Practice for Final Exam
15.063 Communicating with Data
Summer 2003

Problem 1 (20 points. a, b: 7 points each. c: 6 points)

Gill Bates has an opportunity to purchase a lot of electronic components for resell from a local manufacturer. He can purchase the entire lot and resale it at a profit of $10,000.00, but only if the components in the lot are of good quality. If the components in the lot are defective, Robert will lose $3,000.00 in the transaction.
In the past, four out of five lots from this manufacturer have been of good quality (one out of five has been defective).
Robert has an opportunity to hedge his bet by having the lot inspected before making the purchase decision: For $200, he can have the lot inspected. He believes that if the lot is good, the inspector will, with a probability of 0.15, declare the lot defective. Further, if the lot is defective, the inspector will, with a probability of 0.1 declare the lot good.

Let G represent the event that the lot is good;
Let D represent the event that the lot is defective;
Let Q represent the event that the inspector declares the lot to be of good quality;
Let N represent the event that the inspector declares the lot to be of poor quality.

(a) Fill out the Conditional Probability Table.

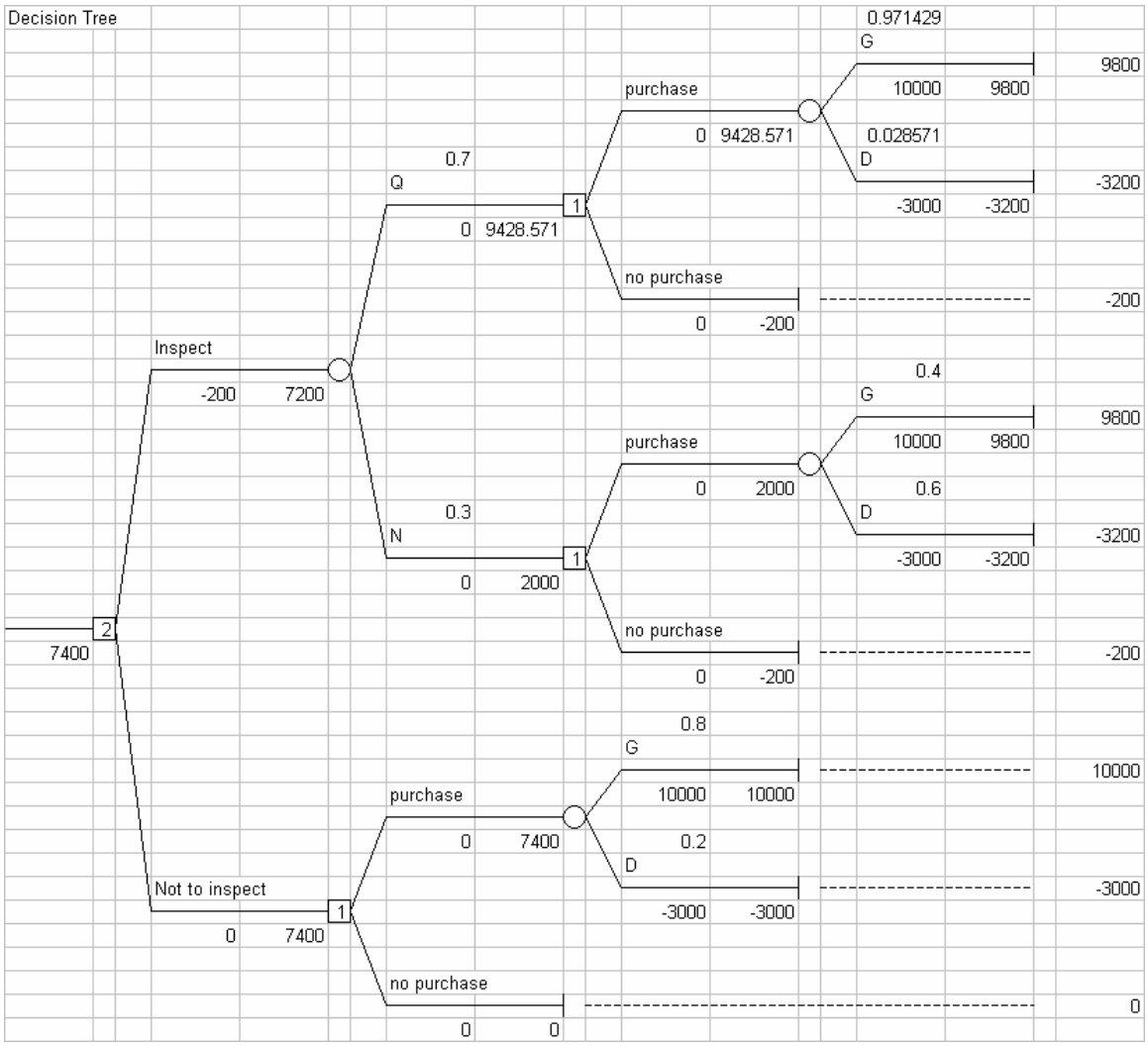
	G	D	Total
Q	0.68	0.02	0.7
N	0.12	0.18	0.3
Total	0.8	0.2	1

(b) Find the following probabilities: P(Q), P(N), P(D|N), P(G|N), P(D|Q), P(G|Q).

From the above table, we see that

P(Q)=0.7, P(N)=0.3, P(D|N)=0.6, P(G|N)=0.4, P(D|Q)=1/35, P(G|Q)=34/35

(c) Build (but do not solve) the decision tree. Be sure to clearly mark each node (whether it is a decision node or event node), the decision or event at each branch, the probability for each event branch, and the payout associated with each leaf (i.e., the outcome at the end of each path in the tree)



Problem 2 (20 points. a, b: 7 points each. c: 6 points)

The Human Resources department at *MOIRTC Technologies* is planning to interview 7 software engineers for the 3 job openings that emerged last week. According to job market survey results, the salary asked by software engineers with 3 years' working experiences is uniformly distributed between $65,000 and $85,000. The 7 interviewees all have 3 years' working experience and the salaries they will ask are independent and in line with the probability distribution indicated by the survey. Suppose *MOIRTC* decided to set the starting salary for the 3 openings at $70,000 each.

(a) What is the probability that all 7 of the interviewees ask for a salary that is less than or equal to $70,000?

Let X be the random variable that represents how much a software engineer will ask. As it is uniform:

$P(X<70000)=(70000-65000)/(85000-65000)=0.25$

Let Y be the random variable that represents the number of interviewees among the 7 that will ask below 70,000. It is clear that Y satisfies all the assumptions of a binomial random variable. Then,

$P(Y=7)=(0.25)^7$

(b) Compute the probability that there are at least 3 among the 7 interviewees (i.e., 3 or more interviewees) whose salary expectation is less than or equal to $70,000.

First, we reverse the inequality because the other is easier to compute.

$P(Y>=3)= 1 - P(Y<=2)$.

To compute $P(Y<=2)$, we sum $P(y=0) + P(y=1) + P(y=2)$, computed using the binomial formula.

$P(Y=0) = (0.25)^0(0.75)^7 = 0.133$
$P(Y=1) = 7 \times (0.25)^1(0.75)^6 = 0.311$
$P(Y=2) = 21 \times (0.25)^2(0.75)^5 = 0.311$

Totaling 0.756, from where the answer is $P(Y>=3)= 0.244$.

(c) There are 300 software engineers with 3 years' working experiences on the job market right now. What is the probability that at least 30 of them (i.e., 30 or more) ask for a salary that is less than or equal to $70,000? (Hint: use an appropriate approximation, and explain why you may do so.)

Let W be the random variable that represents the number of software engineers, among the 300 whose salary expectation is below 70,000. Note that W can be approximated by a normal random variable W' with mean 75 (=np) and standard deviation 7.5 ($=[np(1-p)]^{1/2}$). Thus,

$P(W>=30) \approx P(W'>=30) = P((W'-75)/7.5 >= (30-75)/7.5) = P(Z>=-6) = 1.$

Problem 3 **(20 points. a, b: 7 points each. c: 6 points)**

The following data represent the number of years of education for each employee in a random sample of 15 employees in a large corporation

Employee ID Number	*Years of Education*
01	6
02	7
03	7
04	8
05	12
06	12
07	12
08	13
09	13
10	13
11	14
12	15
13	16
14	18
15	21

(a) What is your estimate of the mean ($\bar{x}$) and standard deviation (s) of the number of years of education for the employees in this corporation?

$$\bar{x} = \frac{(6+7+7+8+12+12+12+13+13+13+14+15+16+18+21)}{15} = 12.46667$$

$$s = \sqrt{\frac{(6-12.47)^2 + ... + (21-12.47)^2}{14}} = 4.2065$$

(b) Construct a 95 % confidence interval for the true mean.

$$\left[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right] = [10.34, 14.60]$$

Actually, a more accurate answer can be given using *t*-distributions. Looking for the value of a 95% confidence interval with 14 degrees of freedom (*n*-1) in the table, gives that *c*=2.145. Then, a better approximation to the confidence interval is [10.14,14.80].

(c) What sample size would be needed to obtain a 95% confidence interval whose tolerance level is plus or minus half a year?

$$1.96\frac{s}{\sqrt{n}} = 0.5 \Rightarrow n = 272$$

Or, again, using a *c* value of 2.145, *n* = 325.

Problem 4 (20 points. a, b, c: 4 points each d: 8 points)

CWD Engineering employs a large number of engineers. Mary wants to analyze the extent to which the salaries of these engineers are related to their education and experience. She has obtained data for a sample of 20 engineers: (1) their yearly salary; (2) their undergraduate GPA (grade point average); and (3) their number of years of experience.

Engineer Number	Yearly Salary ($)	Undergrad. GPA	Years of Experience	Engineer Number	Yearly Salary ($)	Undergrad. GPA	Years of Experience
1	118,683	4.00	14	11	50,960	2.47	6
2	116,420	3.80	17	12	106,694	3.79	11
3	56,021	2.82	4	13	95,850	2.81	8
4	71,274	2.85	6	14	80,346	2.67	10
5	99,204	3.68	13	15	67,010	2.17	7
6	91,695	3.26	13	16	47,545	2.12	4
7	79,564	2.48	14	17	108,438	3.48	11
8	100,771	4.00	16	18	85,605	3.82	5
9	62,091	2.93	2	19	71,877	2.68	8
10	96,634	3.42	12	20	88,541	2.60	12

Mary used a multiple regression model to predict yearly salary on the basis of these data. The resulting computer output is shown below.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.91534
R Square	0.83784
Adjusted R Square	0.81876
Standard Error	9,038.06
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	2	7,174,865,080	3,587,432,540	43.917032
Residual	17	1,388,672,010	81,686,589	
Total	19	8,563,537,091		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4,012.80	10,691.55	0.37532	0.71206
Undergraduate GPA	18,188.59	4,016.87	4.52804	0.00030
Years of Experience	2,538.87	573.40	4.42773	0.00037

(a) Write a complete equation for the simple linear regression model that incorporates the estimated coefficients provided by this computer output. Make sure to define *in words* all the variables used in this equation, and the *units* in which each is expressed.

Let Y be the dependent variable that expresses salary (in dollars). The independent variables are: X1 which is undergraduate GPA (in 0-5 points) and X2 which is the experience (in years).

The regression model tells us Y = 4,012.80 + 18,188.59 x X1 + 2,538.87 x X2

(b) Use the regression model in (a) to predict the annual yearly salary of engineer Joe Smith, who has a 3.50 undergraduate GPA and 10 years of experience.

y = 18,188.59 x 3.5+2,538.87 x 10 + 4,012.8 = 93,061.57

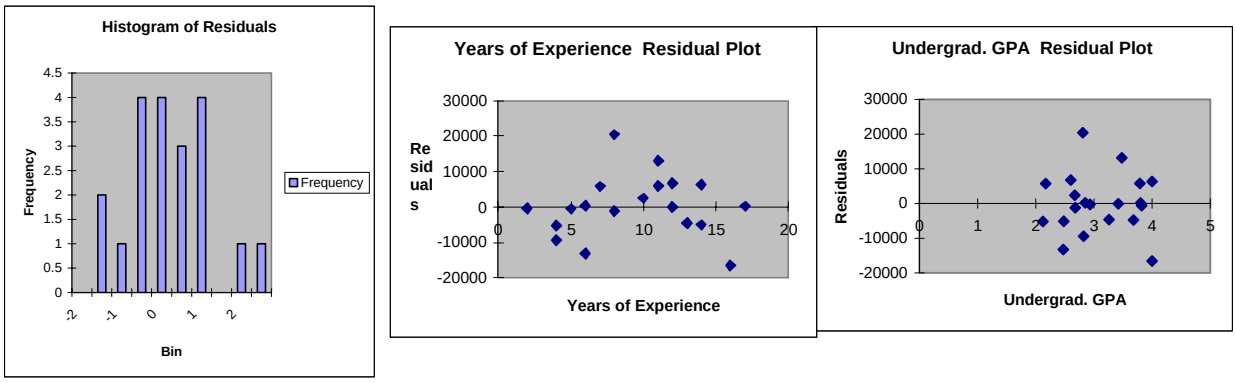
(c) According to the regression model in (a) above, the *actual* yearly salary of engineer Joe Smith is *most* likely to differ from the predicted value found in question (b) above by: (choose the BEST answer. Hint: consider the SD of the error term in your regression model) cE

(cA) about ±$1.00 (cB) about ±$10.00 (cC) about ±$100.00
(cD) about ±$1,000.00 (cE) about ±$10,000.00 (cF) about ±$100,000.00

The correlation matrix of the data in this problem is:

	Yearly Salary ($)	*Undergrad. GPA*	*Years of Experience*
Yearly Salary ($)	1		
Undergrad. GPA	0.806741	1	
Years of Experience	0.801412	0.543407	1

Some more output of the regression is:



(d) In general, does the model satisfy the assumptions that a linear regression makes? Are both variables in the model statistically significant predictors? Do you feel comfortable using this regression model?

R Square is large enough; Residuals look normally distributed; "Years of Experience" residual plot seems to indicate some heterocedasticity; GPA plot seems fine; the correlation matrix looks fine. Variables seem significant (i.e., zero is not in the confidence intervals of the coefficients of the independent variables because t-statistics are large enough).

There is no graph that can help us determine if autocorrelation is present or not.