STAT 217
Assignment #5

Goodness of fit test.

Computer instruction
Enter observed values in column 1 and expected values in column 2.
Click on Calc<Calculator.
Type c3 in the box Store Results.
Type the formula (c1-c2)**2 /c2 in the large box. Hit enter.
Click on Calc<column statistics
Click on sum.
Type c3 in the input variable box.

1. An official of a plastics industry claimed that the industry employed 30% white women, 5% minority women, 50% white men, and 15% minority men. To test the claim, an affirmative action committee randomly sampled 150 employees and obtained the following information:

Category	observed
White females	40
Minority females	15
White males	80
Minority males	15

Test the official's claim at a 5% level of significance and find the p-value [10.89, Rho, p-value = 0.0123]]

2. A computer science major claimed to have written a program that would randomly generate integers from 1 to 100. The program generated the following data. Use a 5% level of significance to test the claim. Find the p-value [6.40, Aho, p-value = .6993]

Integers	Observed
1-10	6
11-20	6
21-30	13
31-40	9
41-50	13
51-60	11
61-70	8
71-80	12
81-90	10
91-100	12

3. Given below are the frequencies observed from 310 tosses of a die. Do theses data cast doubt on the fairness of the die at the 5% significance level? Find the p-value [13.225, Rho, p-value = .0214]

Face No.	1	2	3	4	5	6
Frequency	38	61	54	65	55	37

4. A shipment of assorted nuts is labeled as having 45% walnuts, 20% hazelnuts, 20% almonds, and 15% pistachios. By randomly picking several scoops of nuts from this shipment, an inspector find the following counts.

	Walnuts	Hazelnuts	Almonds	Pistachios	Total
Counts	92	69	32	42	235

Could these findings be a strong basis for an accusation of mislabeling? Test at the 5% significance level. Find the p-value [18.165, Rho, p-value = .0004]

Chi-Square Tests of Independence

Minitab will perform all necessary calculations for chi-square tests on contingency tables, presenting the expected values, the value of the test statistic, degrees of freedom and the p-value.

1. Enter the observed frequencies into rows and columns just as they are given in the contingency table.
2. Go to the main header and click on **STAT>Tables>Chi-Square Test**
3. In **Columns containing the tables,** enter the columns at which your contingency table is contained.
4. Click on **OK.**

1. Over the years pollsters have found that the public's confidence in big business has been closely tied to the economic climate of the country. When businesses are growing and employment is increasing public confidence is high. When the opposite occurs, public confidence is low. In one study, Harvey Kahalas (1981) explored the relationship between confidence in big business and job satisfaction. He hypothesized that there is a relationship between the level of confidence and job satisfaction and that this relationship holds true for both union and nonunion workers. To test his hypotheses he used the sample data given in the tables below:

Union Members

Job Satisfaction

Confidence in Major Corporations	Very Satisfied	Moderately Satisfied	Little dissatisfied	Very Dissatisfied
A great deal	30	19	6	6
Only some	99	77	20	9
Hardly any	38	32	14	15

NonUnion Members

Job Satisfaction

Confidence in Major Corporations	Very Satisfied	Moderately Satisfied	Little dissatisfied	Very Dissatisfied
A great deal	111	52	13	5
Only some	246	142	37	18
Hardly any	73	51	19	9

Perform a hypothesis test on each of these data sets at the 5% significance level. Does the data support Kahalas's theory? [13.359, Rho] [8.298, Aho]

2. A personnel administrator provided the following data as an example of hiring to fill 12 positions from among 40 male and 40 female applicants.

Applicant	Selected	Not Selected	Total
Male	7	33	40
Female	5	35	40

Does this sample indicate a selection bias in favour of males? Use a p-value in your conclusion. [0.392, 0.5312]

3. Applicants for public assistance are allowed an appeals process when they feel unfairly treated. At such a hearing, the applicant may choose self-representation or representation by an attorney. The appeal may result in an increase, decrease, or no change of the aid recommendation. Court records of 320 appeals cases provided the following data.

Amount of Aid

Type of Representation	Increased	Unchanged	Decreased
Self	59	108	17
Attorney	70	63	3

Are the patterns of the appeals decision significantly different between the two types of representation? Test at $\alpha = 0.05$ [15.734, Rho]

4. A survey was conducted by sampling 400 persons who were questioned regarding union membership and attitude toward decreased national spending on social welfare programs. The cross-tabulated frequency counts are presented.

	Support	Indifferent	Opposed
Union	112	36	28
NonUnion	84	68	72
Total	196	104	100

Can these observed differences be explained by chance or are there real differences of attitude between the populations of members and non-members at the 5% significance level? [27.847, Rho]

5. In a Study of possible genetic influence of parental hand preference, a sample of 400 children was classified according to each child's handedness and the handedness of the biological parents. Do these findings demonstrate an association between the handedness of parents and their biological offspring at the 5% significance level? [10.653,Rho]

Handedness of Biological Offspring

Parents' Handedness Father x Mother	Right	Left	**Total**
Right x Right	303	37	340
Right x Left	29	10	39
Left x Right	16	6	22
Total	348	52	401

6. In a genetic study of chromosome structures, 132 individuals are classified according to the type of structural chromosome aberration and carriers in their parents. The following counts are obtained.

Carrier

Type of Aberration	One Parent	Neither Parent	Total
Presumably innocuous	28	19	47
Substantially unbalanced	35	50	85
Total	63	69	132

Test the null hypothesis that type of aberration is independent of parental carrier. Use p-value [4.106, p-value = .043, Rho]

The F-distribution and Simple Linear Regression
Note: there may be slight differences in the answers due to rounding.

1. Here is a set of data showing the historic yearly rates of return in seven randomly selected years, for Stock Y and the New York Stock Exchange Index (the predictor variable).

Year	1	2	3	4	5	6	7
Stock Y	2.0%	7.9%	-6.0%	-9.5%	13.5%	7.5%	1.2%
NYSE Index	4.9%	13.0%	-2.5%	-10.6%	11.0%	14.5%	4.3%

(a) Write down the linear regression model expressing the yearly rate of return on Stock Y as a linear function of the yearly rate of return of the NYSE Index.

(b) Estimate the intercept and slope term in the model. (Note: the slope term is referred to as Stock Y's "beta", or β. This is a measure that stock analysts uses to evaluate the past performance of a stock. Stocks possessing β's greater than 1 tend to have larger expected rates of return compared to stocks with smaller β's. [βo=-1.7470, β1= 0.8332]

Minitab instructions

1. Enter Stock Y data in column C1
2. Enter NYSE Index data in column C2
3. Click on **STAT>Regression>Regression**
4. Enter C1 in the **Response** box
5. Enter C2 in the **Predictor** box
6. Click on **Graphs,** click on **residuals versus fits**, click **OK**
7. Click **OK**

You will now get a graph of the residuals vs fits for the data. There will also be a printout of the regression equation and the ANOVA table on the screen. If you want to see a scatter plot of the data with the fitted line,
Click on **STAT>Regression>Fitted Line Plot**
Enter C1 in **Response (Y)**
Enter C2 in **Predictor (X)**
Highlight **Linear** for **Type of Regression**.
Click **OK**.

(c) Construct a 95% confidence interval estimate for Stock Y's β (β1). Interpret the meaning of this interval. [$0.4560 \le \beta 1 \le 1.2105$

(d) Is the rate of return on Stock Y positively related to the rate of return on the NYSE Index? Test at a level of significance of 0.05. [T=5.676, reject the null hypothesis]

(e) Construct an analysis of variance table for the above regression. In addition, perform the same test in (d) using a different test. (again, $\alpha = 0.05$). Are the results in (c) and (d) consistent?

(f) Find the standard error of the regression and interpret its significance. [Se^2= 10.563]

(g) Find the coefficient of determination and interpret its meaning. [$r^2 = 0.8656$]

(h) Find a 94% confidence interval estimate for the mean rate of return on Stock Y if the rate of return on the NYSE Index is 4.6%. [-0.8916, 5.063]

(i) Find a 99% confidence interval estimate for this year's rate of return on Stock Y if the New York Stock Exchange Index has a rate of return of 8.1%. (or a 99% prediction interval). Interpret this interval. Would you invest in this stock, based on your interval? [-9.1316, 19.1354]

(j) Find the coefficient of correlation between the rate of return on Stock Y and the rate of return on the NYSE [r = +0.930]

2. The Director of Management Information Systems at a conglomerate must prepare his long-range forecasts for the company's 3-year budget. In particular, he must develop staffing ratios to predict the

number of managers and project leaders based on the number of programmers. The results of a sample of the electronic data processing staffs of 10 companies within the industry are displayed below.

# of applications Programmers	15	7	20	12	16	20	10	9	18	15
# of Managers and Project leaders	6	2	10	4	7	8	4	6	7	9

(a) Find the regression coefficients. State the least squares linear regression equation. [$\hat{y} = -0.0885 + 0.45x$]
(b) Interpret the meaning of the slope and intercept.
(c) Compute Se^2 and interpret this value. [Se = 1.42]
(d) Compute the coefficient of determination and interpret its meaning in this problem. [$r^2 = 0.7018$]
(e) At the 0.05 level of significance, is there a linear relationship between the number of managers and the number of application programmers? Use T-test [T= 4.339, Rho]
(f) At the 0.05 level of significance, test for the appropriateness of the simple linear regression model. Use F-test [F=18.834, Rho]
(g) Set up a 95% confidence interval estimate of the true population slope. [$0.2109 \leq \beta 1 \leq 0.6891$]
(h) Set up a 95% confidence interval estimate of the true population intercept. [$-3.6374 \leq \beta o \leq 3.4604$]
(i) Set up a 95% confidence interval estimate of the average number of managers at companies where there are 10 programmers. [2.9711, 5.8559]
(j) Set up a 95% prediction interval estimate of the number of managers for a particular cmpany in which there are 10 programmers. [0.8354, 8.1716]
(k) Construct a residual plot of the above data. What can you conclude from this residual plot? Does the linear model seem appropriate? Explain.

3. High salaries for presidents and high executives of charitable organizations have been in the news from time to time. Consider the information in the table below for the United Way in 10 major cities in Canada.

City	Salary of President	Money Raised (per capita)
Ottawa	$161,396	$17.35
Montreal	$189,808	$15.81
Toronto	$201,490	$16.74
Winnipeg	$171,798	$31.49
Halifax	$108,364	$15.51
St.John's	$126,002	$23.87
Regina	$146,641	$15.89
Saskatoon	$155,192	$9.32
Edmonton	$169,999	$29.84
Vancouver	$143,025	$24.19

(a) Find the least-squares regression equation that expresses the presidents' annual salary as a linear function of the amount of money raised (per capita). Interpret the meaning of the slope term in the context of the question. [$\hat{y} = 152657 + 235.71x$]
(b) This past year, the City of Lethbridge (with a population of approximately 70,000) raised a total of 1.9 million dollars. Estimate the salary of the president of the United Way Lethbridge Chapter. [x=27.14, $\hat{y}$ = $159,024.95]
(c) Find the ANOVA table. What percentage of the variation in presidents' salary is explained by the fact that some raised more money per capita than others? [$r^2 = 0.0035$, very small]
(d) Is there a significant linear relationship between the president's salary and per capita money raised? Use the p-value approach and interpret the p-value in the context of the question. [t=0.1677, Aho]
(e) Does there appear to be a significant linear relationship between the amount of money raised per capita and the presidents' salary? Conduct this test using both the t-test and the F-test. What is your conclusion? [F=0.0281, Aho]

(f) This past year the United Way in Calgary raised $34.94 per capita. Construct a 99% confidence interval for the mean salary of the president of the United Way in Calgary. [$107,862.93, $213,922.71]
(g) Construct a 95% prediction interval for the salary of the president of the United Way in Calgary. [$74,227.12, $247,588.52]
(h) Estimate, with 90% level of reliability, the average (mean) salary of a United Way president who raised $29.00 per capita. [$130,189.84, $188,795.56]

4. The following is a MINITAB output for a random sample of 8 employees at Tackey Toy Manufacturing Company. The company wanted to see if there was a relationship between aptitude test results and output (dozens of units produced).

The regression equation is
Output= 1.03 +5.14 test results

Predictor	Coef	St.Dev	T	P
Constant		2.070		
Test result		0.2831		

S=1.695 R-sq= R-sq(adj)=97.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression					
Residual Error					
Total		968.00			

(a) Fill in the above tables and find R-sq.
(b) State the least squares regression equation. [$\hat{y} = 1.03 + 5.14x$]
(c) What percentage of the variation in output is explained by the fact that some had higher test results on the aptitude tests than others? [$r^2 = 0.982$, very large]
(d) Is there a significant linear relationship between output and aptitude test results? Use the p-value approach and interpret the p-value in the context of the question. [t=18.16, Rho, p-value~0]
(e) Does there appear to be a significant linear relationship between the aptitude test results and output? Conduct this test using both the t-test and the F-test. What is your conclusion? [t=18.16, F=329.61, Rho]