AP FRQ Review – Mr. Rich Name:_________________________________ Per:_____ Seat:____

AP Stats Topic 7: Chi-Square Hypothesis Testing

Product advertisers studied the effects of television ads on children's choices for two new snacks. The advertisers used two 30-second television ads in an experiment. One ad was for a new sugary snack called Choco-Zuties, and the other ad was for a new healthy snack called Apple-Zuties.

For the experiment, 75 children were randomly assigned to one of three groups, A, B, or C. Each child individually watched a 30-minute television program that was interrupted for 5 minutes of advertising. The advertising was the same for each group with the following exceptions.

- The advertising for group A included the Choco-Zuties ad but not the Apple-Zuties ad.
- The advertising for group B included the Apple-Zuties ad but not the Choco-Zuties ad.
- The advertising for group C included neither the Choco-Zuties ad nor the Apple-Zuties ad.

After the program, the children were offered a choice between the two snacks. The table below summarizes their choices.

Group	Type of Ad	Number Who Chose Choco-Zuties	Number Who Chose Apple-Zuties
A	Choco-Zuties only	21	4
B	Apple-Zuties only	13	12
C	Neither	22	3

(a) Do the data provide convincing statistical evidence that there is an association between type of ad and children's choice of snack among all children similar to those who participated in the experiment?

(b) Write a few sentences describing the effect of each ad on children's choice of snack.

Part (a):

Step 1: States a correct pair of hypotheses.

H_0 : The proportion of children who would choose each snack is the same regardless of which type of ad is viewed.

H_a : The proportion of children who would choose each snack differs based on which type of ad is viewed.

Step 2: Identifies a correct test procedure (by name or formula) and checks appropriate conditions.

The appropriate procedure is a chi-square test of homogeneity.

The conditions for this test are satisfied because (1) the question states that the children were randomly assigned to groups, and (2) expected counts for the six cells of the table are all at least 5, as seen in the following table that lists expected counts beside observed counts.

Group	Choco-Zuties	Apple-Zuties	Total
A	21 (18.67)	4 (6.33)	25
B	13 (18.67)	12 (6.33)	25
C	22 (18.67)	3 (6.33)	25
Total	56	19	75

Step 3: Calculates the appropriate test statistic and *p*-value.

The test statistic is calculated as $\chi^2 = \sum \frac{(O-E)^2}{E}$, which is

$$\chi^2 \approx \\ 0.292 + 0.860 + \\ 1.720 + 5.070 + \\ 0.595 + 1.754 \approx 10.291.$$

The *p*-value is $P(\chi^2_{\text{df2}} \geq 10.291) \approx 0.006$.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the *p*-value is very small (for instance, much smaller than $\alpha = 0.05$), we reject the null hypothesis at the 0.05 level (and at the 0.01 level). The data provide convincing statistical evidence that the proportions who would choose each snack differ based on which ad is viewed.

Part (b):

When neither ad was viewed, $\frac{22}{25}$ or 88 percent of the children chose Choco-Zuties whereas only 12 percent chose Apple-Zuties.

When the Choco-Zuties ad was viewed, 84 percent of the children chose Choco-Zuties, which was very similar to the percentage that chose them without viewing any ad. So watching the Choco-Zuties ad did not affect the snack choice very much.

When the Apple-Zuties ad was viewed, only $\frac{13}{25}$ or 52 percent of the children chose Choco-Zuties, and 48 percent chose Apple-Zuties. Watching the Apple-Zuties ad seemed to increase the proportion of children choosing Apple-Zuties.

The Behavioral Risk Factor Surveillance System is an ongoing health survey system that tracks health conditions and risk behaviors in the United States. In one of their studies, a random sample of 8,866 adults answered the question "Do you consume five or more servings of fruits and vegetables per day?" The data are summarized by response and by age-group in the frequency table below.

Age-Group (years)	Yes	No	Total
18–34	231	741	972
35–54	669	2,242	2,911
55 or older	1,291	3,692	4,983
Total	2,191	6,675	8,866

Do the data provide convincing statistical evidence that there is an association between age-group and whether or not a person consumes five or more servings of fruits and vegetables per day for adults in the United States?

Solution

Step 1: States a correct pair of hypotheses.

The null hypothesis is that fruit and vegetable consumption is independent of (that is, it is not associated with) age group for the population of adults in the United States.

The alternative hypothesis is that fruit and vegetable consumption is not independent of (that is, it is associated with) age group for the population of adults in the United States.

Step 2: Identifies a correct test procedure (by name or by formula) and checks appropriate conditions.

The appropriate test is a chi-square test of independence.

The conditions for this test were satisfied because:

1. The question states that the sample was randomly selected.
2. The expected counts for all six cells of the table were all at least 5, as seen in the following table that lists expected counts in parentheses beside the observed counts:

	Five or more servings of fruit and vegetables	Four or fewer servings of fruit and vegetables	Total
18–34 years	231 (240.2)	741 (731.8)	972
35–54 years	669 (719.4)	2242 (2191.6)	2911
55+ years	1291 (1231.4)	3692 (3751.6)	4983
Total	2191	6675	8866

Step 3: Correct mechanics, including the value of the test statistic and p-value (or rejection region).

The test statistic is calculated from $\chi^2 = \sum \frac{(O-E)^2}{E}$; that is,

$$\chi^2 = 0.353 + 0.116 + 3.528 + 1.158 + 2.883 + 0.946 = 8.983.$$

The p-value is $P(\chi^2 \geq 8.983) = 0.011$, based on $(3-1) \times (2-1) = 2$ degrees of freedom.

Step 4: States a correct conclusion in the context of the study, using the result of the statistical test.

Because the p-value is very small (for instance, much smaller than $\alpha = 0.05$), we would reject the null hypothesis at the 0.05 level and conclude that the sample data provide strong evidence that there is an association between age group and consumption of fruits and vegetables for adults in the United States. In particular, older (55+ years of age) people were more likely to eat five or more servings of fruits and vegetables, and middle-aged people (35–54 years of age) were less likely to eat five or more servings of fruits and vegetables.

A random sample of 200 students was selected from a large college in the United States. Each selected student was asked to give his or her opinion about the following statement.

> "The most important quality of a person who aspires to be the President of the United States is a knowledge of foreign affairs."

Each response was recorded in one of five categories. The gender of each selected student was noted. The data are summarized in the table below.

	Response Category				
	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Male	10	15	15	25	25
Female	20	25	25	25	15

Is there sufficient evidence to indicate that the response is dependent on gender? Provide statistical evidence to support your conclusion.

Part 1: States a correct pair of hypotheses

H_o: Response and gender are independent
H_a: Response and gender are not independent
OR
H_o: There is no association between response and gender
H_a: There is an association between response and gender

Part 2: Identifies a correct test (by name or by formula) and checks appropriate conditions.

Chi-Square test (for independence)

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

Conditions: Random sample and large sample size

Expected counts are

	Strongly Disagree	Somewhat Disagree	Neither Agree or Disagree	Somewhat Agree	Strongly Agree
Male	13.5	18.0	18.0	22.5	18.0
Female	16.5	22.0	22.0	27.5	22.0

All expected counts are greater than 5 (or 10), so the sample size is large enough for the Chi-Square test to be appropriate.

(Or, all expected counts are ≥ 1, and no more than 20% of expected counts < 5.)

Part 3: Correct mechanics, including the value of the test statistic, df, and P-value (or rejection region)

χ^2=0.907+0.500+0.500+0.278+2.722+0.742+0.409+0.409+0.227+2.227 = 8.921

df = 4   P-value = 0.063

(Or, using tables, 0.05 < P-value < 0.10, or rejection regions: $\alpha = 0.05$ is 9.48, $\alpha = 0.01$ is13.27)

Part 4: Stating a correct conclusion in the context of the problem, using the result of the statistical test.

Because P-value > selected α (or because χ^2 is not in the rejection region, or because the P-value is large), fail to reject the null hypothesis. There is not sufficient evidence to conclude that response is dependent on gender (or that response and gender are not independent, or that response and gender are associated)

OR

Because results this extreme would occur about 6 times in 100 by chance alone, there is marginal evidence to reject the null hypothesis and conclude that there is an association between response and gender.

The Colorado Rocky Mountain Rescue Service wishes to study the behavior of lost hikers. If more were known about the direction in which lost hikers tend to walk, then more effective search strategies could be devised. Two hundred hikers selected at random from those applying for hiking permits are asked whether they would head uphill, downhill, or remain in the same place if they became lost while hiking. Each hiker in the sample was also classified according to whether he or she was an experienced or novice hiker. The resulting data are summarized in the following table.

	Direction		
	Uphill	Downhill	Remain in Same Place
Novice	20	50	50
Experienced	10	30	40

Do these data provide convincing evidence of an association between the level of hiking expertise and the direction the hiker would head if lost?

Give appropriate statistical evidence to support your conclusion.

Solution:

H_o: There is no association between level of hiking experience and direction
H_a: There is an association between level of hiking experience and direction
or
H_o: Level of hiking experience and direction are independent
H_a: Level of hiking experience and direction are not independent

Chi-Square test for independence

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

Requirements: Need to check expected counts using some accepted rule (textbooks differ).
Table of Observed and (Expected Counts):

20 (18)	50 (48)	50 (54)
10 (12)	30 (32)	40 (36)

Requirements Check: All expected counts are greater than or equal to 5. So the Chi-Square test of independence is appropriate.

$$x^2 = \frac{(20 - 18)^2}{18} + \frac{(50 - 48)^2}{48} + \frac{(50 - 54)^2}{54} + \frac{(10 - 12)^2}{12} + \frac{(30 - 32)^2}{32} + \frac{(40 - 36)^2}{36} = 1.5046$$

df = 2, P-value = 0.471 **OR** (using tables) P-value > 0.25
Since the P-value is large for any reasonable level of significance, we fail to reject H_o. (A picture showing an appropriate rejection region and the test statistic value is acceptable). There is not convincing evidence that an association exists between level of hiking expertise and direction.

Note:
For rejection region approach, rejection regions are
$x^2 > 9.21$ or significance level 0.01,
$x^2 > 5.99$ for significance level 0.05,
$x^2 > 4.61$ for significance level 0.10

A rural county hospital offers several health services. The hospital administrators conducted a poll to determine whether the residents' satisfaction with the available services depends on their gender. A random sample of 1,000 adult county residents was selected. The gender of each respondent was recorded and each was asked whether he or she was satisfied with the services offered by the hospital. The resulting data are shown in the table below.

	Male	Female	Total
Satisfied	384	416	800
Not Satisfied	80	120	200
Total	464	536	1,000

(a) Using a significance level of 0.05, conduct an appropriate test to determine if, for adult residents of this county, there is an association between gender and whether or not they were satisfied with services offered by the hospital.

(b) Is $\frac{800}{1,000}$ a reasonable estimate for the proportion of all adult county residents who are satisfied with the services offered by this hospital? Explain why or why not.

Part a:

H_0: gender and satisfaction with health services offered by the hospital are independent (OR not associated)

H_a: gender and satisfaction with health services offered by the hospital are dependent (OR associated)

Chi-square test for association

Test statistic: $\chi^2 = \sum_{all\,cells} \frac{(\text{observed - expected})^2}{\text{expected}}$

Conditions: A random sample has been taken. The expected cell counts are large enough so that the chi-square approximation can be used. (See the table below for the expected cell counts.) That is, all four of the expected cell counts are at least 5 (or the smallest expected cell count is 92.8 which is greater than 5). We can use the chi-square approximation.

```
Expected counts are printed below observed counts

          Male   Female    Total
    1      384      416      800
        371.20   428.80



    2       80      120      200
         92.80   107.20

Total      464      536     1000

Chi-Sq =  0.441 +  0.382 + 1.766 +  1.528 = 4.117
DF = 1, P-Value = 0.042
```

Because the p-value, 0.042, is less than 0.05, we can reject H_0 at significance level 0.05, and conclude that there is evidence of an association between gender and satisfaction with health services offered by the hospital for adult residents of this county.

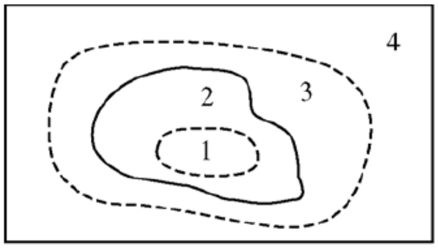
Part b:

Because a random sample has been taken from the population of all county residents, 0.8 is a reasonable estimate for the proportion of all county residents who are satisfied with the services offered by this hospital.

A study was conducted to determine where moose are found in a region containing a large burned area. A map of the study area was partitioned into the following four habitat types.

(1) Inside the burned area, not near the edge of the burned area,
(2) Inside the burned area, near the edge,
(3) Outside the burned area, near the edge, and
(4) Outside the burned area, not near the edge.

The figure below shows these four habitat types.



Note: Figure not drawn to scale.

The proportion of total acreage in each of the habitat types was determined for the study area. Using an aerial survey, moose locations were observed and classified into one of the four habitat types. The results are given in the table below.

Habitat Type	Proportion of Total Acreage	Number of Moose Observed
1	0.340	25
2	0.101	22
3	0.104	30
4	0.455	40
Total	1.000	117

(a) The researchers who are conducting the study expect the number of moose observed in a habitat type to be proportional to the amount of acreage of that type of habitat. Are the data consistent with this expectation? Conduct an appropriate statistical test to support your conclusion. Assume the conditions for inference are met.

(b) Relative to the proportion of total acreage, which habitat types did the moose seem to prefer? Explain.

Part (a):

Step 1: States a correct pair of hypotheses.

H_0 : Moose have no preference for habitat type.
H_a : Moose have a preference for habitat type.

OR

H_0 : The number of moose in each habitat type is proportional to the amount of acreage of that habitat type.
H_a : The number of moose in at least one habitat type is not proportional to the amount of acreage of that habitat type.

OR

$H_0 : p_1 = 0.340, p_2 = 0.101, p_3 = 0.104, p_4 = 0.455$, where p_i = the proportion of moose in habitat type *i*.
H_a : At least one of these proportions is incorrect.

Step 2: Identifies a correct test (by name or formula) and checks appropriate conditions.

- Chi-square goodness-of-fit test (or test for more than two proportions)
$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$
- The stem of the problem stated that conditions for inference are met.

Step 3: Correct mechanics, including the value of the test statistic, df, and *p*-value (or rejection region).

- The test statistic, with df = 4 − 1 = 3, is
$$\chi^2 = \frac{(25-39.780)^2}{39.780} + \frac{(22-11.817)^2}{11.817} + \frac{(30-12.168)^2}{12.168} + \frac{(40-53.235)^2}{53.235} = 43.6893 .$$
- The *p*-value is $P(\chi_3^2 \geq 43.6893) < 0.0005$ (a calculator gives the *p*-value as 1.7569×10^{-9}).

Step 4: States a correct conclusion in the context of the problem, using the result of the statistical test.

The data are not consistent with the researchers' expectation. Because the *p*-value is less than $\alpha = 0.05$, we reject H_0. There is strong evidence that moose have a preference for habitat type.

OR

The data are not consistent with the researchers' expectation. If the null hypothesis is true and the number of moose in each of the habitat types is proportional to the acreage in that habitat type, then we would observe a test statistic of 43.69 or one more extreme less than 0.05 percent of the time. There is strong evidence that moose have a preference for habitat type.

Part (b):

The moose seem to prefer habitat types 2 and 3. Relative to the proportion of total acreage, a higher proportion of moose were observed in each of these habitat types than expected. In habitat types 1 and 4, the observed proportion of moose was less than the expected proportion of moose, indicating that these two habitat types are less desirable.

OR

Habitat type 3 seems to be the most preferred—it has a positive difference between the observed (30) and expected (12.168) counts of moose and the largest contribution to the chi-square statistic (26.1325). Alternatively, habitat type 3 has the largest positive difference between the observed proportion of moose (0.256) and the expected proportion of moose (0.104).

A parent advisory board for a certain university was concerned about the effect of part-time jobs on the academic achievement of students attending the university. To obtain some information, the advisory board surveyed a simple random sample of 200 of the more than 20,000 students attending the university. Each student reported the average number of hours spent working part-time each week and his or her perception of the effect of part-time work on academic achievement. The data in the table below summarize the students' responses by average number of hours worked per week (less than 11, 11 to 20, more than 20) and perception of the effect of part-time work on academic achievement (positive, no effect, negative).

		Average Time Spent on Part-Time Jobs		
		Less Than 11 Hours per Week	11 to 20 Hours per Week	More Than 20 Hours per Week
Perception of the Effect of Part-Time Work on Academic Achievement	Positive Effect	21	9	5
	No Effect	58	32	15
	Negative Effect	18	23	19

A chi-square test was used to determine if there is an association between the effect of part-time work on academic achievement and the average number of hours per week that students work. Computer output that resulted from performing this test is shown below.

```
CHI-SQUARE TEST
Expected counts are printed below observed counts

              <11     11–20     >20     Total

Positive       21         9       5        35
           16.975    11.200   6.825

No effect      58        32      15       105
           50.925    33.600  20.475

Negative       18        23      19        60
           29.100    19.200  11.700

Total          97        64      39       200

Chi-Sq = 13.938, DF = 4, P-Value = 0.007
```

(a) State the null and alternative hypotheses for this test.

(b) Discuss whether the conditions for a chi-square inference procedure are met for these data.

(c) Given the results from the chi-square test, what should the advisory board conclude?

(d) Based on your conclusion in part (c), which type of error (Type I or Type II) might the advisory board have made? Describe this error in the context of the question.

Solution

Part (a):

H_0: There is no association between perceived effect of part-time work on academic achievement and average time spent on part-time jobs.
H_a: There is an association between perceived effect of part-time work on academic achievement and average time spent on part-time jobs.

Part (b):

The following conditions for inference are met:

1. The students were randomly selected.
2. The expected cell counts should be at least 5. The computer output indicates that all expected counts are greater than 5. The smallest expected cell count is 6.825.

Part (c):

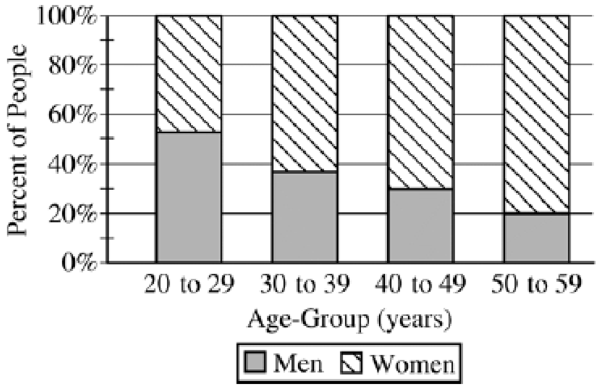
Because the *p*-value 0.007 is less than 0.05, H_0 should be rejected. There is convincing evidence that there is an association between the perceived effect of part-time work on academic achievement and average time spent on part-time jobs.

Part (d):

Because the null hypothesis was rejected, a Type I error may have been made. A Type I error is concluding that there is an association between the perceived effect of part-time work on academic achievement and the average time spent on part-time jobs when, in reality, there is no association between the two variables.

The table and the bar chart below summarize the age at diagnosis, in years, for a random sample of 207 men and women currently being treated for schizophrenia.

	Age-Group (years)				
	20 to 29	30 to 39	40 to 49	50 to 59	Total
Women	46	40	21	12	119
Men	53	23	9	3	88
Total	99	63	30	15	207



Do the data provide convincing statistical evidence of an association between age-group and gender in the diagnosis of schizophrenia?

Solution

Step 1: State a correct pair of hypotheses.

The null hypothesis is that age group at diagnosis and gender are independent (that is, they are not associated) for the population of people currently being treated for schizophrenia.

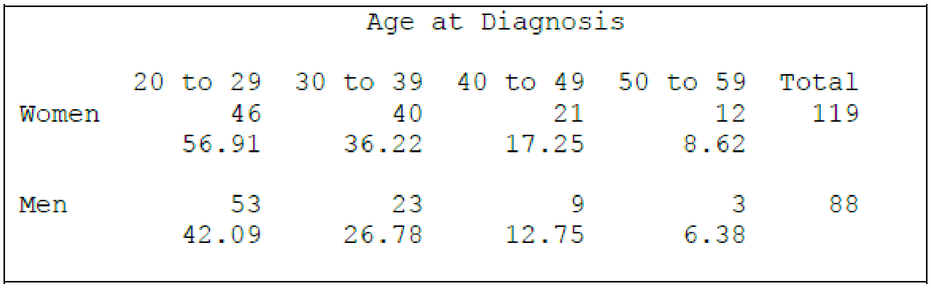
The alternative hypothesis is that age group at diagnosis and gender are not independent for the population of people currently being treated for schizophrenia.

Step 2: Identify a correct test procedure (by name or formula) and check appropriate conditions.

The appropriate test is a chi-square test of independence.

The conditions for this test are satisfied because:

1. The question states that the sample was randomly selected.
2. The expected counts for the eight cells of the table are at least 5, as seen in the following table, with expected counts shown below observed counts.



```
                          Age at Diagnosis

       20 to 29  30 to 39  40 to 49  50 to 59  Total
Women        46        40        21        12    119
          56.91     36.22     17.25      8.62

Men          53        23         9         3     88
          42.09     26.78     12.75      6.38
```

Step 3: Find the value of the test statistic and the p-value.

The test statistic is calculated as $\chi^2 = \sum \frac{(O-E)^2}{E}$, or

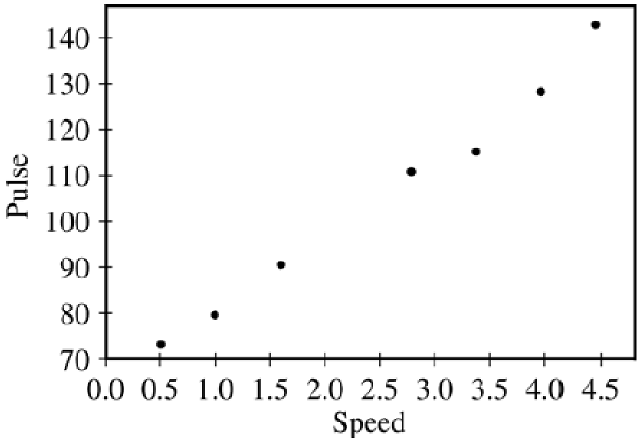
$$\begin{aligned}\chi^2 &= 2.093 + 0.395 + 0.817 + 1.322 \\ &\quad + 2.830 + 0.534 + 1.105 + 1.788 \\ &= 10.884.\end{aligned}$$

The p-value is $P(\chi^2 \geq 10.884) = 0.012,$ based on $(4-1) \times (2-1) = 3$ degrees of freedom.

Step 4: State the conclusion in context, with linkage to the p-value.

Because the p-value is very small (for instance much smaller than $\alpha = 0.05$), we reject the null hypothesis and conclude that the sample data provide strong evidence that there is an association between age group at diagnosis and gender for the population currently being treated for schizophrenia.

John believes that as he increases his walking speed, his pulse rate will increase. He wants to model this relationship. John records his pulse rate, in beats per minute (bpm), while walking at each of seven different speeds, in miles per hour (mph). A scatterplot and regression output are shown below.



```
Regression Analysis:  Pulse Versus Speed

Predictor         Coef          SE Coef              T            P
Constant          63.457        2.387                26.58        0.000
Speed             16.2809       0.8192               19.88        0.000

S = 3.087         R-Sq = 98.7%          R-Sq (adj) = 98.5%

Analysis of Variance

Source            DF            SS           MS           F             P
Regression        1             3763.2       3763.2       396.13        0.000
Residual          5               47.6          9.5
Total             6             3810.9
```

(a) Using the regression output, write the equation of the fitted regression line.

(b) Do your estimates of the slope and intercept parameters have meaningful interpretations in the context of this question? If so, provide interpretations in this context. If not, explain why not.

(c) John wants to provide a 98 percent confidence interval for the slope parameter in his final report. Compute the margin of error that John should use. Assume that conditions for inference are satisfied.

Solution

Part (a):

Predicted Pulse = 63.457 + 16.2809 (Speed)

Part (b):

The intercept (63.457 bpm) provides an estimate for John's mean resting pulse (walking at a speed of zero mph).

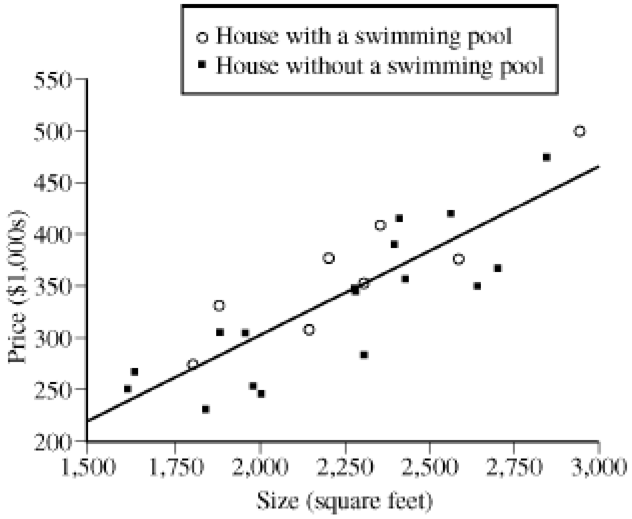
The slope (16.2809 bpm/mph) provides an estimate for the mean increase in John's heart rate as his speed is increased by one mile per hour.

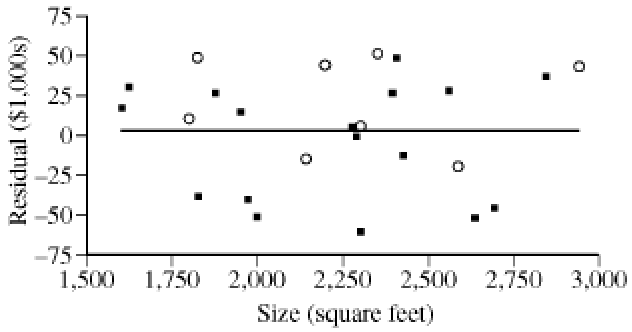
Part (c):

The margin of error for the confidence interval for the slope parameter is $t_{n-2}^* \times s_b$, where s_b is the standard error of the slope parameter. For a 98% confidence interval, the margin of error is $3.365 \times 0.8192 = 2.7566$ bpm.

A real estate agent is interested in developing a model to estimate the prices of houses in a particular part of a large city. She takes a random sample of 25 recent sales and, for each house, records the price (in thousands of dollars), the size of the house (in square feet), and whether or not the house has a swimming pool. This information, along with regression output for a linear model using size to predict price, is shown below and on the next page.

Price ($1,000s)	Size (square feet)	Pool	Residual ($1,000s)
274	1,799	yes	6
330	1,875	yes	49
307	2,145	yes	–18
376	2,200	yes	42
352	2,300	yes	1
409	2,350	yes	50
375	2,589	yes	–23
498	2,943	yes	42
248	1,600	no	13
265	1,623	no	26
228	1,829	no	–45
303	1,875	no	22
303	1,950	no	10
251	1,975	no	–46
244	2,000	no	–57
347	2,274	no	1
345	2,279	no	–2
282	2,300	no	–69
389	2,392	no	23
413	2,410	no	44
353	2,428	no	–19
419	2,560	no	26
348	2,639	no	–58
365	2,701	no	–52
474	2,849	no	33





Linear Fit

Price = –28.144 + 0.165 Size

Summary of Fit

RSquare 0.722

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–28.144	48.259	–0.58	0.5654
Size	0.165	0.0213	7.72	<.0001

(a) Interpret the slope of the least squares regression line in the context of the study.

(b) The second house in the table has a residual of 49. Interpret this residual value in the context of the study.

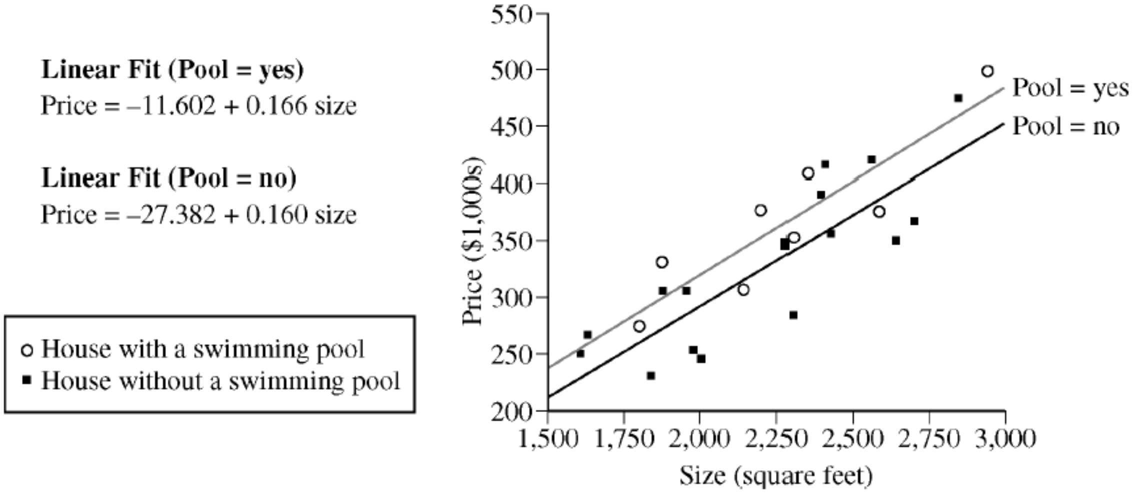
The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(c) Use the residuals from all 25 houses to estimate how much greater the price for a house with a swimming pool would be, on average, than the price for a house of the same size without a swimming pool.

To further investigate the effect of having a swimming pool on the price of a house, the real estate agent creates two regression models, one for houses with a swimming pool and one for houses without a swimming pool. Regression output for these two models is shown below.

Linear Fit (Pool = yes)
Price = –11.602 + 0.166 size

Linear Fit (Pool = no)
Price = –27.382 + 0.160 size



(d) The conditions for inference have been checked and verified, and a 95 percent confidence interval for the true difference in the two slopes is $(-0.099, 0.110)$. Based on this interval, is there a significant difference in the two slopes? Explain your answer.

(e) Use the regression model for houses with a swimming pool and the regression model for houses without a swimming pool to estimate how much greater the price for a house with a swimming pool would be than the price for a house of the same size without a swimming pool. How does this estimate compare with your result from part (c) ?

Part (a):

The slope coefficient is 0.165. This means that for each additional square foot of size, the predicted price of the house increases by 0.165 thousand dollars, which is $165. In other words, this model predicts that the average price of a house increases by $165 for each additional square foot of a house's size.

Part (b):

The residual value of 49 for this house indicates that its actual price is 49 thousand dollars higher than the model would predict for a house of its size.

Part (c):

The average residual value for the eight houses with a swimming pool is:

$$\frac{(6 + 49 + (-18) + 42 + 1 + 50 + 9 + (-23) + 42)}{8} = \frac{149}{8} = 18.6 \text{ thousand dollars.}$$

The average residual value for the 17 houses with no swimming pool is:

$$\frac{(13 + 26 + (-45) + \ldots + (-58) + (-52) + 33)}{17} = \frac{-150}{17} = -8.8 \text{ thousand dollars.}$$

The residual averages suggest that the regression line tends to underestimate the price of homes with a swimming pool by about 18.6 thousand dollars and to overestimate the price of homes with no pool by about 8.8 thousand dollars. The difference between these two residual averages is $18.6 - (-8.8) = 27.4$ thousand dollars. This suggests that, for two houses of the same size, the house with a swimming pool would be estimated to cost $27,400 more than the house with no swimming pool.

Part (d):

No, this confidence interval does *not* indicate a significant difference (at the 95 percent confidence level, equivalent to the 5 percent significance level) between the two slope coefficients because the interval includes the value zero.

Part (e):

If the two population regression lines do in fact have the same slope, the impact of a swimming pool is the (constant) vertical distance between the two lines. However, because the two fitted lines do not have the same slope, the distance between the two fitted lines depends on the size of the house. Using the available information, there are two acceptable approaches to estimating the impact of having a swimming pool.

Approach 1: Use the two fitted lines to predict the price of a house with and without a pool for a particular house size. For example, using the value of size = 2,250 square feet (which is near the middle of the distribution of house sizes), we find:

Predicted price for a 2,250 square-foot house with a swimming pool =
$-11.602 + 0.166 \times 2,250 = 361.898$ thousand dollars.

Predicted price for a 2,250 square-foot house with no swimming pool =
$-27.382 + 0.160 \times 2,250 = 332.618$ thousand dollars.

The difference in these predicted prices is $361.898 - 332.618 = 29.280$ thousand dollars, which is an estimate of the impact of a swimming pool on the predicted price of a 2,250 square-foot house. This is quite similar to the estimate based on residuals in part (c).

Approach 2: Because the slopes of the two sample regression lines were judged not to be significantly different, another acceptable approach would be to use the difference in the intercepts of the two fitted lines as an estimate of the vertical distance between the two population regression lines.

The difference in the intercepts of the two fitted lines is $-11.602 - (-27.382) = 15.780$ thousand dollars, which is an estimate of the impact of a swimming pool on the predicted price of a house, assuming this difference does not change with the size of the house. This is quite different from the estimate based on residuals in part (c).