$H_0 : \mathbf{m} = 0$ , where  $\mathbf{m}$  is the mean difference in height (in inches), cross-fertilized – self-fertilized

$H_a : \mathbf{m} > 0$

### Notes

*The experimental design is critical in determining the appropriate analysis. From the description it is not clear whether the two seedlings in each pot came from the same plant, so we cannot include the original plant as a factor in our analysis. But we know that the two seedlings within each pot will experience the same conditions, so we do include that information. That leads us to treat this as a matched pairs design and analyze the results with a paired *t*-test, using the differences in heights as our data.*

*For this to be a two-sample problem, the design would have to be different. For instance, if Darwin had used one group of plants to obtain the self-fertilized seeds and another group to obtain the cross-fertilized seeds, and then he placed each seed in its own separate pot, he would have two independent samples of seeds. We would then analyze the results using the two-sample *t*-test.*

*In order to conduct a one-sample *t*-test on the differences, students must calculate the differences and examine their distribution. In this case, the data do not satisfy the condition of approximate normality—there is at least one high outlier. With such a small sample size ( $n = 15$ ) this is a problem. Careful students would not proceed with the test.*

*The most common error students will make on this problem is to treat it as a two-sample  $t$  problem. But the two sets of seedlings are not independent, which is a requirement of the two-sample test. Another error students may make would be to create a scatterplot of the data and fit a line or perform a significance test on the slope of the regression line. While that may prove interesting, it would not address the question asked.*

## PART 2

### SCENARIOS WITH MULTIPLE INFERENCE PROBLEMS

#### Introduction

This section presents two scenarios, including data sets, each followed by a group of inference problems and their solutions. Teachers may use these scenarios in a number of ways. One method would be to break the class into small groups and assign one problem to each group. In the follow-up discussion, teachers could note that the inference procedure and analysis are highly dependent upon the question posed, not just on the data's format or the way the data were obtained. Since the problems cover a variety of inference procedures, teachers should not use these problems until after covering all inference topics.

An alternative method would be for teachers to select a subset of the problems for students to solve at different points in the year. This would be particularly effective with the second scenario, because it covers more inference procedures. Teachers could remind students of the scenario and then assign a new problem. As in the first method, teachers should use the problems to show that the question—not just the scenario—will determine the inference procedure.

## PART 2

### INFERENCE SCENARIOS

#### Scenario 1

The table below shows **relative brain weights** (brain weight divided by body weight) for 51 species of mammal whose average litter size is less than 2 and for 45 species of mammal whose average litter size is greater than or equal to 2.

#### Relative Brain Weights x 1000 for 96 Species of Mammals

##### *For 51 species with average litter size < 2*

0.42	0.86	0.88	1.11	1.34	1.38	1.42	1.47	1.63	1.73	2.17	2.42	2.48	2.74	2.74	2.79	2.90
3.12	3.18	3.27	3.30	3.61	3.63	4.13	4.40	5.00	5.20	5.59	7.04	7.15	7.25	7.75	8.00	8.84
9.30	9.68	10.32	10.41	10.48	11.29	12.30	12.53	12.69	14.14	14.15	14.27	14.56	15.84	18.55	19.73	20.00

##### *For 45 species with average litter size = 2*

0.94	1.26	1.44	1.49	1.63	1.80	2.00	2.00	2.56	2.58	3.24	3.39	3.53	3.77	4.36
4.41	4.60	4.67	5.39	6.25	7.02	7.89	7.97	8.00	8.28	8.83	8.91	8.96	9.92	11.36
12.15	14.41	16.00	18.61	18.75	19.05	21.00	21.41	23.27	24.71	25.00	28.75	30.23	35.45	36.35

1. Suppose you are a statistician working for the San Diego Zoo. You are asked to estimate the mean relative brain weights of mammals with small litters (fewer than 2) and of those with larger litters (at least 2). Report your estimates and justify your response.
2. Do these data provide evidence that mean relative brain weights tend to be different for the two groups of mammals? Include the appropriate statistical justification in your response.
3. Use appropriate statistical techniques to estimate the difference between the mean relative brain weights of the two groups of mammals. Explain your response in carefully worded sentences.

(Problem adapted from *The Statistical Sleuth*, F. L. Ramsey and D. W. Schafer, Belmont, CA: Duxbury Press, 1997.)

### Solutions for Scenario 1 Questions

1.

**Procedure type:** confidence intervals for a mean

#### Notes

Students should recognize that when they must estimate a parameter they report a confidence interval. The question seeks two different confidence intervals, one for the mean of each class of mammal. In both cases the sample size is large, so the shape of the data distribution need not appear approximately normal. Students should comment on the sampling method; the procedures require a simple random sample, and we do not know whether the data meet this requirement.

The 95% confidence interval for the mean relative brain weight of mammals with litters < 2 is (5.130, 8.180), in the same units as the data in the table. The 95% means that the **method** we used to construct the interval will produce an interval that contains the true mean 95% of the time in repeated sampling. We do not know whether our particular interval contains the true mean.

The 95% confidence interval for the mean relative brain weight of mammals with litters < 2 is (8.01, 13.92), in the same units as the data in the table.

2.

**Procedure type:** two-sample t-test

$H_0: \mathbf{m}_s = \mathbf{m}_l$ , where  $\mathbf{m}_s$  is the mean relative brain weight of mammals with small litters and  $\mathbf{m}_l$  is the mean relative brain weight of mammals with larger litters

$H_a: \mathbf{m}_s \neq \mathbf{m}_l$

#### Notes

As in problem 1, in both cases the sample size is large, so the shape of the data distribution need not appear approximately normal. Students should comment on the sampling method; the procedures require a simple random sample, and we do not know whether the data meet this requirement. We also require two independent samples, and that seems a reasonable assumption in this problem.

Because the test is two-sided, students can carry it out using either a test statistic or a confidence interval for the difference between the two means.

With a test statistic: The test statistic is  $t = -2.61$ , with a  $p$ -value of 0.011, smaller than  $\alpha = 0.05$ , leading students to reject the null hypothesis and conclude that we have strong evidence that the mean relative brain weights differ for the two classes of mammals.

With a confidence interval: The 95% confidence interval for the difference between the two mean relative brain weights is  $(-7.61, -1.0)$ , and since this interval does not include 0 we can conclude that we have strong evidence that the mean relative brain weights differ for the two classes of mammals.

3.

**Procedure type:** confidence interval for the difference between two means

### Notes

As in problem 1, students should recognize that when they must estimate a parameter they report a confidence interval. As in both previous problems, in both cases the sample size is large, so the shape of the data distribution need not appear approximately normal. Once again, students should comment on the sampling method; the procedures require simple random samples, and we do not know whether the data meet this requirement. We also require two independent samples, and that seems a reasonable assumption in this problem.

The 95% confidence interval for the difference between the two mean relative brain weights is  $(-7.61, -1.0)$ . Students should interpret this to mean that they believe the true mean amount by which the relative brain weight of species with large litters exceeds the relative brain weight of species with small litters is between 1 and 7.61. The 95% means that the **method** used to construct the interval will produce an interval that contains the true mean 95% of the time in repeated sampling. We do not know whether our particular interval contains the true mean difference.

Often students feel they need to give non-statistical reasons why they obtained a particular result. For instance, they may discuss animals they are familiar with, such as cats, dogs, horses, and humans. But those comments do not add to the statistical content of the response and generally should be omitted.

**Scenario 2**

John Gottman (1994) has done 20 years of research on why marriages succeed or fail. His basic finding is that there must be a ratio of at least five positive acts to each negative act for the marriage to last. Suppose you decide to replicate this study. You recruit 13 couples and make videotapes of each couple discussing an important issue in their relationship. You then count the number of positive acts (compliments, smiles, agreements, and so forth) in each interaction, as well as the number of negative acts (scowls, insults, and so on). From these numbers you then calculate the ratio of positive to negative acts, as well as the total number of acts. Three years later you relocate the original couples and see whether they have broken up (1 = broke up, 0 = did not break up). The data are shown below.

**Positive and negative actions between couples, and whether they break up within 3 years**

POSITIVE	NEGATIVE	RATIO	TOTAL	BREAK UP
25	6	4.1667	31	1
23	5	4.6000	28	0
26	10	2.6000	36	1
21	15	1.4000	36	1
8	2	4.0000	10	0
10	1	10.0000	11	0
13	1	13.0000	14	0
11	5	2.2000	16	1
46	10	4.6000	56	0
43	5	8.6000	48	0
51	15	3.4000	66	1
55	25	2.2000	80	1
75	15	5.0000	90	0

1. According to the given data, is there a statistically significant difference between the couples who broke up and those who stayed together in terms of the mean number of positive acts observed? Justify your answer with appropriate statistical evidence and clearly written sentences.
2. According to the given data, is there a statistically significant difference between the couples who broke up and those who stayed together in terms of the mean number of negative acts observed? Justify your answer with appropriate statistical evidence and clearly written sentences.

3. According to the given data, is there a statistically significant difference between the couples who broke up and those who stayed together in terms of the mean ratio of positive to negative acts observed? Justify your answer with appropriate statistical evidence and clearly written sentences.
4. Calculate the differences between the numbers of positive acts and negative acts for each couple. Do these data provide evidence that the mean difference differs significantly between couples who broke up and those who did not? Is performing the analysis using the **differences** more reasonable or less reasonable than using the **ratios** of positive to negative acts? Explain your answer and include statistical justification.
5. Use the data to estimate the mean ratio for couples who stay together. Include statistical justification in your clearly written response.
6. Do these data provide evidence that couples who broke up had a mean ratio of positive to negative acts less than 5? Support your response with appropriate statistical justification.
7. Is there a significant relationship between the number of negative acts and the number of positive acts for all the pairs of couples? Give a statistical justification to support your response.
8. Consider the couples who broke up and those who did not separately. In either case, is there a significant relationship between the number of negative acts and the number of positive acts? Is the relationship stronger for one set of couples? Justify your answer with appropriate statistical evidence and carefully worded sentences.

(Problem adapted from *Statistics and Data Analysis; An Introduction, Second Edition*, A. F. Siegel and C. J. Morgan, New York: John Wiley & Sons, Inc., 1996.)

**Solutions for Scenario 2 Questions**

1.

**Procedure type:** Two-sample t-test

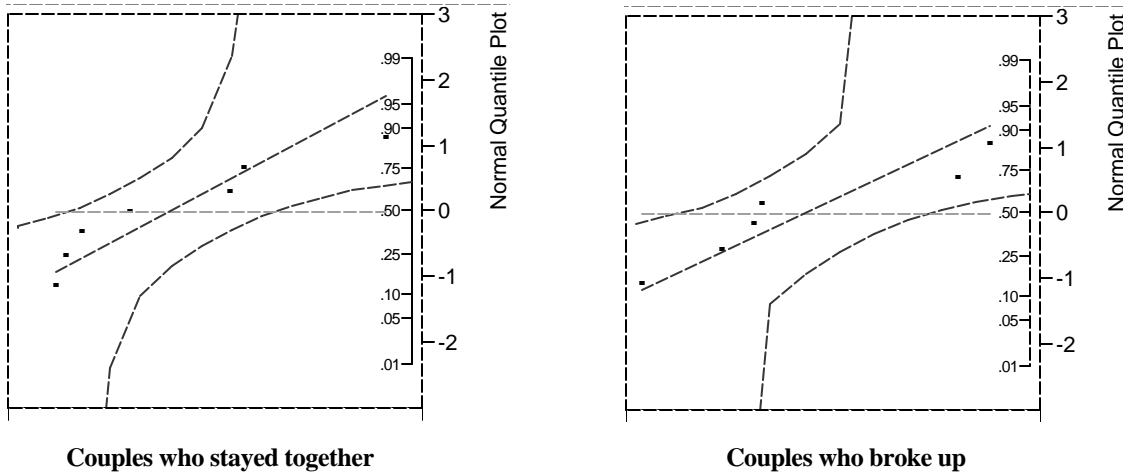
$H_0 : \mathbf{m}_b = \mathbf{m}_s$ , where  $\mathbf{m}_b$  is the mean number of positive acts for couples who broke up and  $\mathbf{m}_s$  is the mean number of positive acts for couples who stayed together

$H_a : \mathbf{m}_b \neq \mathbf{m}_s$

**Notes**

We will treat this as two independent samples, but clearly the researchers could not have known which couples would break up in the future. In addition, we do not have any indication of how the researchers incorporated randomization in their study. Students should comment on these issues in their responses.

Students should graphically examine the two data sets. Both are very small samples, so it is difficult to judge the normality of them. Below are the normal quantile plots for the two data sets.



The first normal quantile plot is of the number of positive acts by the couples who did not break up, and it is clearly skewed. But  $n = 7$  for that sample, and if only one data value changed the plot would change dramatically. The second normal quantile plot does not cause us concern, but  $n = 6$  for that group. For such small samples, normal quantile plots are the best choice for these data sets. Boxplots for such small samples would be a poor choice.

Since this is a two-sided test, students may choose to calculate either a test statistic or a confidence interval for the difference between the two means. If they use a confidence interval, however, they still must clearly state their hypotheses.

With a test statistic: The test statistic is  $t = 0.03$ , with a  $p$ -value of 0.98, so we have no evidence that the mean number of positive acts differs for the two types of couples.

With a confidence interval: The 95% confidence interval for the difference between the two mean numbers of positive acts is  $(-25.9, 26.6)$ , and since this interval contains 0 we have no evidence of a difference between the mean numbers of positive acts for the two types of couples.

2.

**Procedure type:** Two-sample  $t$ -test

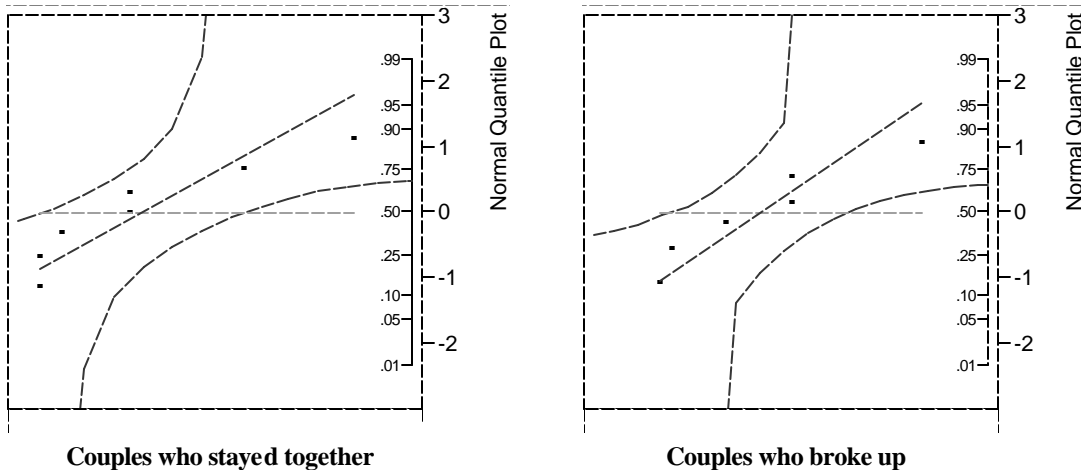
$H_0 : \mathbf{m}_b = \mathbf{m}_s$ , where  $\mathbf{m}_b$  is the mean number of negative acts for couples who broke up and  $\mathbf{m}_s$  is the mean number of negative acts for couples who stayed together

$H_a : \mathbf{m}_b \neq \mathbf{m}_s$

**Notes**

We will treat this as two independent samples, but clearly the researchers could not have known which couples would break up in the future. In addition, we do not have any indication of how the researchers incorporated randomization in their study. Students should comment on these issues in their responses.

Students should graphically examine the two data sets. Both are very small samples, so it is difficult to judge the normality of them. Below are the normal quantile plots for the two data sets.



Neither normal quantile plot shows strong skewness, so the normality assumption does not appear to be violated. Normal quantile plots are the best choice for these data sets. Boxplots for such small samples would be a poor choice.

Since this is a two-sided test, students may choose to calculate either a test statistic or a confidence interval for the difference between the two means. If they use a confidence interval, however, they still must clearly state their hypotheses.

With a test statistic: The test statistic is  $t = 1.97$ , with a  $p$ -value of 0.085, so we have some evidence that the mean number of negative acts differs for the two types of couples. Some students will use  $\alpha = 0.05$  and others will use  $\alpha = 0.10$ . The conclusion will vary depending on those values.

With a confidence interval: The 95% confidence interval for the difference between the two mean numbers of positive acts is  $(-1.3, 15.4)$ , and since this interval contains 0 we do not have strong evidence of a difference between the mean numbers of negative acts for the two types of couples. Students using a 90% confidence interval will reach the opposite conclusion.

### 3.

**Procedure type:** Two-sample  $t$ -test

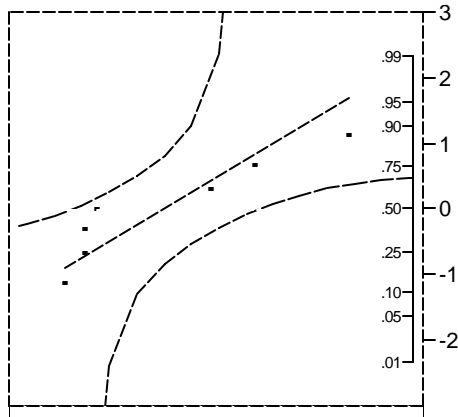
$H_0 : \mathbf{m}_b = \mathbf{m}_s$ , where  $\mathbf{m}_b$  is the mean ratio of positive to negative acts for couples who broke up and  $\mathbf{m}_s$  is the mean ratio of positive to negative acts for couples who stayed together

$H_a : \mathbf{m}_b \neq \mathbf{m}_s$

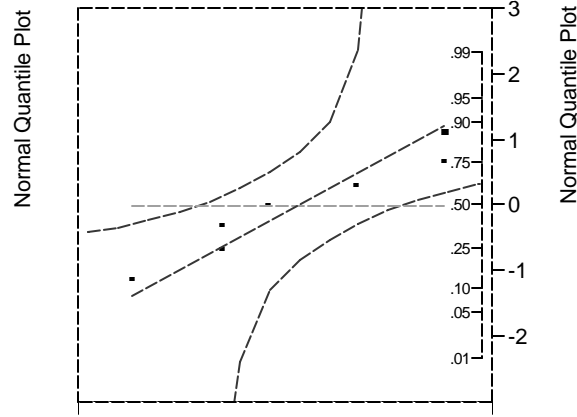
#### Notes

We will treat this as two independent samples, but clearly the researchers could not have known which couples would break up in the future. In addition, we do not have any indication of how the researchers incorporated randomization in their study. Students should comment on these issues in their responses.

Students should graphically examine the two data sets. Both are very small samples, so it is difficult to judge their normality. Below are the normal quantile plots for the two data sets.



**Couples who stayed together**



**Couples who broke up**

In the first normal quantile plot, the curvature is a problem, but with no points outside the confidence bounds, we will proceed cautiously with our analysis. The second plot appears fairly linear, so we have no qualms about proceeding. Boxplots for such small samples would be a poor choice.

*Since this is a two-sided test, students may choose to calculate either a test statistic or a confidence interval for the difference between the two means. If they use a confidence interval, however, they still must clearly state their hypotheses.*

With a test statistic: The test statistic is  $t = -3.25$ , with a  $p$ -value of 0.014, so we have strong evidence (if  $\alpha = 0.05$ ) that the mean ratio of positive acts to negative acts differs for the two types of couples.

With a confidence interval: The 95% confidence interval for the difference between the two mean numbers of positive acts is  $(-7.69, -1.2)$ , and since this interval does not contain 0 we have good evidence of a difference between the mean ratios of positive acts to negative acts for the two types of couples.

**4.**

**Procedure type:** Two-sample  $t$ -test

$H_0 : \mathbf{m}_b = \mathbf{m}_s$ , where  $\mathbf{m}_b$  is the mean difference between the number of positive acts and the number of negative acts for couples who broke up and  $\mathbf{m}_s$  is the mean difference between the number of positive acts and the number of negative acts for couples who stayed together

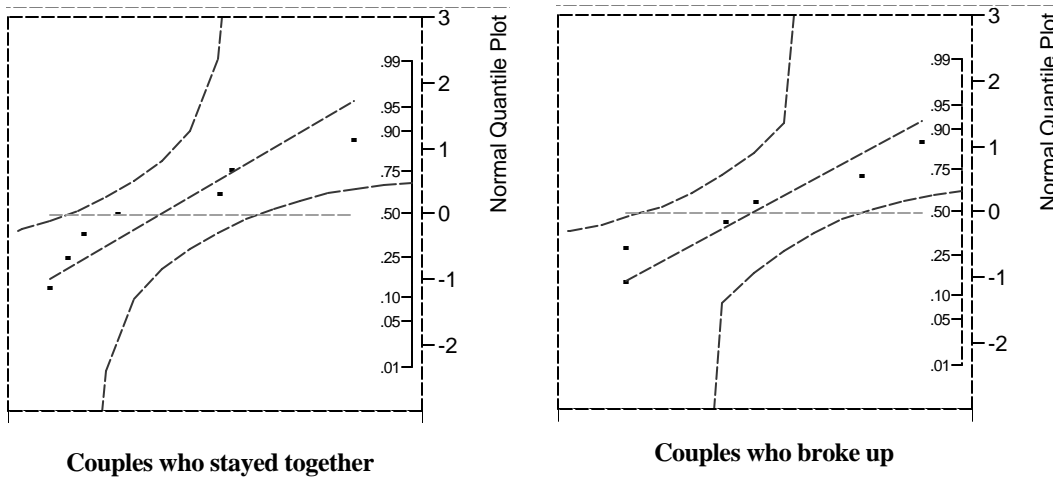
$H_a : \mathbf{m}_b \neq \mathbf{m}_s$

**Notes**

Students must first calculate the differences. Some may compute positive – negative and others may compute negative – positive. This solution uses the first option.

We will treat this as two independent samples, but clearly the researchers could not have known which couples would break up in the future. In addition, we do not have any indication of how the researchers incorporated randomization in their study. Students should comment on these issues in their responses.

Students should graphically examine the two data sets. Both are very small samples, so it is difficult to judge their normality. Below are the normal quantile plots for the two data sets.



Neither plot shows strong skewness, so the normality assumption does not appear to be violated. (The first plot shows a slightly skewed distribution, but no points are outside the confidence bands.) Boxplots for such small samples would be a poor choice.

Since this is a two-sided test, students may choose to calculate either a test statistic or a confidence interval for the difference between the two means. If they use a confidence interval, however, they still must clearly state their hypotheses.

With a test statistic: The test statistic is  $t = -0.75$ , with a  $p$ -value of 0.47, so we do not have evidence that the mean difference between the numbers of positive acts and negative acts differs for the two types of couples.

With a confidence interval: The 95% confidence interval for the difference between the two mean differences between the numbers of positive acts and negative acts is  $(-26.8, 13.3)$ , and since this interval contains 0 we do not have evidence of a difference between the mean differences between the numbers of positive acts and negative acts for the two types of couples.

As for the appropriateness of this analysis, students should recognize that the differences do not account for the range of total numbers of acts for the thirteen couples. One couple

committed a total of 90 acts, while two other couples committed only 11 acts. Thus the difference between the numbers of positive and negative acts is necessarily limited by the total number of acts for the couple. The ratios account for the different totals by computing the relative difference.

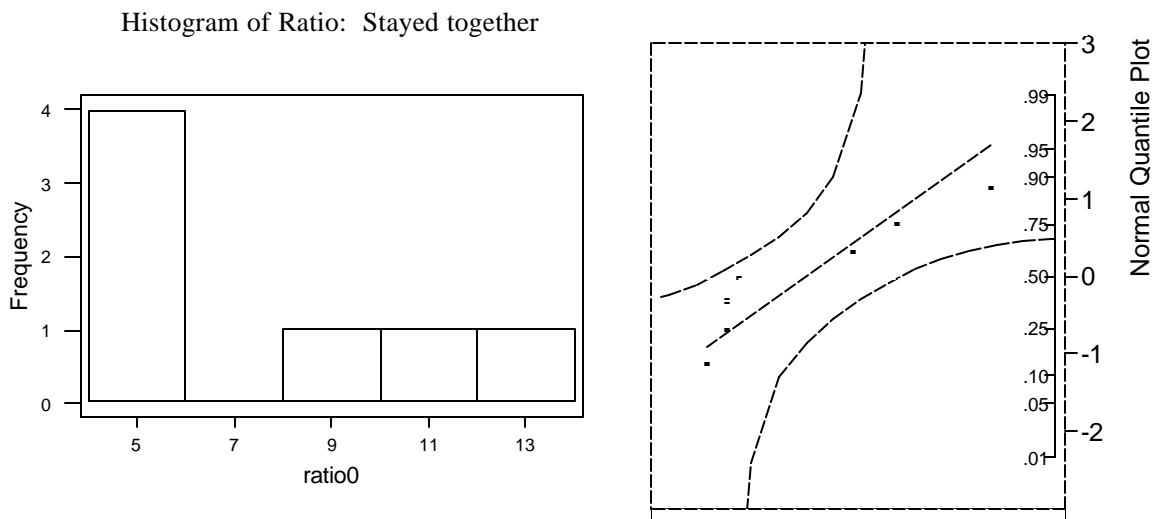
5.

**Procedure type:** Confidence interval for a mean

**Notes**

Students should recognize that when they must estimate a parameter they report a confidence interval. They should also note that we do not have any indication of how the researchers incorporated randomization in their study.

Before computing the confidence interval students should examine the data set graphically. Below are a histogram and a normal quantile plot of the ratios for couples who stay together.



The distribution is slightly skewed, but with the small sample size ( $n = 7$ ) it is difficult to assess normality. Since no data points appear outside the confidence bands, we will proceed with our analysis.

The 95% confidence interval for the mean ratio of positive to negative acts for couples who stayed together is (3.91, 10.32). The 95% means that the **method** we used to construct the interval will produce an interval that contains the true mean 95% of the time in repeated sampling. We do not know whether our particular interval contains the true mean.

6.

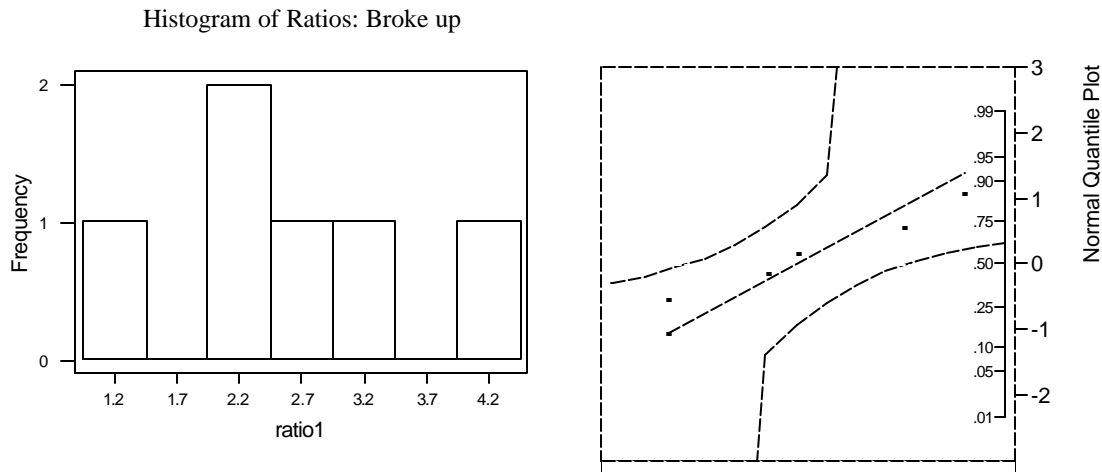
**Procedure type:** One-sample t-test

$H_0 : \mathbf{m} = 5$ , where  $\mathbf{m}$  is the mean ratio of the number of positive acts to the number of negative acts for couples who broke up

$H_a : \mathbf{m} < 5$

**Notes**

Students should note that we do not have any indication of how the researchers incorporated randomization in their study. Before carrying out the test students should examine the data set graphically. Below are a histogram and a normal quantile plot of the ratios for couples who break up.



Neither display indicates skewness, so we may proceed with our analysis. Notice that since the test is one-sided, students should not use a confidence interval to complete the test.

The test statistic is  $-5.83$ , with a corresponding  $p$ -value of  $0.0011$ . With any reasonable value of  $\alpha$  students would reject the null hypothesis and conclude that we have strong evidence that the mean ratio of numbers of positive to negative acts for couples who broke up is less than 5. But this does not necessarily mean that the low ratio causes the break-up. This study was not an experiment and so does not provide evidence of causation.

7.

**Procedure type:** *t*-test on the correlation or the slope of the regression line

$H_0 : \mathbf{r} = 0$ , where  $\mathbf{r}$  is the correlation between the number of positive acts and the number of negative acts for all couples

$H_a : \mathbf{r} \neq 0$

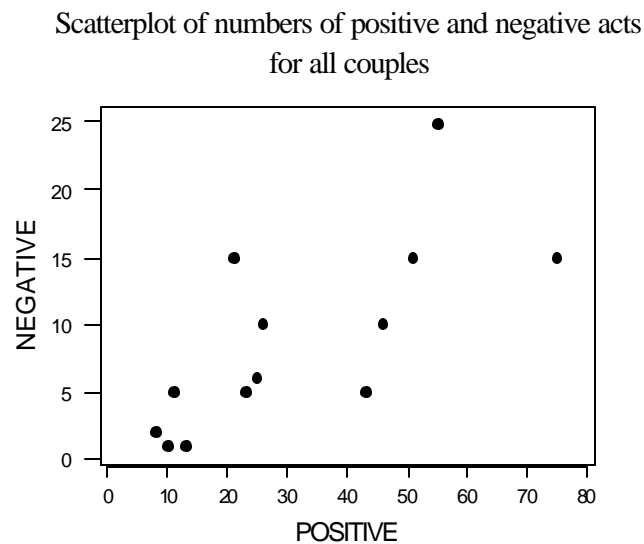
or

$H_0 : \mathbf{b}_1 = 0$ , where  $\beta_1$  is the slope of the population regression line for the number of positive acts and the number of negative acts for all couples

$H_a : \mathbf{b}_1 \neq 0$

### Notes

Students should recognize that seeking a relationship between two quantitative variables means that they should consider a scatterplot of the data. Below is an appropriate scatterplot.



Students should note that the data appear to be positively associated, but the relationship may not be strong.

The test statistic (for either set-up) is  $t = 3.45$ , with a corresponding  $p$ -value of 0.005. Using any reasonable  $\alpha$ , we have strong evidence that there is a linear relationship between the number of positive acts and the number of negative acts for all the couples.

8.

**Procedure type:**  $t$ -test on the correlation or the slope of the regression line

For the couples who broke up:

$H_0: \mathbf{r} = 0$ , where  $\mathbf{r}$  is the correlation between the number of positive acts and the number of negative acts for the couples who broke up

$H_a: \mathbf{r} \neq 0$

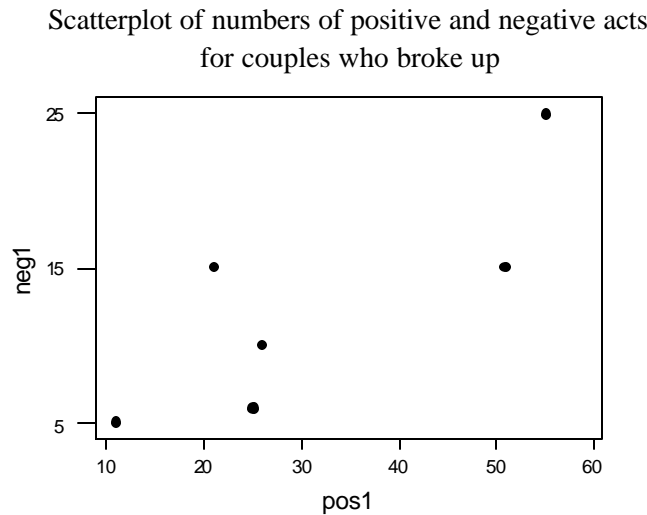
or

$H_0: \mathbf{b}_1 = 0$ , where  $\beta_1$  is the slope of the population regression line for the number of positive acts and the number of negative acts for the couples who broke up

$H_a: \mathbf{b}_1 \neq 0$

### Notes

Students should recognize that seeking a relationship between two quantitative variables means that they should consider a scatterplot of the data. Below is an appropriate scatterplot.



With so few data points, it is difficult to assess the relationship, but students should notice the positive association.

The test statistic (for either set-up) is  $t = 2.78$ , with a corresponding  $p$ -value of 0.05. For those who use  $\alpha = 0.05$ , this result may be confusing. Many textbook authors consider a test to be significant at  $\alpha = 0.05$  if the  $p$ -value is less than **or equal to** 0.05; other authors require the  $p$ -value to be less than 0.05. Teachers may choose to follow their own textbook's rule. Students who use  $\alpha = 0.10$  will find evidence of a linear relationship between the numbers of positive and negative acts for the couples who broke up.

For the couples who stayed together:

$H_0 : \mathbf{r} = 0$ , where  $\mathbf{r}$  is the correlation between the number of positive acts and the number of negative acts for the couples who stayed together

$H_a : \mathbf{r} \neq 0$

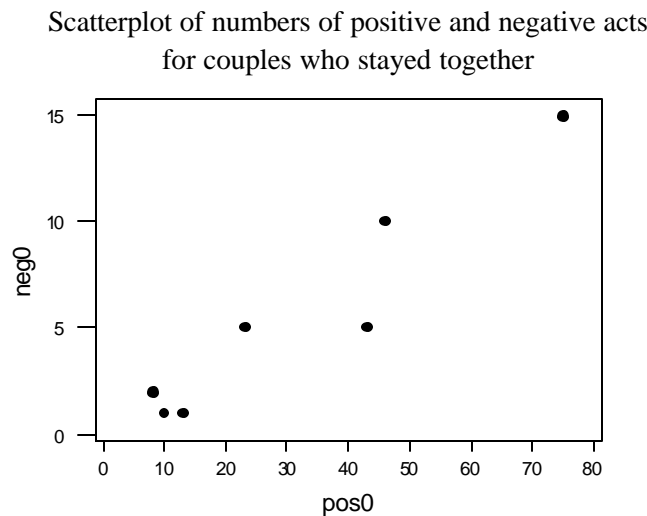
or

$H_0 : \mathbf{b}_1 = 0$ , where  $\beta_1$  is the slope of the regression line for the number of positive acts and the number of negative acts for the couples who stayed together

$H_a : \mathbf{b}_1 \neq 0$

### Notes

Students should recognize that seeking a relationship between two quantitative variables means that they should consider a scatterplot of the data. Below is an appropriate scatterplot.



The scatterplot shows the data are positively associated, and in fact the relationship appears to be linear.

The test statistic (for either set-up) is  $t = 7.19$ , with a corresponding  $p$ -value of 0.001. Using any reasonable  $\alpha$ , we have strong evidence that there is a linear relationship between the number of positive acts and the number of negative acts for these couples. This confirms our observations of the scatterplot.

## **PART 3**

### **INFERENCE SET-UP PROBLEMS**

#### **Introduction**

This section presents a group of inference problems without any data. Identifying the appropriate inference procedure continues to be one of the most difficult tasks for AP statistics students, so teachers should provide students with many opportunities to practice this technique. For each of the following problems, students must interpret the goals of the study, state the hypotheses for the appropriate test, and identify the correct inference procedure to use after data collection. Students should also evaluate the study's design and comment on whether it is reasonable or flawed. Teachers should encourage students to recognize the important connection between experimental design and inference.

Following the sheet of questions are solutions and notes to teachers.

## PART 3

### WORKSHEET OF INFERENCE SET-UP PROBLEMS

1. A teacher wants to know if the method of instruction affects how well students learn. Using two classes of the same level of statistics, she teaches one class using lecture only and the other class using lecture and group work. She measures the level of learning by giving both classes the same test. Assuming that the two classes are representative of all statistics students, what type of inference procedures should be used? State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
2. A student of political science wished to determine whether there is a statistically significant relationship between the gender of a student and their political affiliation. A simple random sample of forty-five students was selected from the school and asked on a survey whether their political affiliation was more likely to be liberal, conservative or moderate. Students were also categorized as male or female. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
3. A student was interested in whether an on-line SAT review course increased students' math scores. To answer his question he randomly selects 30 students from the population of all the students who took the course and records their SAT scores before the course and after finishing the course. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
4. A student wishes to know if SUV drivers at his school are more likely to be male than female. He randomly selects 50 students from a large list who are SUV drivers and records their gender. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
5. Your friend in Portland claims that it seems like lots of drivers who pass her while she awaits the school bus are talking on a cell phone. You think it's a worse problem in your hometown. One randomly selected day, you and your friend agree to keep track of the number of cell phone users and total cars that pass by you between 7 and 7:30 a.m. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
6. In your psychology class, your group (5 students) wants to investigate the relative intelligence of mice. You claim that male mice are more intelligent than female mice, but your group members disagree. You decide to perform an experiment on mice, using mazes. Each of you has one male and one female mouse at home (for a total of

- 10 mice), and you each build a different maze. Each of you will allow each mouse one trial and record the time to reach the cheese at the end of the maze. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
7. A high school physical education teacher is curious whether doing plyometrics three days a week will help athletes sprint faster. On a particular day of class in the fall, the teacher times all the students in a 40-yard sprint. For the next six weeks, the students perform 10 minutes of plyometric exercises three days a week. At the end of the six weeks, the teacher times all the students in another 40-yard sprint. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
  8. Your friend studies German and you study French. Your friend is always complaining about how long the words are, but you think French words are longer, on average, than German words. You decide to determine which student is correct. You each find a different novel in the appropriate language, randomly choose a page in the novel, and count the lengths of all the words on the page. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.
  9. Researchers have noted that sleep deprivation leads to car accidents and other mistakes, often due to inattention or slower reaction time. In order to examine the level of sleep deprivation in high school students, a researcher performs the following study. At 10 a.m. on a particular school day, students in two classes play a computer game that is actually recording the time it takes them to negotiate a mental obstacle course. At 2 p.m. that day, one of the classes is given 30 minutes in a silent, dark room with comfortable furniture, and the students are allowed to sleep. The other class has regular classes. At 3 p.m., both classes play the computer game again. The researcher records the differences in the times it takes each student to complete the game. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

## PART 3

### Worksheet Solutions and Notes

1. A teacher wants to know if the method of instruction affects how well students learn. Using two classes of the same level of statistics, she teaches one class using lecture only and the other class using lecture and group work. She measures the level of learning by giving both classes the same test. Assuming that the two classes are representative of all statistics students, what type of inference procedures should be used? State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

#### Solution

*Procedure type: Two sample t-test*

$H_o: \mathbf{m}_L = \mathbf{m}_G$  Where  $\mathbf{m}_L$  represents the mean score of the tests in the class where lecture only  
 $H_a: \mathbf{m}_L \neq \mathbf{m}_G$  was used and  $\mathbf{m}_G$  represents the mean score of the tests in the class where lecture and group work were both used.

#### **Notes**

*The response variable, scores of individual students, is numerical, and there are two independent groups, classes with lecture only and classes with both lecture and group work. This leads us to conclude it is a difference of means two-sample t problem. The teacher is looking to see if one method is different from the other which would indicate a two-tailed test.*

*There are several problems with the design of this experiment. First is that students are not randomly assigned to the two classes. There may be differences between the students in the two classes which would affect the results. Also, since the same teacher teaches both classes, the classes must meet at different times. Students in classes which meet at a later time in the day may respond differently than students meeting earlier in the day. Another problem with the design is that there is no true replication in the experiment. Since the students are all in the same class and therefore not independent of each other, only the average of the class can be considered the response variable. To get replication one must have several classes. Another possible problem is that the teacher may hate group work which may have an effect on the results. Because of the flaws in the design of the experiment, results from the hypothesis test may not be valid.*

2. A student of political science wished to determine whether there is a statistically significant relationship between the gender of a student and their political affiliation. A simple random sample of forty-five students was selected from the school and asked on a

survey whether their political affiliation was more likely to be liberal, conservative or moderate. Students were also categorized as male or female. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

**Solution**

**Procedure type:** Chi-Square test of independence

$H_o$  : Gender and political affiliation are independent

$H_a$  : Gender and political affiliation are not independent

**Notes**

The data are categorical, so displaying the data using a two-way table would be appropriate. To test a relationship or an association between categorical data that form a two-way table, one uses the Chi-Square test of independence.

The sampling design appears appropriate. One must be careful to state the conclusions for this school only and not the population of all high school students.

3. A student was interested in whether an on-line SAT review course increased students' math scores. To answer his question he randomly selects 30 students from the population of all the students who took the course and records their SAT scores before the course and after finishing the course. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

**Solution**

**Procedure type:** Two sample t-test

$H_o$  :  $\mathbf{m}_d = 0$

$H_a$  :  $\mathbf{m}_d < 0$

Where  $\mathbf{m}_d$  is the mean difference between the before SAT scores and after SAT scores. (SAT before – SAT after)

**Notes**

Because students were taking the course individually we can assume that the before – after differences are independent. This is a paired design because a difference between a before and after SAT score was taken for each student. The alternative hypothesis is one sided because we hope to show that the SAT course improved scores.

*The design of the experiment is appropriate because each student is taking the course independently and they were randomly selected from a list of all students who took the course.*

4. A student wishes to know if SUV drivers at his school are more likely to be male than female. He randomly selects 50 students from a large list who are SUV drivers and records their gender. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

### **Solution**

**Procedure type:** *Large sample test for proportion*

$$\begin{aligned} H_o: p &= .5 \\ H_a: p &> .5 \end{aligned} \quad \text{Where } p \text{ is the proportion of SUV-driving students who are male}$$

### **Notes**

*Students may confuse this example with difference of proportions ( $H_o: p_m = p_f$  where  $p_f$  is the proportion of females and  $p_m$ ). This would be incorrect because the males and females are being selected from the same 50 people who are all SUV drivers. The proportion of males is simply  $1 -$  the proportion of females. The difference of proportion test stated above requires that the two proportions come from different populations.*

*Since we are asking whether SUV drivers are more likely to be male we must test the hypothesis that the proportion of males is 50% (anything greater than 50% would be a majority) against the alternative hypothesis that the proportion of males is more than 50%. This requires students to use a one-tailed test.*

*If the “large list” includes all SUV drivers and since a Simple Random Sample was used to select students from this list, the design of the study would be appropriate. Again, students should be reminded that conclusions should be drawn only about SUV drivers at this particular school.*

5. Your friend in Portland claims that it seems like lots of drivers who pass her while she awaits the school bus are talking on a cell phone. You think it’s a worse problem in your hometown. One randomly selected day, you and your friend agree to keep track of the number of cell phone users and total cars that pass by you between 7 and 7:30 a.m. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

**Solution**

**Procedure type:** two-proportion z test

$H_0 : p_p = p_h$ , where  $p_p$  is the proportion of drivers in Portland who are talking on a cell phone and  $p_h$  is the proportion of drivers in your hometown who are talking on a cell phone

$H_a : p_p < p_h$

**Notes**

We have two independent samples here, the drivers in Portland and the drivers in your hometown. You want to see whether your hometown has proportionately more cell-phone-using drivers than Portland, so the test is one-sided.

Students may mention that perhaps the drivers passing the two bus stops are not representative of all drivers in the two cities, and students may also be uncomfortable with using **all** the cars in the chosen half hour. A random sample of cars on the road at that time would be better, but it may be extremely difficult to obtain. Students should also note that the results could only be applied to that particular time of day.

6. In your psychology class, your group (5 students) wants to investigate the relative intelligence of mice. You claim that male mice are more intelligent than female mice, but your group members disagree. You decide to perform an experiment on mice, using mazes. Each of you has one male and one female mouse at home (for a total of 10 mice), and you each build a different maze. Each of you will allow each mouse one trial and record the time to reach the cheese at the end of the maze. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

**Solution**

**Procedure type:** paired t-test

$H_0 : \mu = 0$ , where  $\mu$  is the mean difference in times for the five mazes, time for male mouse – time for female mouse

$H_a : \mu < 0$

**Notes**

The key here is to recognize the question seeks to test the claim that male mice are smarter than female mice, and the researchers are measuring intelligence using maze completion times. The five mazes are different, so we must compare the times for the two mice for each

*maze separately. If we treated this as a two-sample t-test, we would ignore the fact that the five times came from different mazes. The mazes may differ in difficulty, for instance.*

*Students should recognize that the sample size here is very small (5). A better design may be to use one maze and a larger number of mice. That would lead to a two-sample t-test.*

7. A high school physical education teacher is curious whether doing plyometrics three days a week will help athletes sprint faster. On a particular day of class in the fall, the teacher times all the students in a 40-yard sprint. For the next six weeks, the students perform 10 minutes of plyometric exercises three days a week. At the end of the six weeks, the teacher times all the students in another 40-yard sprint. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

### **Solution**

***Procedure type:*** *paired t-test*

$H_0 : \mathbf{m} = 0$ , where  $\mu$  is the mean difference in 40-yard sprint times for the students,  
*final time – initial time*

$H_a : \mathbf{m} < 0$

### ***Notes***

*Here each student has two sprint times, and all students undergo the training. Thus this is a matched pairs scenario, and the data will be the differences in the students' times. Because the different students may have very different running speeds, we cannot ignore those inherent differences. A two-sample t-test would cause us to lose that piece of information.*

*In terms of the design, students may comment that including another class that does not do the plyometrics would be a good idea, because perhaps some other class activity leads to improvement in sprint times. In addition, we do not know how the students were assigned to this class. Are they all varsity athletes? Are they all healthy at the start of the study? Many variables such as these are important but not commented upon.*

8. Your friend studies German and you study French. Your friend is always complaining about how long the words are, but you think French words are longer, on average, than German words. You decide to determine which student is correct. You each find a different novel in the appropriate language, randomly choose a page in the novel, and count the lengths of all the words on the page. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

**Solution**

**Procedure type:** two-sample *t*-test

$H_0 : \mathbf{m}_G = \mathbf{m}_F$ , where  $\mu_G$  is the mean length of a German word in the chosen novel and  $\mu_F$  is the mean length of a French word in the chosen novel

$H_a : \mathbf{m}_G < \mathbf{m}_F$

**Notes**

*The data for this problem would be two sets of word lengths. The two samples are independent, because they are taken from two different novels. This cannot be a matched pairs problem, because there is no guarantee that the sample sizes are equal and there is no pairing in the design.*

*Students may comment that an improved design would be to take a simple random sample of the same number of words in each novel, but that would be extremely difficult in practice. They may also comment that the two writing selections should be of equivalent reading level and subject matter. As the problem stands the conclusion can only apply to those two novels.*

9. Researchers have noted that sleep deprivation leads to car accidents and other mistakes, often due to inattention or slower reaction time. In order to examine the level of sleep deprivation in high school students, a researcher performs the following study. At 10 a.m. on a particular school day, students in two classes play a computer game that is actually recording the time it takes them to negotiate a mental obstacle course. At 2 p.m. that day, one of the classes is given 30 minutes in a silent, dark room with comfortable furniture, and the students are allowed to sleep. The other class has regular classes. At 3 p.m., both classes play the computer game again. The researcher records the differences in the times it takes each student to complete the game. State the hypotheses for the appropriate test and identify the inference procedure you would use. Justify your response and include comments on the design of the study.

**Solution**

**Procedure type:** two-sample *t*-test

$H_0 : \mathbf{m}_S = \mathbf{m}_N$ , where  $\mu_S$  is the mean difference in game time (time at 3 p.m. – time at 10 a.m.) for the students who took naps and  $\mu_N$  is the mean difference in game time (time at 3 p.m. – time at 10 a.m.) for the students who did not take naps

$H_a : \mathbf{m}_S < \mathbf{m}_N$

### **Notes**

*The data for this problem would be two sets of differences in game times. The two samples are the two classes of students. Although the data are differences (and may lead students to assume this is a matched pairs problem), the study seeks to determine the effect of a nap on students' ability to negotiate a mental obstacle course, not how each student's time changes during the course of the school day.*

*In terms of design, students might note some important issues. First, the students in the study were not randomly assigned to the classes. In addition, we do not know if the students all had the same experiences during the time between 10 and 2 (or 3). Some students may have had a physical education class that tired them out. Some students may have skipped lunch and others may have gorged themselves. Some students may have taken a grueling test and tired themselves out mentally. These are some of the issues students can comment upon.*

## PART 4

### NOTES ON INFERENCE ON THE 2001 AP EXAM

In looking at the 2001 exam we find that two questions, question #2 (the two copier question) and question #5 (the name brand and generic brand prescription problem) have similar wording in the directions but require students to perform very different tasks. In question #2 students are asked, “*Give a statistical justification to support your recommendation,*” and question #5 asked students, “*Give appropriate statistical evidence to support your response.*” In reading these two statements students may think that a hypothesis test is required for both problems when this in fact is not the case.

#### Analysis of question #5

In question #5 students should be aware that the data are a sample from the population. The problem asks to report findings for the population, thus suggesting inference. Also, students need to interpret the third sentence in the problem; “*An independent consumer advocacy group wanted to determine if there was a difference, in milligrams, in the amount of active ingredient between a certain “name” brand drug and its generic counterpart.*” This is implying that we are looking at a difference. Looking at the data values we find they are numerical as opposed to categorical which leads us to look at a t-test. Finally students need to recognize that it is a paired t-test because the observations are paired by pharmacy.

#### Analysis of question #2

Question #2 asks students, “*give a recommendation based on overall cost as to which machine, A or B, should be purchased.*” The given information is shown as a probability distribution as opposed to a list of categorical or numerical data. Since we do not have data, nor do we have the summary statistics for the raw data, no inference can be done or should be done.

#### Summary

As this analysis has shown, it is crucial that students carefully read the entire question before writing a solution. Often students will embark on a course of action based on what the data look like or a single key word in the problem. This may lead students to use a test of significance when none is required or to use the wrong test of significance. Looking at the presentation of the data and seeking out key words are useful but should not take the place of analyzing the problem as a whole.