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.

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?

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.

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.

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

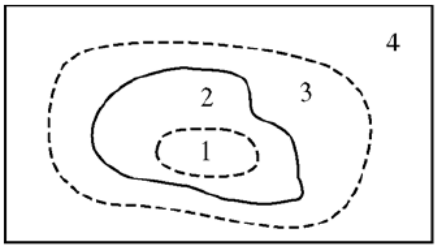
(a) Using a significance level of 0.05, conduct an appropriate test to determine if, for adult residents of this county, there is an association between gender and whether or not they were satisfied with services offered by the hospital.

(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.

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.

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 16.975	9 11.200	5 6.825	35
No effect	58 50.925	32 33.600	15 20.475	105
Negative	18 29.100	23 19.200	19 11.700	60
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.

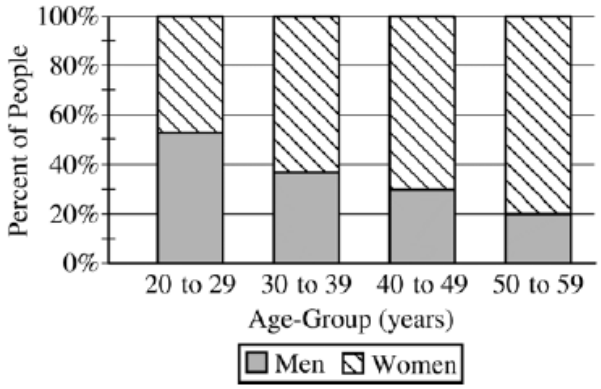
(b) Discuss whether the conditions for a chi-square inference procedure are met for these data.

(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.

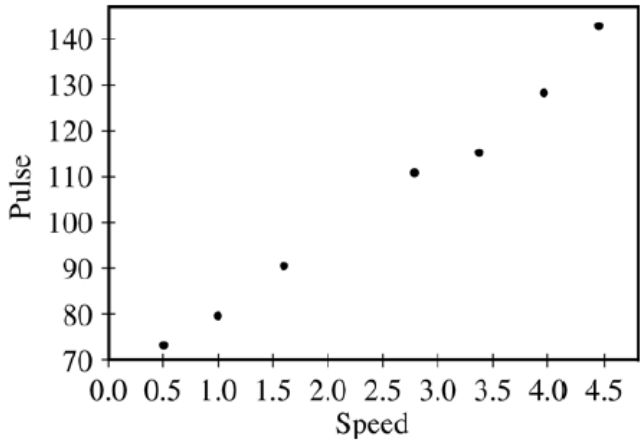
The table and the bar chart below summarize the age at diagnosis, in years, for a random sample of 207 men and women currently being treated for schizophrenia.

	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?

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			

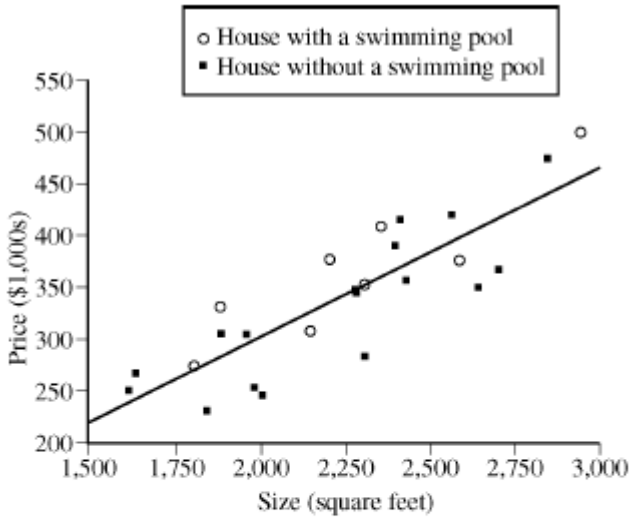
(a) Using the regression output, write the equation of the fitted regression line.

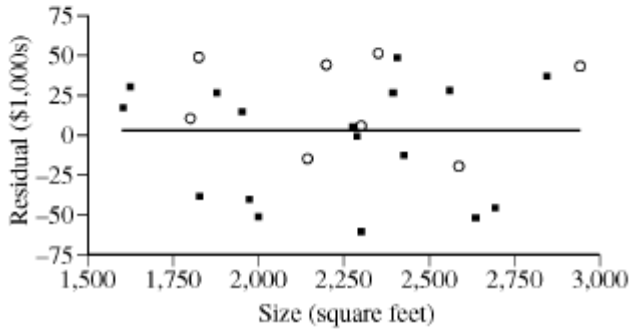
(b) Do your estimates of the slope and intercept parameters have meaningful interpretations in the context of this question? If so, provide interpretations in this context. If not, explain why not.

(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.

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.

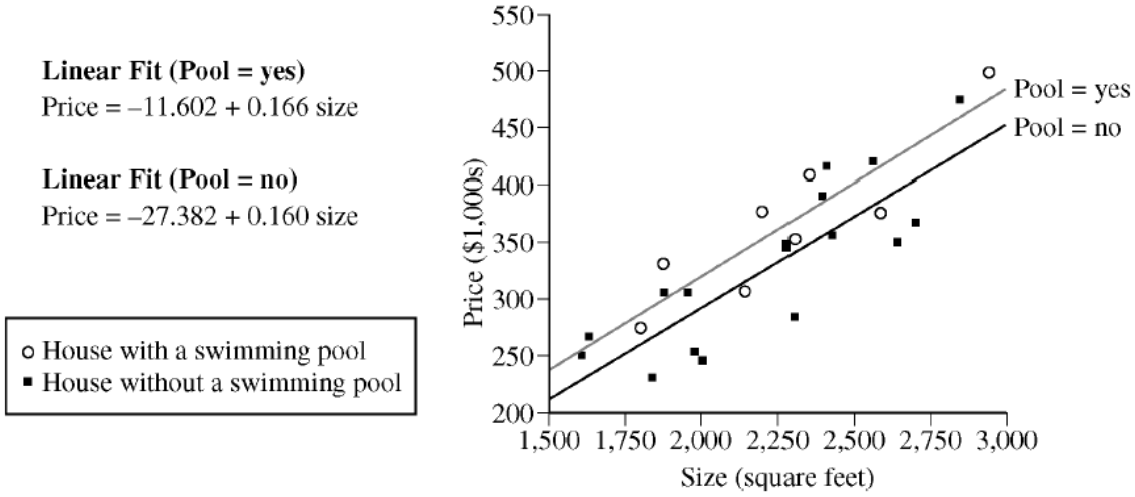
The real estate agent is interested in investigating the effect of having a swimming pool on the price of a house.

(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) ?