Place your name on ALL pages NOW. THANKS!

Name _____

IE 303 — Statistics I
Mark S. Daskin

Dept. of IE/MS

Winter, 2003
Northwestern University

FINAL EXAM -- SOLUTIONS

Problem 1: (20 percent; a = 5% b=5%; c=10%)

WCAS assigns students to freshman seminars based on their preferences (and based on a program that I wrote – really!) Suppose we have the following data for 67 seminars, each of which has exactly 16 students.

No Women	No Seminars
0	0
1	0
2	2
3	0
4	2
5	2
6	11
7	9
8	13
9	11
10	5
11	1
12	6
13	2
14	2
15	1
16	0

Total	67

In other words, there were 2 seminars that had only 2 women, 2 seminars that had 4 women, 2 seminars with 5 women, 11 seminars that had 6 women, and so on. There were a total of 556 women assigned in this particular quarter out of a total of 1,072 students.

a) Which of the following probability distributions would best reflect the number of women in a seminar assuming choices are random with respect to gender. (Check one only)

_____	Uniform	_____	Geometric
XXXX	Binomial	_____	Exponential
_____	Normal	_____	Chi-Squared

b) Suppose you believe that the number of women in each seminar should follow a Binomial distribution (and this **may** or **may not** be the correct answer to part (a)). In that case, find the parameters of the Binomial distribution assuming that men and women have the same probability of being assigned to any individual seminar.

Binomial with *n=67* and *p=556/1072 = 0.51865672*

c) To test the hypothesis that the underlying distribution is Binomial, you set up the following table.

No. Women	No. Seminars	Probability	Expected	Aggregate Groups -- No. Women	Expected	Observed	(Ei-Oi)2/Ei
0	0	0.000008	0.0006				
1	0	0.000143	0.0096				
2	2	0.001157	0.0775				
3	0	0.005818	0.3898				
4	2	0.020373	1.3650				
5	2	0.052685	3.5299	≤ 5	5.3723	6	0.073338
6	11	0.104077	6.9731	6	6.9731	11	2.3254451
7	9	0.160207	10.7338	7	10.7338	9	0.2800676
8	13			8		13	
9	11	0.186008	12.4625	9	12.4625	11	0.1716282
10	5	0.140299	9.4000	10	9.4000	5	2.0595834
11	1	0.082459	5.5247	11	5.5247	1	3.7057469
12	6	0.037021	2.4804	≥ 12	3.5218	11	
13	2	0.012274	0.8224				
14	2	0.002834	0.1899				
15	1	0.000407	0.0273				
16	0	0.000027	0.0018				
Total	**67**	1.0000000	67			67	

Find the values of the six shaded (highlighted) cells and test the null hypothesis that the number of women in a seminar follows a Binomial distribution. Use α=0.05. Use the next page for your work, but fill in the answers in the table above.

No Women	No Seminars	Prob (Binomial)	Expected	No Women	Expected	Observed	(Ei-Oi)^2 / Ei
0	0	0.000008	0.0006				
1	0	0.000143	0.0096				
2	2	0.001157	0.0775				
3	0	0.005818	0.3898				
4	2	0.020373	1.3650				
5	2	0.052685	3.5299	≤ 5	5.3723	6	0.073338
6	11	0.104077	6.9731	6	6.9731	11	2.325445
7	9	0.160207	10.7338	7	10.7338	9	0.280068
8	13	**0.194204**	**13.0117**	8	**13.0117**	13	**0.000010**
9	11	0.186008	12.4625	9	12.4625	11	0.171628
10	5	0.140299	9.4000	10	9.4000	5	2.059583
11	1	0.082459	5.5247	11	5.5247	1	3.705747
12	6	0.037021	2.4804	≥ 12	3.5218	11	**15.879272**
13	2	0.012274	0.8224				
14	2	0.002834	0.1899				
15	1	0.000407	0.0273				
16	0	0.000027	0.0018				

Total	**67**	1	67		**67**	67	**24.49509**

We compare the computed test statistic (24.495) to the critical value of the chi-squared distribution with 6 degrees of freedom (12.592) and conclude that we should reject the null hypothesis.

Problem 2: (20 percent; a=10%; b=10%)

The Center for Disease Control (CDC) reports on a question that asked "Do you have any kind of health care coverage?" Results of the question by education level are shown below for Illinois.

Education	Yes	No	Total
Less than H.S.	295	64	359
H.S. or G.E.D.	1053	124	1177
Some post-H.S.	1084	93	1177
College Graduate	1233	50	1283

Education	Yes	No
Less than H.S.	82.2%	17.8%
H.S. or G.E.D.	89.5%	10.5%
Some post-H.S.	92.1%	7.9%
College Graduate	96.1%	3.9%

(Source = http://apps.nccd.cdc.gov/brfss/education.asp?cat=HC&yr=2001&qkey=868&state=IL)

a. For the data shown above, test the following hypothesis with α=0.05.

$$H_0: \quad p_{\text{HS or GED}} = p_{\text{Some post HS}}$$
$$H_1: \quad p_{\text{HS or GED}} \neq p_{\text{Some post HS}}$$

where $p_{\text{HS or GED}}$ is the fraction of the population who have a High School or GED degree who answered yes to this question and other probabilities have the obvious meaning. Clearly show your work.

$$p = \frac{1053 + 1084}{1177 + 1177} = 0.9078$$

$$\text{Test statistic} = \frac{1084/1177 - 1053/1177}{\sqrt{0.9078 \bullet 0.0922 * \left(\frac{1}{1177} + \frac{1}{1177}\right)}} = 2.20868$$

Since this is greater than 1.96, we reject the null hypothesis.

b. For the data shown above, test the following hypothesis with α=0.05.

$$H_0: \quad p_{\text{Less than HS}} = p_{\text{College Grad}}$$
$$H_1: \quad p_{\text{Less than HS}} \neq p_{\text{College Grad}}$$

Clearly show your work.

$$p = \frac{1233 + 295}{1283 + 359} = 0.9306$$

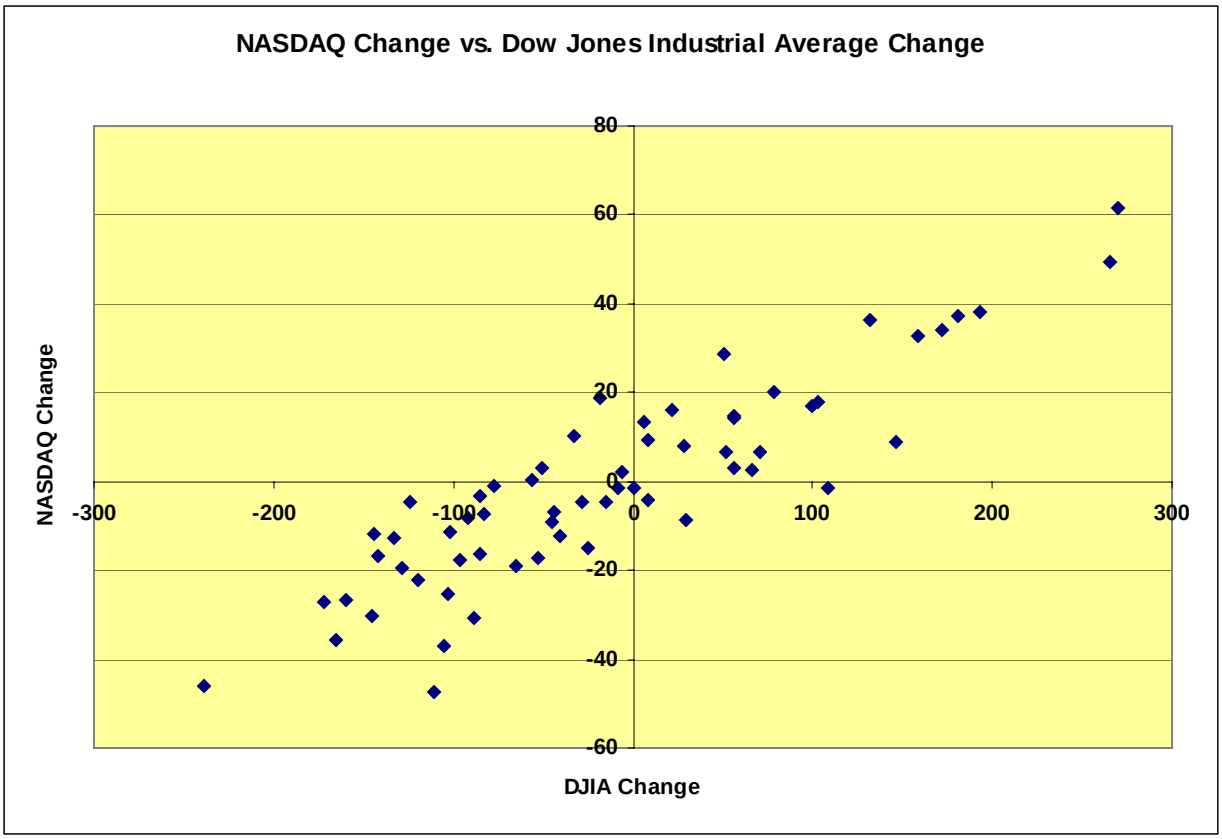
$$\text{Test statistic} = \frac{1233/1283 - 295/359}{\sqrt{0.9306 \bullet 0.0694 * \left(\frac{1}{1283} + \frac{1}{359}\right)}} = 9.179$$

Since this is greater than 1.96, we reject the null hypothesis.

Problem 3: (40%; a=12%; b=12%; c=8%; d=8%)

The closing averages for the NASDAQ (a stock exchange with a large number of high tech stocks) and the Dow Jones Industrial Average (a weighted average of 30 very large companies) were recorded from December 11, 2002 through March 13, 2003. There were a total of 63 business days during this interval during which the markets were open. The change in the two indices were computed from these 63 closing values, resulting in 62 **changes**. These changes are plotted below

Name _____



(source = http://table.finance.yahoo.com/k?s=^dji&g=d and http://table.finance.yahoo.com/k?s=^ixic&g=d)

A summary of the relevant data is as follows, where

Y = NASDAQ Change and

X = DOW JONES Change.

$$\sum_{i=1}^{62} Y_i = -55.82 \qquad \sum_{i=1}^{62} Y_i^2 = 30{,}752.79$$

$$\sum_{i=1}^{62} X_i = -767.39 \qquad \sum_{i=1}^{62} X_i^2 = 755{,}864.41$$

$$\sum_{i=1}^{62} X_i Y_i = 135{,}987.53$$

a) The regression output (from EXCEL) for these data is shown below. Fill in the 12 boxes that are shown with heavy lines. If a P-value is very small, just indicate that it is smaller than a number whose value you can readily estimate.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8938
R Square	0.7988
Adjusted R Squar	0.7955
Standard Error	10.1463
Observations	62

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	24525.73	24525.73	238.24	0.00
Residual	60	6176.80	102.95		
Total	61	30702.53			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.3433	1.2967	1.0359	0.3044	-1.2505	3.9372
DJIA Change	0.1813	0.0117	15.4349	0.0000	0.1578	0.2048

b) In listening to the reports on the closing averages, it seemed to me that you could reasonably well predict the change in the NASDAQ by dividing the change in the Dow Jones Average by 4. Test whether or not this is true using α=0.05. You will need to perform **two** different hypothesis tests to see if this is true. Clearly state the null and alternative hypotheses and show your work.

Do you believe this is a good model? If not, propose an alternative model that you think would work well and that would be easy to use (since using this regression is not easy if I just want to guess the change in the NASDAQ after I hear the change in the Dow Jones. I would only have about 10 seconds to do the guestimation in my head.)

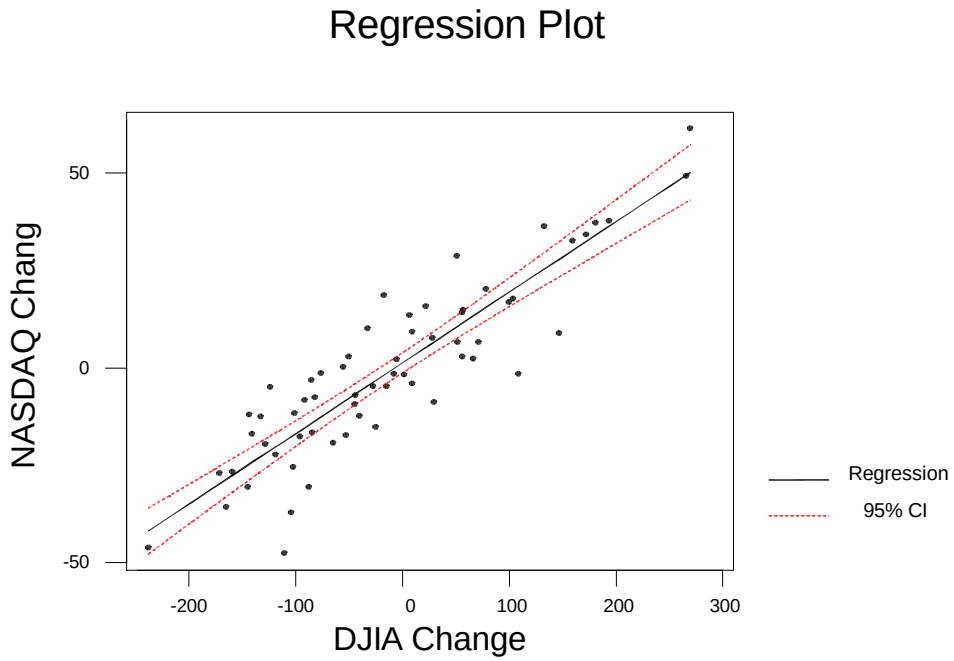
First test is to determine whether or not the intercept is significantly different from 0, which it cannot be if we are to simply divide by 4. The t-stat listed above is smaller than any standard critical value, so we cannot reject the null hypothesis that the intercept is equal to 0.

The second test is $\begin{matrix} H_0: & \beta_1 = 0.25 \\ H_1: & \beta_1 \neq 0.25 \end{matrix}$. The test statistic for this is $\frac{0.1813 - 0.25}{0.0117} = -5.8718$.

Since the absolute value of this exceeds the critical value for the t-test (or the z-test) at α=0.05, we reject the null hypothesis and conclude that this is NOT a good model.

A better model would be to divide by 5. Here the test statistic is $\frac{0.1813-0.20}{0.0117}=-1.5982$ which is smaller than the critical value for α=0.05, so we can conclude that this is a reasonable model.

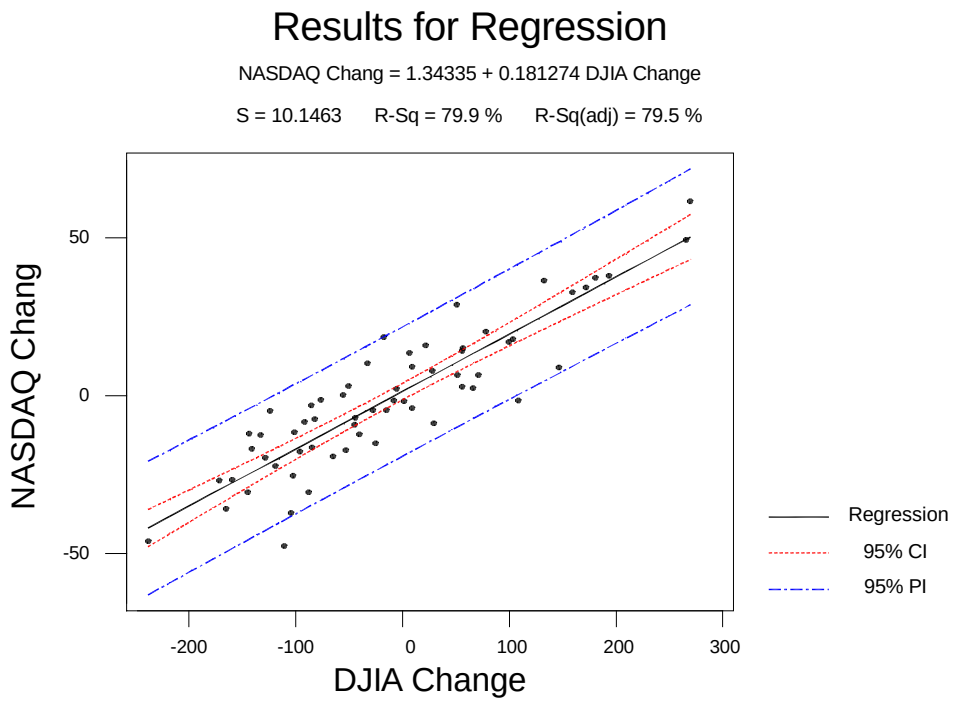
c) When I plot the data and the regression line and the 95% confidence intervals for the line, I found that many of the observations fell outside of the 95% confidence interval. Specifically, you get the following (from MiniTab).



Is this a surprising result? Briefly explain your answer.

No, this is not a surprising result since this is a **confidence** interval and not a **prediction** interval.

The prediction interval and confidence interval are shown below. Note that most of the observations are within the 95% prediction intervals



d) Find a 95% prediction interval for the change in the NASDAQ on a day when the Dow Jones Industrial average lost 100 points.

$$\hat{Y}^* \pm t_{60,0.025} s\sqrt{1+\frac{1}{62}+\frac{\left(-100-\left(-12.3773\right)\right)^2}{755864.41}} = 1.3433 - 0.1813(100) \pm 2.000 \bullet 10.1463 \bullet 1.013058$$
$$= -16.784 \pm 20.5575$$

Problem 4: (20 percent: a=10%; b=10%)

Consider the following data on the amount of time students spend studying for finals. A random sample of 50 students was taken and the following data were obtained.

$$\sum_{i=1}^{50} h_i = 1{,}503 \qquad \sum_{i=1}^{50} h_i^2 = 45{,}572$$

where h_i represents the number of hours that the i[th] student spent studying.

a) Find the sample average, sample variance and sample standard deviation of the number of hours a student spends studying for finals.

$$\bar{X} = \frac{\sum_{i=1}^{50} h_i}{50} = \frac{1503}{50} = 30.06$$

$$s^2 = \frac{\sum_{i=1}^{50} h_i^2 - \left(\sum_{i=1}^{50} h_i\right)^2 \Big/ 50}{49} = \frac{45572 - 1503^2/50}{49} = 7.996$$

$$s = \sqrt{7.996} = 2.828$$

b) Find a 95% confidence interval for the number of hours a student spends studying

We want a 95% confidence interval for the true **mean** number of hours a student spends studying. This is $30.06 \pm t_{49,0.025}\frac{s}{\sqrt{50}} = 30.06 \pm 2.01\frac{2.828}{\sqrt{50}} = 30.06 \pm 0.8036$

This question was ambiguous and many people (possibly correctly) interpreted this as asking for a confidence interval for a **single** observation in which case you would get $30.06 \pm t_{49,0.025}\, s = 30.06 \pm 2.01 \bullet 2.828 = 30.06 \pm 5.684$