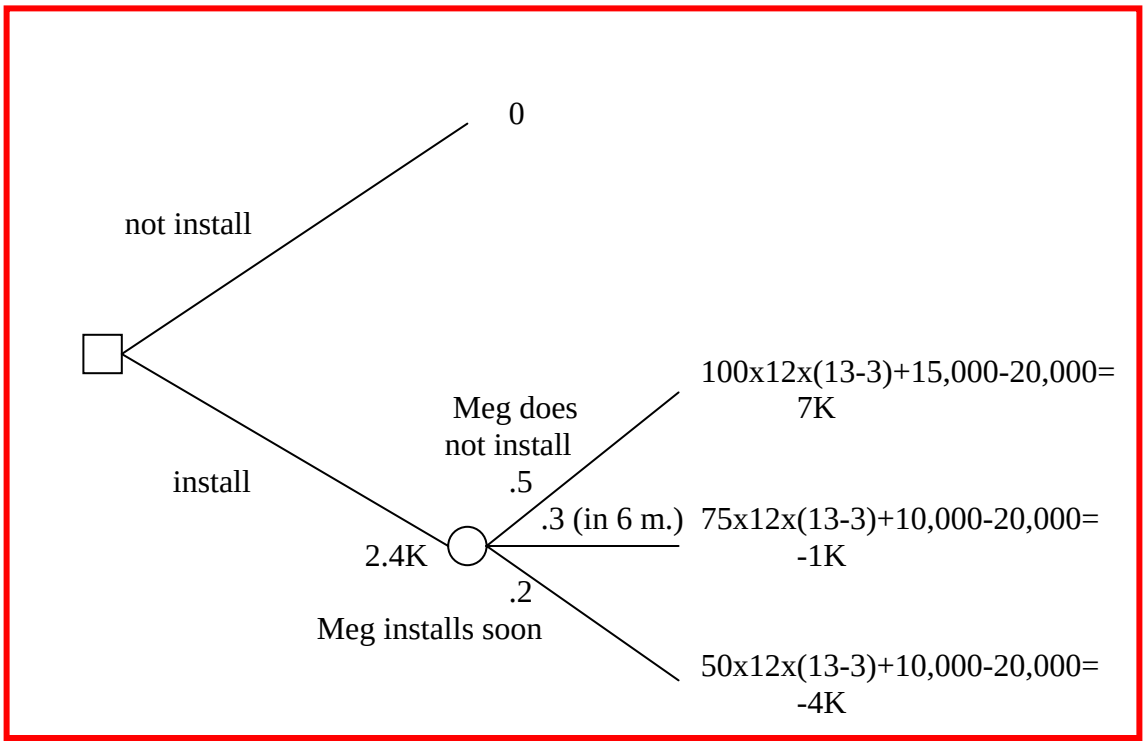
Solution to Final Examination
15.063 Communicating with Data
Summer 2003

Problem 1 (30 points)

Joe Bloke operates a garage in Rockwell, MA. He has just been offered the opportunity to install a car wash at the highly discounted cost of $20,000. This exceptional offer, however, is only available if he accepts it within one week. If he does, he will charge $13.00 per car washed. Labor and materials will cost $3.00 per car washed. Joe expects that monthly demand for car wash will be about 100 cars. A major uncertainty is whether or not his competitor, Meg's Garage and Auto Repair, will also install her own car wash. If Meg does, then Joe expects the demand for his car wash will be halved. Joe figures that there is a 20% chance that Meg will install her car wash very soon, in which case Joe and Meg would split demand for the next 12 months; a 30% chance that Meg will install her car wash in about 6 months, in which case Joe would have the full demand for 6 months and split it with her afterwards; and a 50% chance that Meg will not install a car wash at all**.** Furthermore, Joe's accountant estimates that the value of the car wash after the first 12 months is $15,000 in the absence of competition, and $10,000 if Meg has installed her own car wash.

(a) *(12 points)* Construct a decision tree to represent Joe's problem. Be sure to properly label all branches, and to include all relevant probabilities and payoffs.



(b) *(5 points)* Find the best strategy for Joe, assuming the Expected Monetary Value (EMV) criterion. What is the EMV of this strategy?

As shown in the tree the EMV of the event corresponding to Joe installing the car wash is 2.4K (= 7K x .5 – 1K x .3 – 4K x .2) which should be chosen over 0K corresponding of not installing.

(c) *(5 points)* Is EMV an appropriate decision criterion for Joe in such a situation? Why, or why not? (Please limit your answer to no more than 4 lines of text.)

Although there is a 50% of loosing money, the expected value is positive and there is a probability of 20% of loosing 4K, which is not too risky. (if you answer that it was not appropriate and justified it, no problem).

(d) *(8 points)* If the monthly demand was smaller or bigger than 100, would your recommendation change? For which values of monthly demand should Joe choose to have the car wash installed? Assume that he uses the EMV criterion and that all other data are as originally stated.

The bigger the demand, the more profit Joe makes and vice-versa. To determine the value in which the decision to open or not changes, we let Q be the demand and write the equation that makes the two EMV equal. Namely,

(120 Q – 500) x .5 + (90 Q – 10,000) x .3 + (60 Q – 10,000) x .2 = 0,

or after simplifying, 99 Q = 7500. Then, for Q > 76 we should install or for Q < 75 we should not.

Problem 2 (10 points)

Suppose that, at last, Joe opened the car wash referenced in Problem 1 and Meg did not. Now, Joe has to decide the price *p* that he will charge his customers for washing their cars because he is not sure if $13.00 is the right one. In order to do that, he improved the model he had before (described in Problem 1). Namely, for a certain price *p*, he assumes that the monthly demand is a normal random variable with mean 100 – *p* and constant standard deviation equal to 10. To provide a good quality of service he needs to hire one employee if the expected monthly demand is greater than 80, but otherwise he can manage by himself. The monthly salary that the employee makes is a random variable that is uniformly distributed between $800 and $2,000 (that is in addition to the $3 labor and materials mentioned before).

How would you set up a simulation to determine how much Joe should charge (i.e., *p*) ?

Hint: you do not need to do any math in this problem, just describe briefly how to use simulation to solve the problem.

First, create a cell that will contain the price *p*. Using Crystal Ball, we setup a random variable D representing the demand with normal distribution, mean equal to 100 – *p* and standard deviation equal to 10. In addition, we setup another random variable that will contain the salary S, distributed uniformly between 800 and 2,000. Last, we create the forecast corresponding to the monthly profit using the formula

= D (*p* – 3) + IF(100 – *p* > 80 , S , 0)

To determine the optimal value for *p*, we run different simulations for values of *p* in a range (e.g., 1, 2, 3, …, 20) and choose a value with a good tradeoff between large expected profit and small risk.

Problem 3 (10 points)

The CWD-Mart department store has sampled 225 sales records on July 14.
The average sale was $42.00, with an observed sample standard deviation of $30.00 per customer.

(a) *(5 points)* Construct a 90% confidence interval for the mean sale value.

We read the *c* value 1.645 from Table 2. Then, plugging-in the values into the CI formula:

[42 – 1.645 x 30 / sqrt(225) , 42 + 1.645 x 30 / sqrt(225)] = [38.71 , 45.29]

(b) *(5 points)* How many sales records would need to be sampled for the 90% confidence interval to be within ±$1.00 of the sample mean?

The required sample size n is $c^2 s^2 / L^2$, where L = 1. Then n = 2435.

Problem 4 (30 points)

GuardWare Inc. has installed intrusion-detection devices on the campus of a large university. The devices are very sensitive and, on any given weekend, each one has a 10% chance of being mistakenly activated when no intruder is present. Assume that the mistaken activation of different devices are independent events.

(a) There are six devices in the Administration building. Assume there will be no intruder in that building next weekend.

1. *(6 points)* What is the probability that no device will be activated next weekend?

Let X be a RV that represents the number of device mistakenly activated. X is binomial because each device is independent and all has the same probability of being mistakenly activated.

$P(X=0) = (.9)^6 = .5314$

2. *(6 points)* If two or more of the six devices in the Administration building are activated during the same weekend, the system automatically signals the police. What is the probability that this system will signal the police next weekend? *Hint: remember the assumption.*

$P(X>=2) = 1 - P(X=0) - P(X=1) = 1 - 0.5314 - 6!/5!\ (0.1)(0.9)^5 = 0.1143$

3. *(6 points)* What are the expected value and the standard deviation of the number of devices (mistakenly) activated in the Administration building next weekend?

Using the formulas for binomials: $E(X)=np=6 \times .1 = .6$
$stdev(X) = sqrt(\ np\ (1 - p\ )) = sqrt\ (\ .6 \times .9\ ) = .7348$

(b) *(6 points)* There are a total of 200 such devices on campus. Assume there will be no intruder on any of these locations next weekend. What is the probability that at least 22 of these devices will be (mistakenly) activated? (Hint: use an appropriate approximation, and explain why you may do so.)

The random variable Y can be approximated by a normal random variable because $np = 20 > 5$ and $n(1 - p\ ) >> 5$.

With the new n, $E(X)=np=200 \times .1 = 20$ and
$stdev(X) = sqrt(\ np\ (1 - p\ )) = sqrt\ (\ 20 \times .9\ ) = 4.243$

$P(\ Y >= 22\ ) = 1 - P\ (\ Y <= 22\ ) = 1 - F\ (\ (22 - 20) / 4.243\ ) = 1 - F(\ 0.4714\ ) =$
$1 - 0.6808 = 0.3192$

(c) *(6 points)* An intruder has a 5% chance of not activating the device located in the President's office. Campus police has obtained information about a student prank in preparation and, as a result, estimates that there is a 20% chance that an intruder will visit the President's office during next Columbus Day weekend. If this device is activated during that weekend, what is the probability that there is an intruder in the President's office?

Translating the statement into math, we get:

P(intruder) = 0.2
P(not activate | intruder) = 0.05

Then, P(intruder and not activate) = P(not activate | intruder) P(intruder)
= 0.2 x 0.05 = 0.01

Also, from the initial statement, we know that P(activate | no intruder) = 0.1.
Then, P(activate and no intruder) = P(activate | no intruder) P(no intruder)
= 0.1 x 0.8 = 0.08

AND	activate	not activate	total
intruder	0.19	0.01	0.2
no intruder	0.08	0.72	0.8
total	0.27	0.71	1

The answer P(intruder | activate) = P(intruder and activate) / P(activate)
= 0.19 / 0.28 = 0.7037

Problem 5 (20 points)

CWD-Tel operates mobile telephone systems in a developing country. David, a manager at CWD-Tel, wants to predict the *peak demand* for circuits, on the basis of the *number of portable phones* in a region. He has obtained the following data for 20 regions:

Region	Number of Portable Phones	Peak Demand (circuits)	Region	Number of Portable Phones	Peak Demand (circuits)
1	137,657	39,863	11	54,061	6,878
2	85,184	15,735	12	45,213	6,909
3	105,623	22,778	13	70,496	10,825
4	28,411	869	14	124,375	33,922
5	63,242	9,229	15	34,247	2,823
6	19,712	1,115	16	28,676	2,019
7	20,050	1,821	17	36,806	1,294
8	112,642	26,175	18	31,388	2,367
9	22,127	1,021	19	91,451	16,413
10	159,280	52,299	20	37,364	2,674

(Thus, for example, Region 11 had 54,061 portable phones and its peak demand was 6,878 circuits.) David used linear regression for this purpose. The resulting computer output is shown below:

SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.97533
R Square	0.95127
Adjusted R Square	0.94856
Standard Error	3381.29
Observations	20

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	4,017,125,436	4,017,125,436	351.358278
Residual	18	205,796,369	11,433,131	
Total	19	4,222,921,805		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-9,159.91	1,396.64	-6.55855	3.6647E-06
Number of Phones	0.336564	0.017955	18.74455	2.9406E-13

(a) *(5 points)* 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.

Let Y be the peak demand for circuits and X be the number of portable phones.

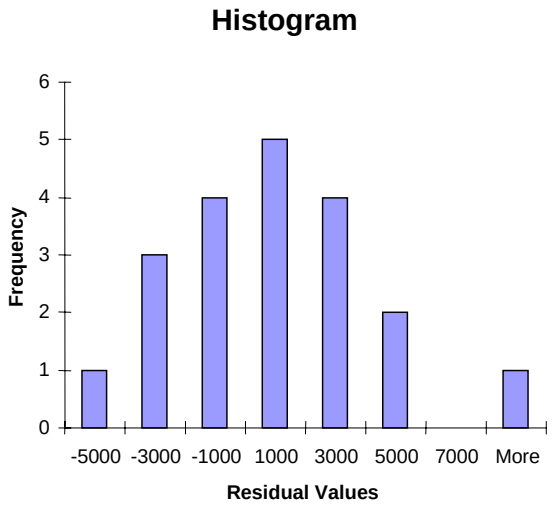
Y = – 9,159.91 + 0.336564 X + error

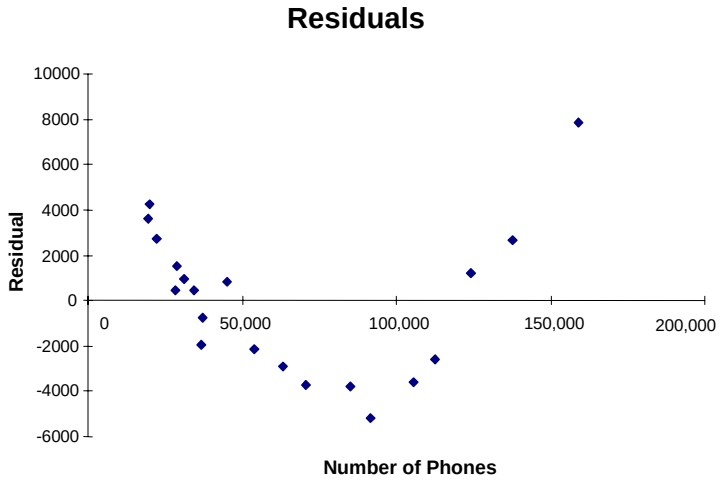
(b) *(5 points)* Interpret, in managerial terms, the meaning of each of the two regression coefficients (including the *units* in which each is expressed).

The intercept coefficient (– 9,159.91 circuits) expresses the number of circuits needed if there are no phones (X=0). With it, we know that for small demand (number of phones), we do not need many circuits.

The coefficient for X (0.336564 circuits / phone) measures how many circuits are needed for each additional phone that the network needs to serve. With it, we can estimate how much we need to invest to improve our network if we expect to incorporate new clients.

David also produced a histogram of the regression residuals, and a plot of the residuals against number of portable phones, as shown below.





For each of the statements (A) to (E) below, indicate, by circling the correct answer, whether you think the statement is true or false. If you answered "TRUE", please provide a brief justification of your answer in the space provided. (No justification is needed if you answered "FALSE").

(c) *(5 points)* These graphs show evidence of:

1. non-Normal noise.	TRUE - **FALSE**	If "TRUE", how do you know?
2. autocorrelation.	TRUE - **FALSE**	If "TRUE", how do you know?
3. overspecification.	TRUE - **FALSE**	If "TRUE", how do you know?
4. multicollinearity.	TRUE - **FALSE**	If "TRUE", how do you know?
5. heteroscedasticity.	**TRUE** - FALSE	If "TRUE", how do you know?

The plot of residuals does not look as a random cloud with the same amplitude along the *x* axis.

(d) *(5 points)* Explain how you would go about correcting the most serious problem of the ones you found in part (c). If applicable, *how* would you modify the data and/or the model to implement this change?

The most likely cause that the effect in the graph is produced is a non-linearity in the relation between the number of phones and the peak number of circuits. The shape suggests that squaring the number of phones might give better results.

TABLE 1: Cumulative distribution function of the standard Normal distribution (for $z \geq 0$).
If Z is a standard Normal random variable, then $F(1.34) = P(Z \leq 1.34) = 0.9099$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

For a negative value of *z*, use the formula $F(z)=1- F(-z)$.
For example, $F(-0.7) = 1 - F(0.7) = 1 - 0.758 = 0.242$.

TABLE 2: The *c* value for a *t*-distribution with *k* degrees of freedom (dof) and a standard Normal distribution.

E.g., for 15 degrees of freedom and 95% confidence level, $c = 2.131$, that is, $P(-2.131 \le T \le 2.131) = 0.95$. When dof is 30 or more, we use the Normal distribution.

Degrees of freedom (k)	90%	95%	98%	99%
1	6.314	12.706	31.821	63.656
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
11	1.796	2.201	2.718	3.106
12	1.782	2.179	2.681	3.055
13	1.771	2.160	2.650	3.012
14	1.761	2.145	2.624	2.977
15	1.753	2.131	2.602	2.947
16	1.746	2.120	2.583	2.921
17	1.740	2.110	2.567	2.898
18	1.734	2.101	2.552	2.878
19	1.729	2.093	2.539	2.861
20	1.725	2.086	2.528	2.845
21	1.721	2.080	2.518	2.831
22	1.717	2.074	2.508	2.819
23	1.714	2.069	2.500	2.807
24	1.711	2.064	2.492	2.797
25	1.708	2.060	2.485	2.787
26	1.706	2.056	2.479	2.779
27	1.703	2.052	2.473	2.771
28	1.701	2.048	2.467	2.763
29	1.699	2.045	2.462	2.756
Standard Normal	1.645	1.960	2.326	2.576