Place your name on ALL pages NOW. THANKS!

Name _____

IE 303 — Statistics I
Mark S. Daskin

Dept. of IE/MS

Spring, 2002
Northwestern University

FINAL EXAM

Problem 1:

Computer chips are produced in batches of 144 chips at a time. Each chip is tested and the number of defective chips per batch is recorded. After the production of 200 batches of chips (28,800 chips in total), the following information is obtained for the total number of defective chips.

$$\sum_{i=1}^{200} d_i = 2{,}933 \qquad \sum_{i=1}^{200} d_i^2 = 46{,}059$$

a. Find the sample average and sample variance of the number of defective chips per batch.

$$\overline{D} = 14.665$$
$$s_D^2 = 15.3093$$

b. The quality engineer in charge of the process believes that the number of defects per batch follows a Poisson distribution. He has summarized the data in the table on the next page. Complete the 6 missing cells in the table. Clearly show your work.

No Defects	No. Groups (Observed value)	Expected	$(E_i-O_i)^2/E_i$
7 or fewer	5	4.36	0.094
8	6	4.54	0.471
9	7	7.39	0.021
10	9	10.84	0.313
11	13	14.46	0.147
12	12	17.67	1.817
13	28	19.929	3.269
14	25	20.88	0.815
15	16	20.410	0.953
16	22	18.71	0.580
17	13	16.14	0.610
18	11	13.15	0.351
19	11	10.15	0.072
20	6	7.44	0.279
21	7	5.20	0.626
22 or more	9	8.75	0.007
TOTAL	**200**	**200**	**10.425**

c. Test the hypothesis that the number of defects per batch comes from a Poisson distribution against the alternate hypothesis that the number of defects does not follow a Poisson distribution with α=0.1. Clearly indicate the test statistic value you compute from the data, the distribution of the test statistic under the null hypothesis, the number of degrees of freedom used in the test and whether or not you reject the null hypothesis.

The test statistic has a value of 10.425 This has a Chi-squared distribution with 14 degrees of freedom. The critical value for α=0.1 is 21.064, so we do not reject the null hypothesis.

Problem 2:

Data were collected in a study in Bollate (a suburb of Milan, Italy) on the incidence of eye disease. A portion of the data is summarized below, giving the number of people and the percentage of the population suffering from senile macular degeneration.

	Males			Females			Total		
Age	Sample Size	Number Affected	Percent Affected	Sample Size	Number Affected	Percent Affected	Sample Size	Number Affected	Percent Affected
60-69	240	33	13.8%	241	34	14.1%	481	67	13.9%
70-79	82	15	18.3%	46	11	23.9%	128	26	20.3%
Total	322	48	14.9%	287	45	15.7%	609	93	15.3%

(Source = http://www.itba.mi.cnr.it/epidemiology/eye_disease.html)

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

$$H_0: \quad p_{male} = p_{female}$$
$$H_1: \quad p_{male} \neq p_{female}$$

where p_{male} is the fraction of the male population with macular degeneration and p_{female} is the fraction of the female population with the disease. Clearly show your work.

$$p = \frac{93}{609} = 0.153$$

$$\text{test statistic} = \frac{0.157 - 0.149}{\sqrt{0.153 \cdot 0.847\left(\frac{1}{322} + \frac{1}{287}\right)}} = 0.271$$

which is not statistically significant at $\alpha = 0.05$

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

$$H_0: \quad p_{60-69} = p_{70-79}$$
$$H_1: \quad p_{60-69} \neq p_{70-79}$$

where p_{60-69} is the fraction of the 60-69 year old individuals with macular degeneration and p_{70-79} is the fraction of the 70-79 year old population with the disease. Clearly show your work.

$$p = \frac{93}{609} = 0.153$$

$$\text{test statistic} = \frac{0.203 - 0.139}{\sqrt{0.153 \cdot 0.847\left(\frac{1}{481} + \frac{1}{128}\right)}} = 1.783$$

which is also not statistically significant at $\alpha = 0.05$ since we are doing a two sided test and the critical value is 1.96

Problem 3:

You are interested in estimating the level of E. coli bacteria in Lake Michigan water. You have available to you three different testing methods as described below:

Method	Biased/unbiased	Variance $(\text{cells}/100\text{ ml})^2$	Cost per test (dollars)	Time per test (minutes)
A	Unbiased	100	$200	30
B	Unbiased	225	$75	15
C	Unbiased	675	$25	10

Because of budget considerations, you elect to test four samples using method B every day and eight samples using method C. Because it is such an expensive test to use, method A is not used for any samples despite the fact that it has the smallest variance.

Let X_j^B be the value of the j[th] sample tested using method B and X_k^C be the value of the k[th] sample tested using method C. Let w_B be the weight that is placed on each sample tested using method B and w_C be the weight placed on each sample tested using method C. In other words, the estimator that is used to find the estimated value of the number of E. coli cells per 100 ml of water is the following:

$$\tilde{X} = w_B\left(X_1^B + X_2^B + X_3^B + X_4^B\right) + w_C\left(X_1^C + X_2^C + X_3^C + X_4^C + X_5^C + X_6^C + X_7^C + X_8^C\right)$$

a. In addition to the obvious conditions that $w_B \geq 0$ and $w_C \geq 0$, write down one more conditions that is (are) needed to ensure that $\tilde{X}$ is an unbiased estimator of the true level of E. coli in the lake.

$$4w_B + 8w_C = 1$$

b. Find the values of w_B and w_C that minimize the variance of the estimator $\tilde{X}$. Recall that the variance of $\tilde{X}$ is given by:

$$Var(\tilde{X}) = 4w_B^2 Var(\text{method B}) + 8w_C^2 Var(\text{method C})$$

$$\begin{aligned} \text{Min} \quad & Var(\tilde{X}) = 4w_B^2 Var(\text{method B}) + 8w_C^2 Var(\text{method C}) \\ & = 4w_B^2 225 + 8w_C^2 675 \\ & = 900w_B^2 + 5400w_C^2 \\ \text{Subject to} \quad & 4w_B + 8w_C = 1 \end{aligned}$$

$$w_C = \frac{1-4w_B}{8}$$

$$\text{Min} \quad Var(\tilde{X}) = 900w_B^2 + 5400\left(\frac{1-4w_B}{8}\right)^2$$

$$\begin{aligned} \frac{dVar(\tilde{X})}{dw_B} &= 1800w_B + 10{,}800\left(\frac{1-4w_B}{8}\right)\left(-\frac{1}{2}\right) = 0 \\ &= 1800w_B - \frac{10800}{16} + 2700w_B = 0 \\ &= 4500w_B - \frac{10800}{16} = 0 \\ w_B &= 0.15 \\ w_C &= 0.05 \end{aligned}$$

c. Using the values of w_B and w_C that you obtained in part (b), find the variance of $\tilde{X}$. If you could not solve for w_B and w_C, use $w_B = 0.1$ and $w_C = 0.075$.

*Note that these are **not** the correct values for part (b).*

$$Var(\tilde{X}) = 4(0.15^2)225 + 8(0.05^2)675 = 33.75 \text{ using correct weights}$$
$$Var(\tilde{X}) = 4(0.1^2)225 + 8(0.075^2)675 = 39.375 \text{ using incorrect weights}$$

d. Suppose now that the observations were weighted equally (i.e., $w_B = w_C = 1/12$). Calculate the percentage increase in the variance of $\tilde{X}$ that would result from this weighting of the observations compared to that of the (optimal) weighting found in part (c).

Note that if you solved part (b) correctly, the variance in (c) will be the minimum variance. If you did not, the answer to part (c) should still be a lower variance for $\tilde{X}$ than you would get from using equal weights on the observations.

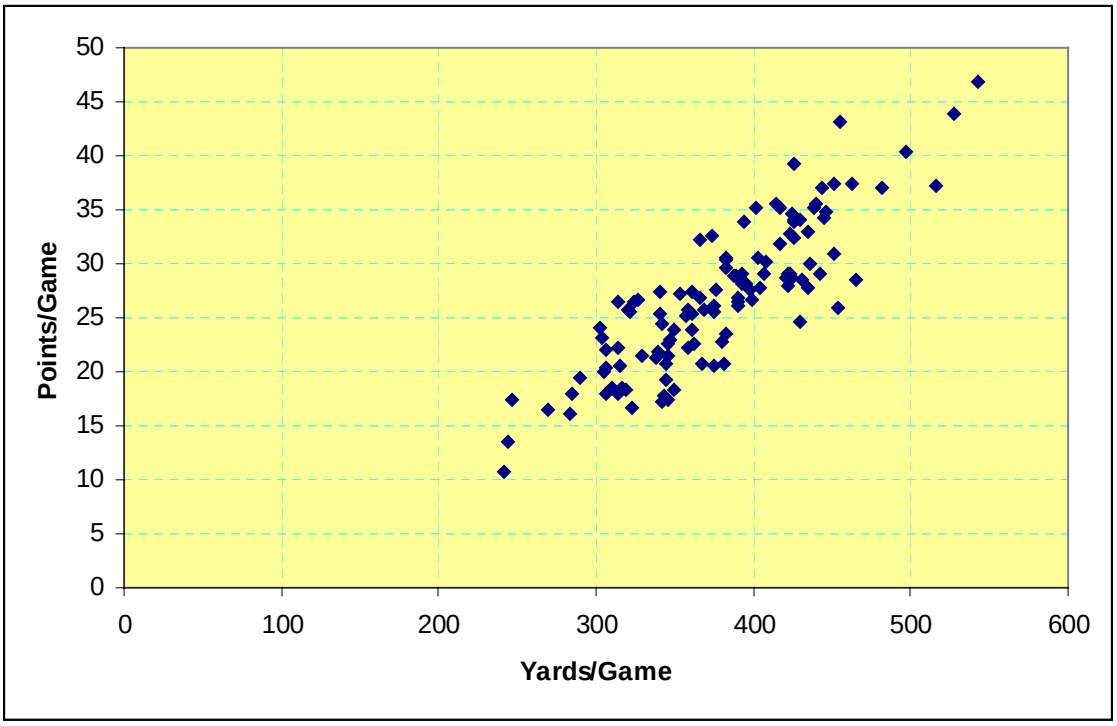
$$Var(\bar{X}) = 4\left(\left(\frac{1}{12}\right)^2\right)225 + 8\left(\left(\frac{1}{12}\right)^2\right)675 = 43.75$$

$$\text{Percentage increase using correct weights} = \frac{43.75 - 33.75}{33.75} \bullet 100 = 29.63\%$$

$$\text{Percentage increase using incorrect weights} = \frac{43.75 - 39.375}{39.375} \bullet 100 = 11.01\%$$

Problem 4:

Data are collected on 117 Division 1-A football teams. We are interested in the relationship between the number of points per game that a team scores and the number of yards per game that they gain. These data are shown below.



(source = http://www.cbs.sportsline.com/u/football/college/stats/tptsNCAAF.htm)

A summary of the relevant data is as follows, where Y = Yards/game and P = Points/game.

$$\sum_{i=1}^{117} Y_i = 44{,}313.2 \qquad \sum_{i=1}^{117} Y_i^2 = 17{,}180{,}049$$

$$\sum_{i=1}^{117} P_i = 3{,}150.6 \qquad \sum_{i=1}^{117} P_i^2 = 90{,}093.8$$

$$\sum_{i=1}^{117} Y_i P_i = 1{,}232{,}567.56$$

Just for your interest, Northwestern ranked 40th from the top of this list in terms of total points with 320 points, 29.1 points/game and 442.9 yards/game.

a. The regression output (from EXCEL) for these data is shown below. Fill in the 11 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	0.8608
R Square	$r^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} = 0.7409$
Adjusted R Square	**Ignore this cell**
Standard Error	$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{11.837} = 3.4405$
Observations	117

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3892.54	3,892.54	328.84	Smaller than 10^{-6}
Residual	115	1,361.26	11.837		
Total	116	5,253.80			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-10.5924	2.0934	-5.0599	1.6054E-06
Yards/Game	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{39292.62}{396632.81} = 0.0991$	0.0055	18.012	Smaller than 10^{-6}

$$S_{xx} = 17180049 - 44313.2^2/117 = 396632.81$$
$$S_{yy} = 90093.8 - 3150.6^2/117 = 5253.797$$
$$S_{xy} = 1232567.56 - 44313.2 \bullet 3150.6/117 = 39292.62$$

b. Consider the following hypothesis related to β_1, the slope term:

$$H_0: \quad \beta_1 = 0.1$$
$$H_1: \quad \beta_1 \neq 0.1$$
$$\alpha = 0.05$$

Can you reject the null hypothesis at the indicated level of significance?

Test statistic = $\dfrac{0.0991 - 0.1}{0.0055} = -0.164$ so you cannot reject the null hypothesis.

c. Find a 95% **prediction** interval the number of points per game scored by a team that gains 440 yards per game (e.g., like Northwestern University).

$$-10.5924 + 0.0991(440) \pm 1.98(3.4405)\sqrt{1 + \frac{1}{117} + \frac{(440 - 378.745)^2}{396632.81}} = 33.0116 \pm 6.8732$$

26.1384 *to* 39.8848

Problem 5:

Group Assessment Form

Your name _____ **Group name** _____

Please evaluate both yourself and your other group members for each of the three activities the group was involved in as well as an overall assessment. Give yourself and each group member a grade (A through F with the obvious meaning). Please feel free to include comments discussing the roles of the different team members in each assignment.

Team Member	Lab 1	Lab 2	Project	Overall Assessment
Yourself				