MSCI:9110 *Advanced Analytics*
Kurt Anstreicher
Spring, 2017

FINAL EXAM

Name: SOLUTION

This is a two and one-half hour, open book and notes exam. There are a total of 85 points. In turning in this exam you attest that all work is your own and you completed the exam without the assistance of another person. Be sure to show your work to receive partial credit. Good Luck!

Question 1 (30 points)

Weekly grocery store scanner data for a metropolitan market area has been obtained for a 3-year period (156 weeks) to analyze consumer behavior with regard to purchasing ketchup. The data consists of average price ($/pound) and market share (coded from 0 to 1) for 4 brands of ketchup. Brand 3 is the leading national brand. (So for example if Share 3=.5, then 50% of the sales volume in a particular week was for Brand 3.)

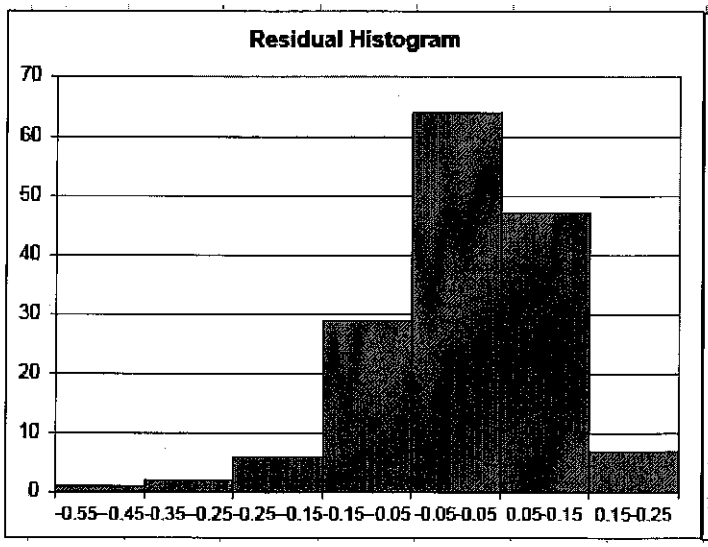
Below is the result of a multiple regression of the natural logarithm of Share3 on the natural logarithms of the prices for the 4 brands of ketchup (plots of the residuals against the 4 independent variables are on the next page).

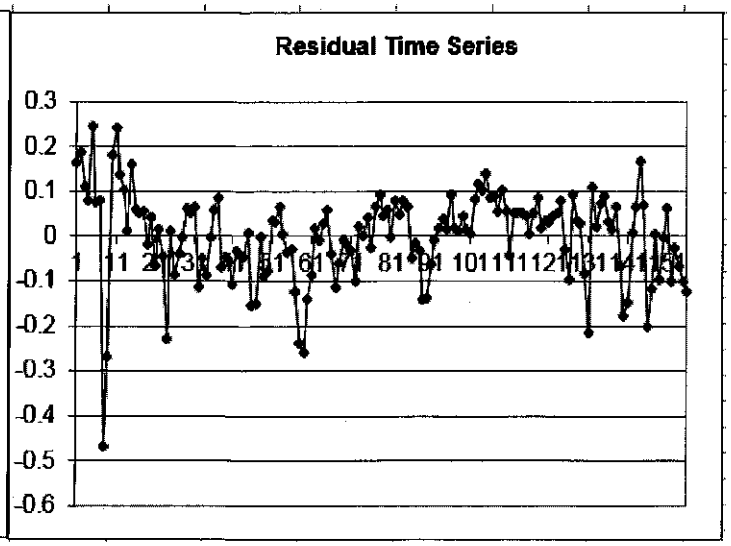
Regression Statistics	
Multiple R	0.7367
R Square	0.5428
Adjusted R Square	0.5307
Standard Error	0.1009
Observations	156

ANOVA

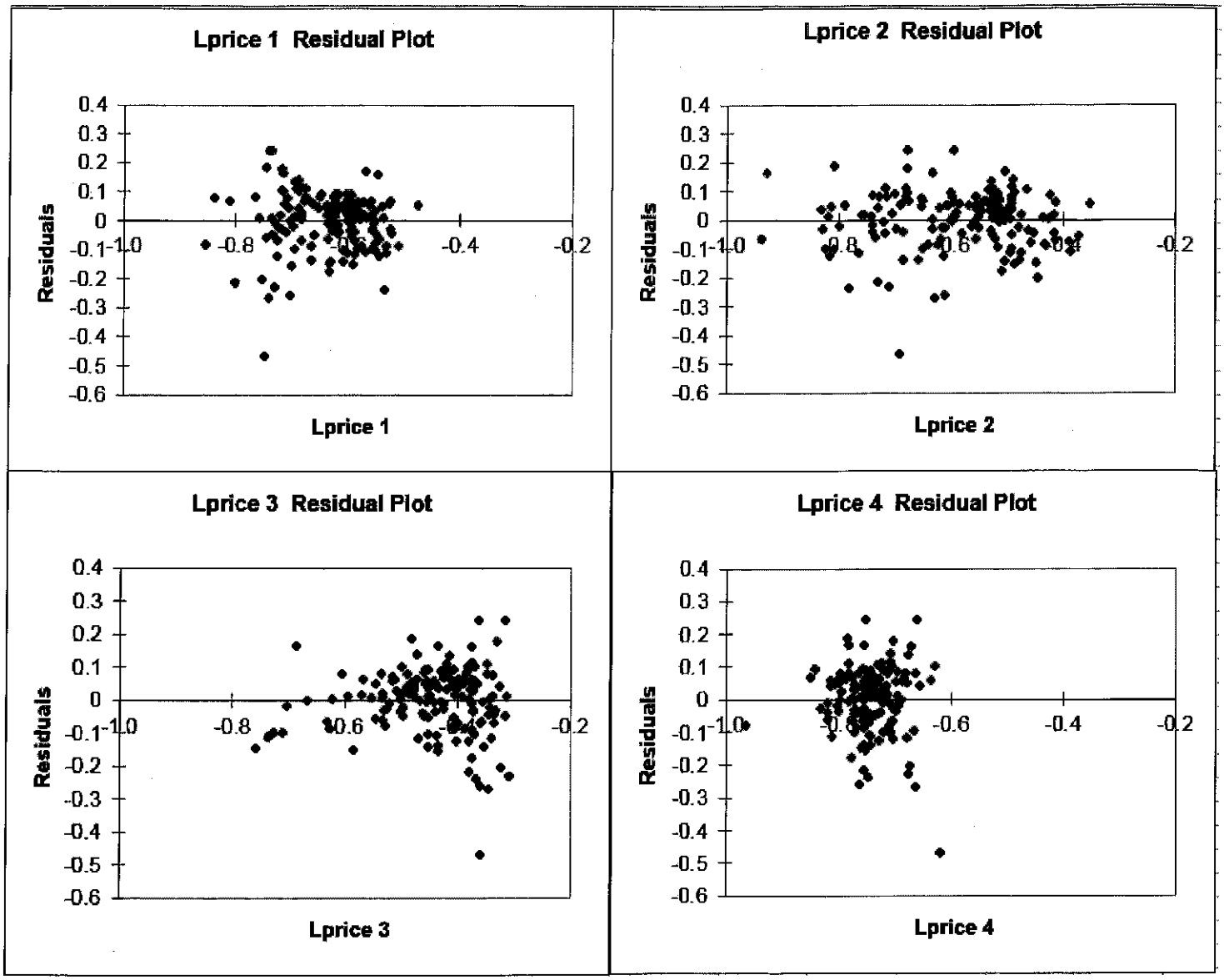
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	1.8233	0.4558	44.8117	0.0000
Residual	151	1.5360	0.0102		
Total	155	3.3593			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.0466	0.1461	0.3193	0.7499	-0.2420	0.3352
Lprice 1	0.7048	0.1142	6.1699	0.0000	0.4791	0.9305
Lprice 2	0.2817	0.0692	4.0721	0.0001	0.1450	0.4184
Lprice 3	-0.9167	0.0894	-10.2510	0.0000	-1.0934	-0.7400
Lprice 4	0.2697	0.1581	1.7062	0.0900	-0.0426	0.5820





Autocorrelation = .464



A. (5 points) Do the signs (positive/negative) of the regression slope coefficients make sense? Explain.

YES. AS THE PRICE OF BRAND 3 GOES UP, THE MARKET SHARE OF BRAND 3 GOES DOWN. ON THE OTHER HAND, AS THE PRICE OF ONE OF THE OTHER BRANDS GOES UP, THE MARKET SHARE OF BRAND 3 GOES UP.

B. (5 points) One of the three other brands (brands 1,2 and 4) is a local "store brand" that is not a national brand. Based on the regression output, which one is the store brand? Explain.

BRAND 4. NOTE THAT THE SLOPE COEFFICIENT FOR LPRICE4 IS THE SMALLEST OF THE THREE OTHER BRANDS, AND IS NOT SIGNIFICANTLY DIFFERENT FROM ZERO. FROM THE RESIDUAL PLOT IT ALSO APPEARS THAT THE AVERAGE PRICE OF BRAND 4 IS THE LOWEST.

C. (5 points) Describe any problems that you see in the residuals of this regression.

THE AUTOCORRELATION OF THE RESIDUALS IS HIGHER THAN DESIRED, AND THE HISTOGRAM SHOWS OBVIOUS SKEW. THE RESIDUALS PLOTTED AGAINST LPRICE3 ALSO LOOK ODD AT THE LOWEST PRICE LEVELS.

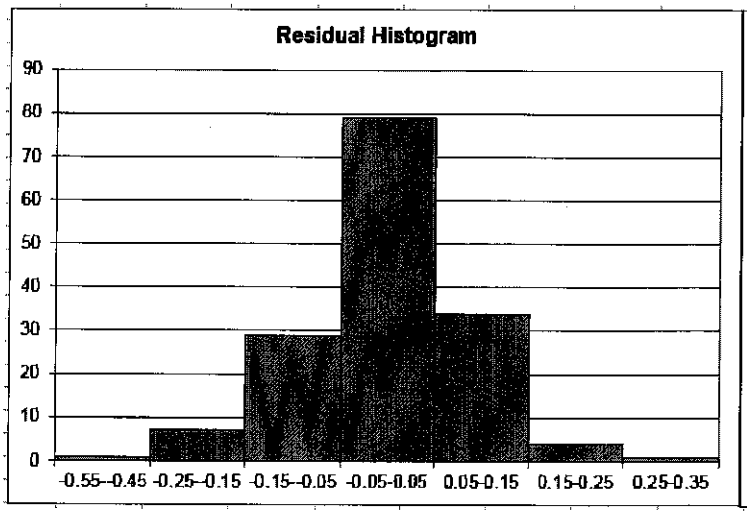
A second analysis adds the lag of LN(Share3) as an additional independent variable. The results of that regressions are shown below (the residual plots against the four LPrice variables are on the next page).

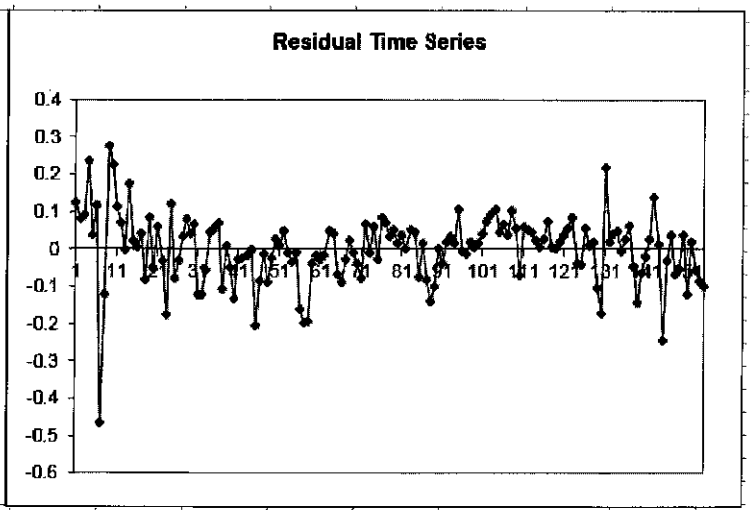
Regression Statistics	
Multiple R	0.78000
R Square	0.60841
Adjusted R Square	0.59527
Standard Error	0.09325
Observations	155

ANOVA

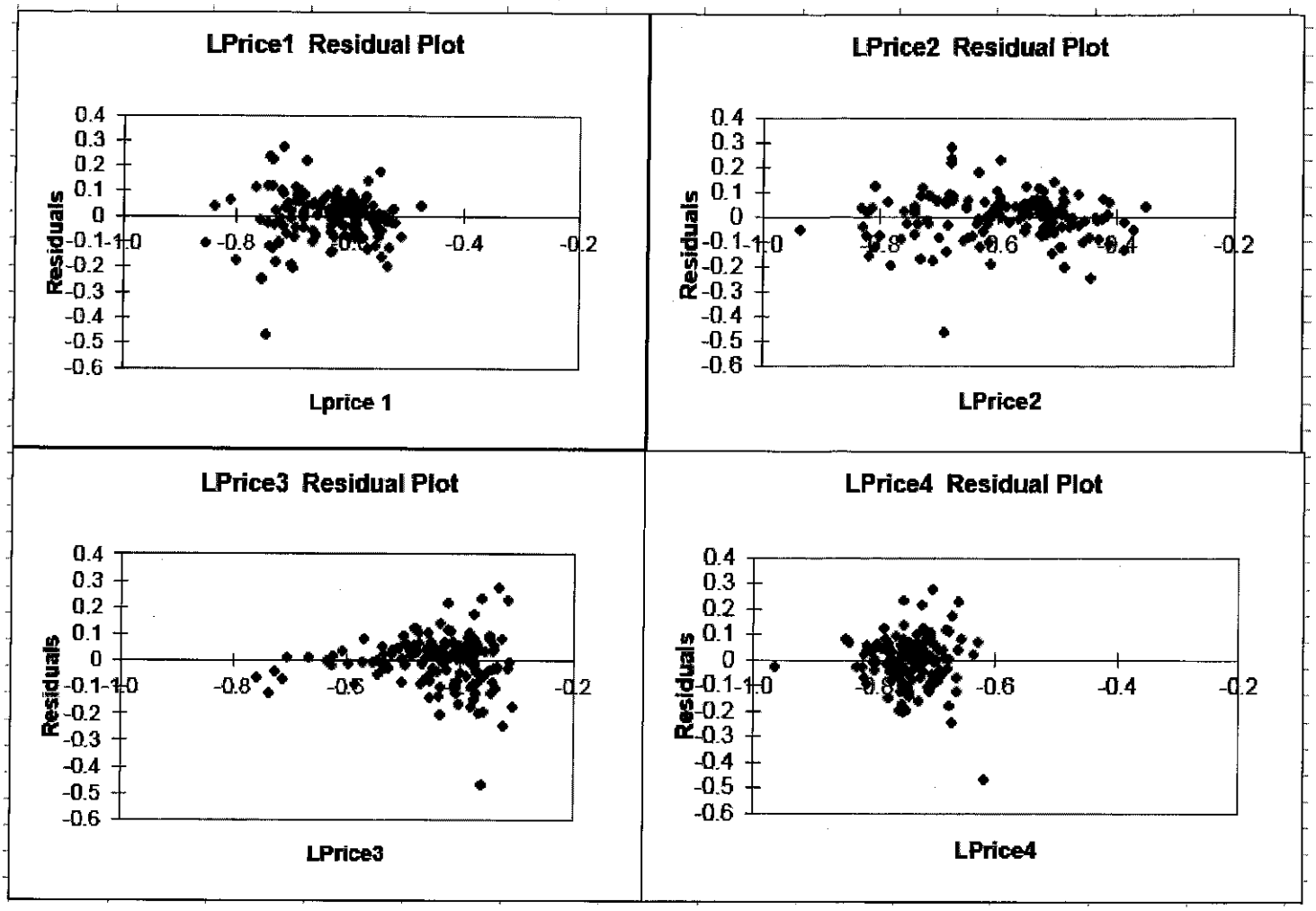
	df	*SS*	*MS*	*F*	*Sig. F*
Regression	5	2.01283	0.40257	46.29952	0.00000
Residual	149	1.29553	0.00869		
Total	154	3.30835			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.12371	0.13780	0.89779	0.37075	-0.14858	0.39600
LagLshare3	0.25891	0.05272	4.91124	0.00000	0.15474	0.36308
LPrice1	0.64303	0.10697	6.01116	0.00000	0.43165	0.85441
LPrice2	0.26901	0.06671	4.03244	0.00009	0.13719	0.40083
LPrice3	-0.84572	0.08544	-9.89875	0.00000	-1.01455	-0.67690
LPrice4	0.27014	0.14625	1.84718	0.06670	-0.01884	0.55913





Autocorrelation = .237



D. (5 points) Describe ways in which the second regression appears to be better than the first.

THE FIT TO THE DATA IS BETTER - R^2 IS HIGHER AND THE STANDARD ERROR IS LOWER.

THE RESIDUAL AUTOCORRELATION IS SUBSTANTIALLY LOWER, AND THE HISTOGRAM IS ALMOST PERFECTLY SYMMETRIC.

E. (5 points) Suppose that there is a sale on Brand 3 where the average price is cut by 20%. Based on the second regression, what effect will this have on the market share for Brand 3? (An approximate answer is sufficient.)

FOR A LOG-LOG REGRESSION, THE SLOPE COEFFICIENT GIVES THE APPROXIMATE % CHANGE IN THE DEPENDENT VARIABLE PER % CHANGE IN THE INDEPENDENT VARIABLE. THEREFORE THE APPROXIMATE % CHANGE IN SHARE 3 WOULD BE

$$-20(-.84572) = 16.91,$$

OR AN INCREASE OF ABOUT 17%. THE EXACT FACTOR CHANGE WOULD BE

$$.8^{-.84572} = 1.208,$$

OR AN INCREASE OF ALMOST 21%.

F. (5 points) Both regressions show a tendency to have negative residuals for the smallest values of Price3. Explain why this might tend to occur.

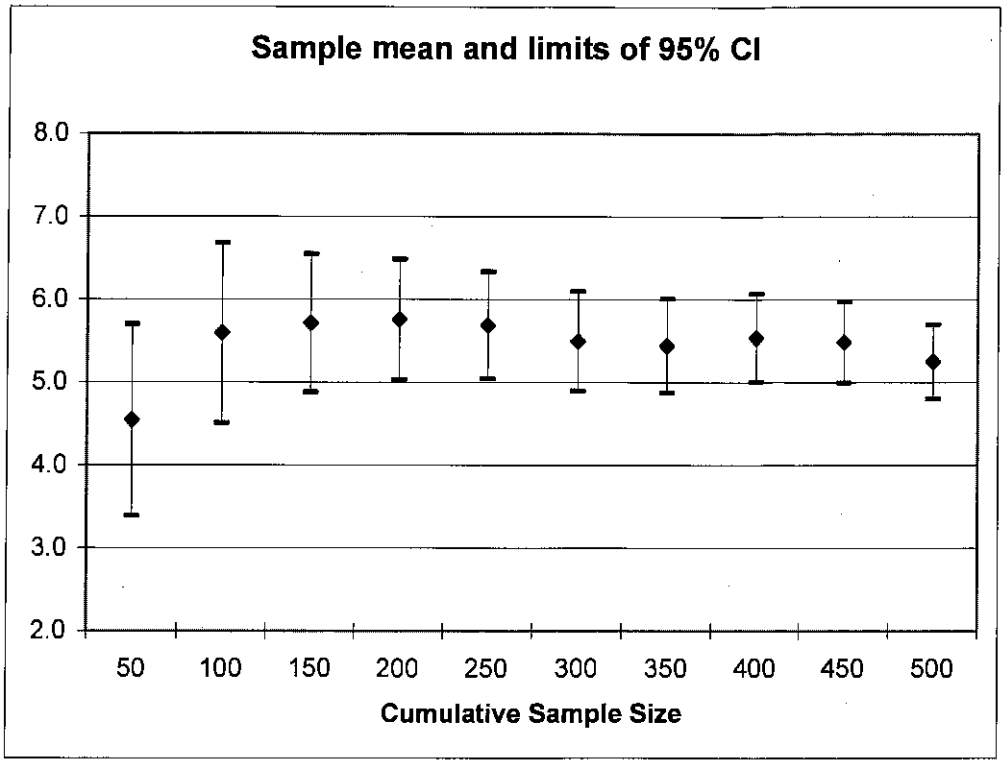
THE MODEL PREDICTS THAT SHARE 3 GOES UP AS PRICE 3 GOES DOWN. HOWEVER, AS PRICE 3 GETS LOWER AND LOWER, SHARE 3 CANNOT EXCEED 1.0 (100%), AND AT THE LOWEST PRICE GIVES SHARE 3 WILL BE CLOSE TO 100% SINCE IT IS THE LEADING BRAND. THE MODEL WILL LIKELY PREDICT VALUES OF SHARE 3 THAT ARE OVER 100% AT THESE LOWEST PRICE LEVELS, RESULTING IN NEGATIVE RESIDUALS.

Question 2 (20 points)

A call center has a contract to handle help calls for an appliance manufacturer. The contract was negotiated with the assumption that the calls would require an average of 5 minutes each. Some managers at the call center believe that the contract needs to be renegotiated because the calls are taking longer than 5 minutes on average. In order to study the problem the following "sequential sampling" scheme is proposed. A random sample of 50 calls will be obtained, and these calls will be used to construct a 95% confidence interval for the true mean length of the calls. If there is a significant difference between the sample mean and the assumed mean of 5 minutes (that is, if the value 5.0 is outside the 95% confidence interval), then the study will be stopped. Otherwise another random sample of 50 calls will be obtained and added to the first 50, making a total sample of 100 calls. If the sample mean for these 100 calls is significantly different from 5.0 (continuing to work at the 5% level of significance) then the sampling process will be stopped. Otherwise *another* sample of 50 random calls will be obtained, making for a total sample of 150 calls. If the sample mean for these 150 calls is significantly different from 5.0 then the process will be stopped. Otherwise the process will be continued by taking another sample of 50 calls, and so on, up to a maximum total sample size of 500 calls.

A quality control engineer claims that this proposed procedure is incorrect because it is far too likely to obtain a "significant" difference even when the true mean duration of help calls does *not* differ from the assumed mean of 5.0 minutes. To demonstrate this she has written a Crystal Ball simulation. The simulation generates 10 samples of 50 calls, where the length of each call has an exponential distribution with mean 5.0 minutes (the exponential distribution is often used to model the time required to complete a task). For each sample of 50 calls the simulated data is used to compute a 95% confidence interval for the mean *using all of the data up to that point*, exactly as in the proposed sequential sampling procedure. For each of these confidence intervals the simulation checks if the assumed mean of 5.0 is outside the confidence interval, corresponding to a significant difference between the sample mean and the assumed mean. The forecast cell in the simulation (C12) is the number of such significant differences. The worksheet containing the simulation (not showing the 500 values for the call durations in the 10 samples of size 50) is shown below, and a chart displaying the 10 confidence intervals is on the next page. Note that for the particular outcomes shown, the sample mean is significantly different from the value 5.0 for the samples with 200, 250 calls and 400 calls, so in this case the sequential sampling procedure would actually have stopped after 200 calls were obtained.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Call Center Simulation**											
2	Underlying distribution assumed to be exponential, mean=5.											
3												
4	Cumulative sample size	50	100	150	200	250	300	350	400	450	500	
5	Sample Mean	4.542	5.593	5.715	5.758	5.688	5.497	5.441	5.539	5.486	5.251	
6	Sample Std. Dev.	4.175	5.547	5.214	5.257	5.202	5.307	5.407	5.432	5.295	5.142	
7	Std. Error	0.590	0.555	0.426	0.372	0.329	0.306	0.289	0.272	0.250	0.230	
8	95% CI Lower Limit	3.384	4.506	4.880	5.029	5.044	4.896	4.875	5.007	4.997	4.800	
9	95% CI Upper Limit	5.699	6.680	6.549	6.486	6.333	6.097	6.007	6.071	5.976	5.702	
10	Significant Difference?	0	0	0	1	1	0	0	1	0	0	
11												
12	Number of Significant Differences		3									
13												



A. (5 points) Explain why in the above chart the confidence intervals are getting smaller going from left to right.

EACH CONFIDENCE INTERVAL HAS THE FORM

$\bar{X} \pm 1.96\, s/\sqrt{n}$,

WHERE $\bar{X}$ IS THE SAMPLE MEAN AND s IS THE SAMPLE STD. DEV. THE CONFIDENCE INTERVALS GET SMALLER DUE TO THE $\sqrt{n}$ IN THE DENOMINATOR

B. (5 points) What formula is in cell H8 of the spreadsheet?

THIS IS $\bar{X} - 1.96\, s/\sqrt{n}$,

↑ ↑ ↑

H5 H6 H4

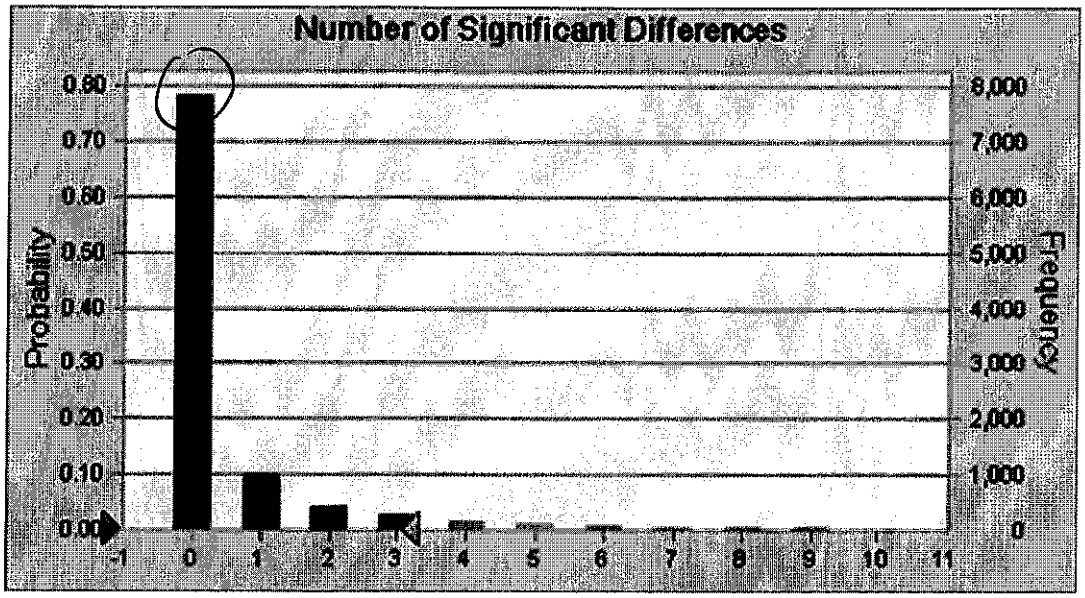
SO THE FORMULA IN H8 IS

= H5 - 1.96 * H6/SQRT(H4)

ALTERNATIVELY ONE COULD REFERENCE THE STD. ERROR (= $s/\sqrt{n}$) IN CELL H7 AND WRITE

= H5 - 1.96 * H7

C. (5 points) Below is the frequency distribution for the forecast cell (Number of Significant Differences) for a CB simulation with 10,000 trials. Based on this simulation, what is the approximate probability that the sequential sampling procedure will terminate with a "significant" difference when the true mean is actually the assumed value 5.0?



APPROXIMATELY 22%

D. (5 points) Describe how you could modify the sequential sampling procedure in a simple way, and use the Crystal Ball simulation, to end up with a procedure that actually has only a 5% chance of terminating with a significant difference when the true mean length of the calls is 5.0 minutes.

THE SIMPLEST WAY TO DO THIS IS TO INCREASE THE CONFIDENCE LEVEL USED FOR EACH OF THE 10 INDIVIDUAL CONFIDENCE INTERVALS UNTIL "ZERO SIGNIFICANT DIFFERENCES" OCCURS 95% OF THE TIME. ALL THAT THIS REQUIRES IS INCREASING THE FACTOR "1.96" THAT IS CURRENTLY USED UNTIL THE SIMULATION OUTPUT OBTAINS A PROBABILITY OF 95% FOR ZERO SIGNIFICANT DIFFERENCES.

Question 3 (35 points)

A manufacturer of birdseed needs to decide on the blend of ingredients to use in producing their Economy Songbird Mix. The possible ingredients are sunflower seeds, white millet, cracked corn, safflower, and canary grass seed. The specifications of Economy Songbird Mix are that it should contain at least 12% protein and 12% fat, and no more than 20% fiber. The content of these three nutrients in each of the possible ingredients, as well as the costs of the ingredients (in $/pound) are given in the spreadsheet below. (For example, one pound of sunflower seeds contains .17 Lb of protein, .26 Lb of fat and .29 Lb of fiber, and costs $.22).

	A	B	C	D	E	F	G
1							
2		Nutrient content of Ingredients					Ingredient amount
3	Ingredient	Protein	Fat	Fiber	Cost $/Lb		per pound of mix
4	Sunflower seeds	0.17	0.26	0.29	0.22		0.368293
5	White millet	0.12	0.04	0.08	0.19		0
6	Cracked corn	0.09	0.04	0.03	0.07		0.512195
7	Safflower	0.18	0.18	0.25	0.26		0
8	Canary grass seed	0.12	0.04	0.11	0.11		0.119512
9							
10	Mix content	0.12000	0.12000	0.13366	0.13002		1.0
11	Constraint for mix	0.12	0.12	0.20			1.0

A linear programming model has been built to determine the least-cost mix of ingredients that meets the specifications for Economy Songbird Mix. The variables, in cells G4:G8, are the amounts of each ingredient to use in one pound of Economy Songbird Mix. The objective, in cell E10, is the cost for one pound of mix. The solution obtained by Solver is shown in the spreadsheet above, and the Solver Sensitivity Report is reproduced below.

Adjustable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
G4	Sunflower seeds per pound of mix	0.368293	0	0.22	0.074231	0.051176
G5	White millet per pound of mix	0	0.078602	0.19	1E+30	0.078602
G6	Cracked corn per pound of mix	0.512195	0	0.07	0.039000	0.034800
G7	Safflower per pound of mix	0	0.046309	0.26	1E+30	0.046309
G8	Canary grass seed per pound of mix	0.119512	0	0.11	0.020714	0.038649

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
B10	Mix content Protein	0.12	1.162602	0.12	0.017182	0.003973
C10	Mix content Fat	0.12	0.235772	0.12	0.010500	0.075600
D10	Mix content Fiber	0.133661	0	0.2	1E+30	0.066339
G10	Mix content per pound of mix	1.0	-0.037780	1	0.056257	0.156328

A. (5 points) What is the cost per pound for the least-cost formulation of Economy Songbird Mix obtained by Solver?

$.13/LB

B. (5 points) What formula is in cell B10 of the spreadsheet?

=SUMPRODUCT(B4:B8, G4:G8)

NOTE THAT THE UNITS FOR EACH TERM IN THIS SUMPRODUCT ARE $\frac{LB\ PROTEIN}{LB\ INGREDIENT} \cdot \frac{LB\ INGREDIENT}{LB\ MIX} = \frac{LB\ PROTEIN}{LB\ MIX}$

C. (10 points) Suppose that the cost of Sunflower seeds increases to $.25/LB while the cost of cracked corn decreases to $.06/LB. Will the current solution remain optimal? Explain. What is the minimum cost per pound for Economy Songbird Mix in this case?

THIS IS AN INCREASE OF .03 IN THE OBJECTIVE COEFF. FOR SUNFLOWER SEEDS, AND A DECREASE OF .01 FOR CRACKED CORN. APPLYING THE 100% RULE,

$$\sum \frac{\text{CHANGE}}{\text{RANGE}} = \frac{.03}{.074} + \frac{.01}{.035} = .691 < 1.0,$$

SO THE SOLUTION REMAINS OPTIMAL. THE MINIMUM COST/LB WILL NOW BE

$$.25(.368293) + .06(.512195) + .11(.119512)$$

$$= .13595$$

D. (10 points) Suppose that the costs of the ingredients are as originally specified, but the nutritional specifications of Economy Songbird Mix are changed so that the fat requirement is 0.06 Lb per pound of mix, with the requirements for protein and fiber unchanged. Determine what the least-cost formulation would cost in this case, and explain why your calculation is valid.

THIS IS A DECREASE IN THE RHS OF THE "MIX CONTENT FAT" CONSTRAINT OF .06 (FROM .12 TO .06). SINCE THE ALLOWABLE DECREASE IS .07576, THE SHADOW PRICE APPLIES AND WILL GIVE THE CHANGE IN THE OBJECTIVE:

$$-.06(.235772) = -.0141.$$

THE LEAST COST FORMULATION WILL THEREFORE COST

$$.13 - .0141 = \$.1159/LB$$

E. (5 points) Suppose that Economy Songbird Mix is actually advertised to contain no more than 50% cracked corn, so the above solution is not a feasible mix. True or False: the effect of adding this constraint on the minimum cost for Economy Songbird Mix can be determined from the given solution output.

True: ____

False: X