Sample Test #1

SHOW ALL WORK
Each problem is worth 25 points
Output follows each problem

1. One possible effect of air pollution is genetic damage. One group of mice are exposed to air near a steel mill and another group to air in a rural area. Gene mutations are then compared.

	Location	
Mutation	Steel Mill Air	Rural Air
Yes	30	23
No	66	127

a) Write a SAS program to determine if there is a relation between location and mutation.

b) Show how the quantity 'chi-square' statistic value is obtained in the output from other values contained in the output and explain where the 'probability' comes from.

c) Does there appear to be a relationship between location and mutation. Explain.

The FREQ Procedure

Table of mutation by air

mutation	air		
Frequency Cell Chi-Square Percent Row Pct Col Pct	SM	R	Total
Y	30	23	53
	4.1971	2.6861	
	12.20	9.35	21.54
	56.60	43.40	
	31.25	15.33	
N	66	127	193
	1.1526	0.7376	
	26.83	51.63	78.46
	34.20	65.80	
	68.75	84.67	
Total	96	150	246
	39.02	60.98	100.00

The FREQ Procedure

Statistics for Table of mutation by air

Statistic	DF	Value	Prob
Chi-Square	1	8.7734	0.0031
Likelihood Ratio Chi-Square	1	8.5908	0.0034
Continuity Adj. Chi-Square	1	7.8570	0.0051
Mantel-Haenszel Chi-Square	1	8.7378	0.0031
Phi Coefficient		0.1888	
Contingency Coefficient		0.1856	
Cramer's V		0.1888	

Fisher's Exact Test

Cell (1,1) Frequency (F)	127
Left-sided Pr <= F	0.9990
Right-sided Pr >= F	0.0027
Table Probability (P)	0.0017
Two-sided Pr <= P	0.0041

Sample Size = 246

2. The following data have been collected on 5 subjects where in the gender category '1' means male and '0' means female:

ID	Age	Gender	GPA	Cscore
1	18	1	3.7	650
2	18	0	3.3	490
3	19	1	2.8	580
4	23	1	2.8	530
5	21	1	3.5	640

a) Write the necessary SAS code with the data included in order to create a SAS data file.

b) The data in the interior of the table is included in a file called 'c:\cscore'. Write the necessary SAS code to create a SAS data set for these data using this file.

c) Give the SAS code to calculate the mean and standard deviation of the 'cscore' variable.

d) Give the SAS code necessary to regress the explanatory variables 'gender', 'age' and 'GPA' on the response variable 'cscore' .

e) Use the output to determine which variables appear to be important to the model. Give reasons.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
gpa	5	3.2200000	0.4086563	2.8000000	3.7000000
cscore	5	578.0000000	69.0651866	490.0000000	650.0000000

The REG Procedure
Model: MODEL1
Dependent Variable: cscore

Number of Observations Read	5
Number of Observations Used	5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	18378	6125.92073	8.72	0.2428
Error	1	702.23781	702.23781		
Corrected Total	4	19080			

Root MSE	26.49977	R-Square	0.9632
Dependent Mean	578.00000	Adj R-Sq	0.8528
Coeff Var	4.58474		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	314.07816	229.24472	1.37	0.4014	0
age	1	-7.41483	7.86898	-0.94	0.5189	1.65772
gpa	1	93.75418	37.19994	2.52	0.2405	1.31637
gender	1	136.05878	33.79026	4.03	0.1550	1.30073

3. The average oral vocabulary size of children at various ages was collected to investigate the development of language in students. The data are given on the following page with the age and number of words. Also the SAS output is given.

   a. Give the estimated coefficient of the number of words.

   b. What would you say about the claim that the estimate of the intercept is zero.

   c. On the output labeled 'Fit Diagnostics for Age'
      i) explain what the graph 'Quantile vs. Residual' gives and
      ii) explain what the graph 'Residual vs. Percent' gives.

   d) (Undergraduates only) Give the estimated number of words for a child whose age is 5 ½ years old.

   e) (Graduate students only) For the output labeled 'Fit Plot for Age' explain what the shaded area and the dotted lines represent.

Obs	age	number
1	1.0	3
2	1.5	22
3	2.0	272
4	2.5	446
5	3.0	896
6	3.5	1222
7	4.0	1540
8	4.5	1870
9	5.0	2072
10	6.0	2562

The REG Procedure
Model: MODEL1
Dependent Variable: number

Number of Observations Read	10
Number of Observations Used	10

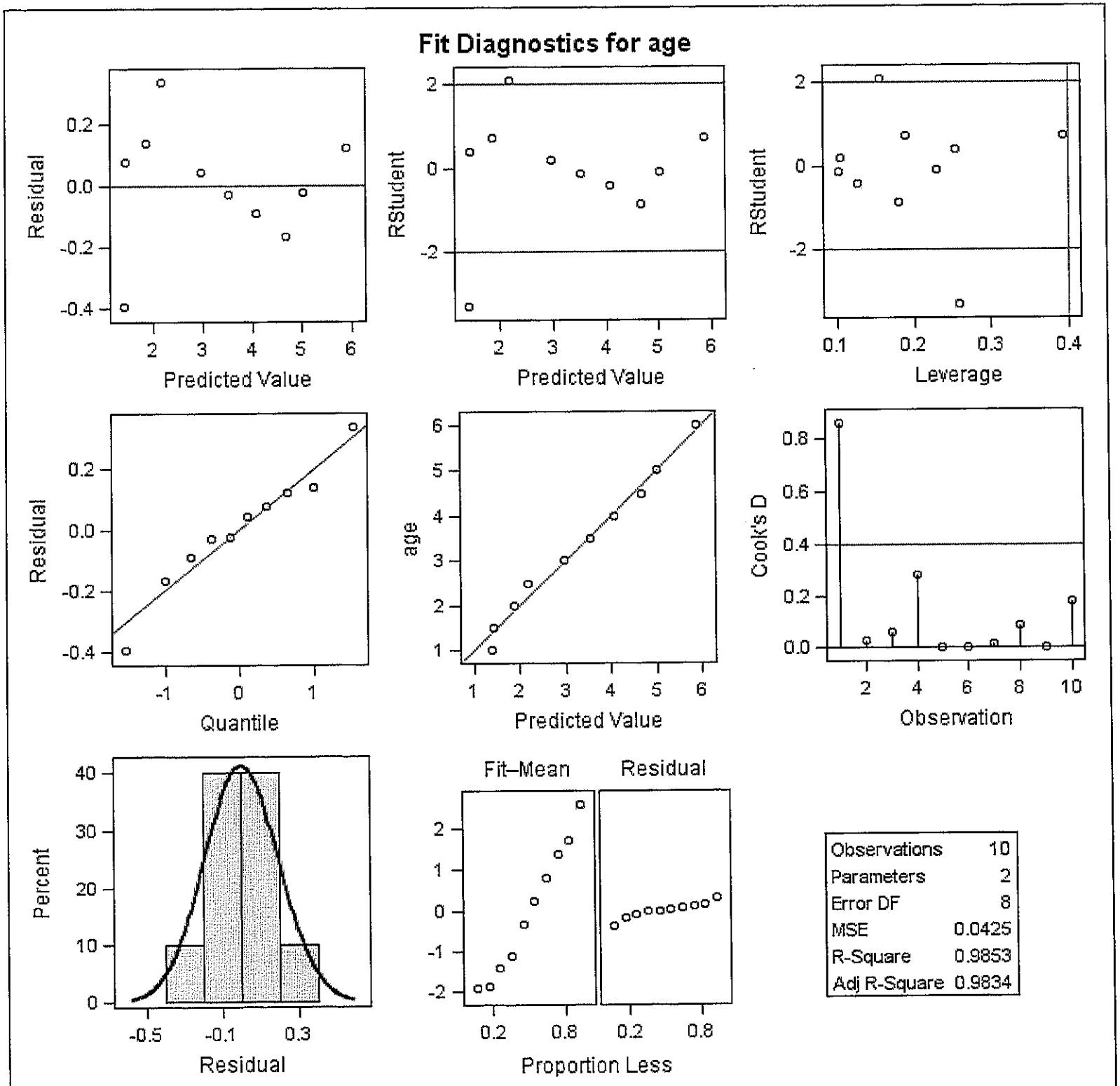
Analysis of Variance

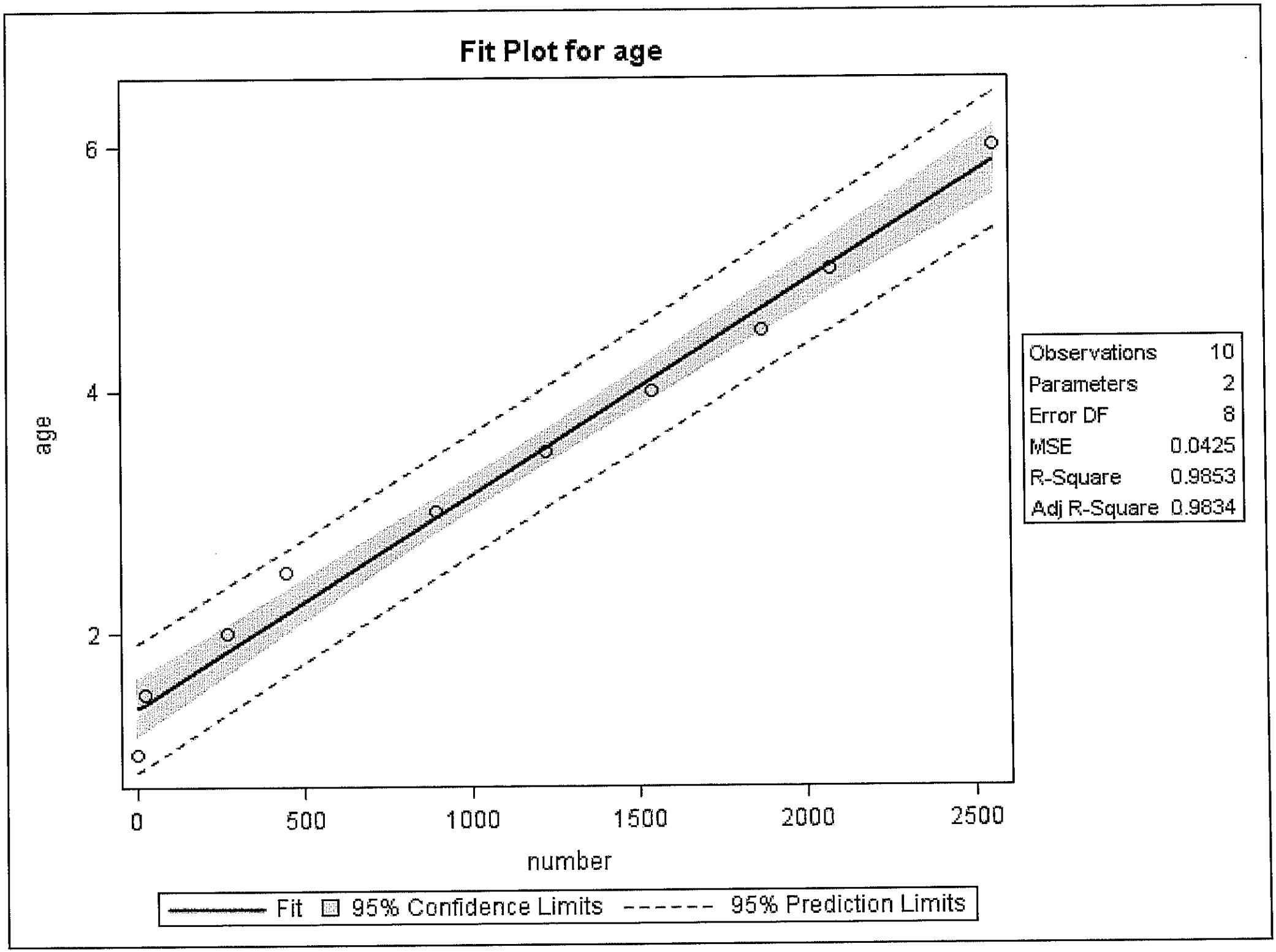
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7294086	7294086	535.18	<.0001
Error	8	109033	13629		
Corrected Total	9	7403119			

Root MSE	116.74372	R-Square	0.9853
Dependent Mean	1090.50000	Adj R-Sq	0.9834
Coeff Var	10.70552		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-763.85714	88.24996	-8.66	<.0001
age	1	561.92641	24.29000	23.13	<.0001


Fit Diagnostics for age
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
age
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 10
Parameters 2
Error DF 8
MSE 0.0425
R-Square 0.9853
Adj R-Square 0.9834


Fit Plot for age
age
number
6
4
2
0
500
1000
1500
2000
2500
Observations 10
Parameters 2
Error DF 8
MSE 0.0425
R-Square 0.9853
Adj R-Square 0.9834
Fit
95% Confidence Limits
95% Prediction Limits

4. We consider the data of the homework problem #2 on page 97for weight gain in rats. The output follows the problem.

   a) Assume the data is stored in 'c:\rats.txt' give the SAS code for the data step.

   b) Write the code for the table given in the second page of the output.

   c) From the first ANOVA table describe why the model has 3 degrees of freedom.

   d) Use the data in the first ANOVA table to describe whether this is a good model or not.

   e) Are the variances equal or not? Explain why.

   f) Explain what is the meaning of the three p-values in the source subtable of the second anova table.

Obs	id	source	amount	weightgain	cell	
1	1	beef	low	90	beef	low
2	2	beef	high	73	beef	high
3	3	cereal	low	107	cereal	low
4	4	cereal	high	98	cereal	high
5	5	beef	low	76	beef	low
6	6	beef	high	102	beef	high
7	7	cereal	low	95	cereal	low
8	8	cereal	high	74	cereal	high
9	9	beef	low	90	beef	low
10	10	beef	high	118	beef	high
11	11	cereal	low	97	cereal	low
12	12	cereal	high	56	cereal	high
13	13	beef	low	64	beef	low
14	14	beef	high	104	beef	high
15	15	cereal	low	80	cereal	low
16	16	cereal	high	111	cereal	high
17	17	beef	low	86	beef	low
18	18	beef	high	81	beef	high
19	19	cereal	low	98	cereal	low
20	20	cereal	high	95	cereal	high
21	21	beef	low	51	beef	low
22	22	beef	high	107	beef	high
23	23	cereal	low	74	cereal	low
24	24	cereal	high	88	cereal	high
25	25	beef	low	72	beef	low
26	26	beef	high	100	beef	high
27	27	cereal	low	74	cereal	low
28	28	cereal	high	82	cereal	high
29	29	beef	low	90	beef	low
30	30	beef	high	87	beef	high
31	31	cereal	low	67	cereal	low
32	32	cereal	high	77	cereal	high
33	33	beef	low	95	beef	low
34	34	beef	high	117	beef	high
35	35	cereal	low	89	cereal	low
36	36	cereal	high	86	cereal	high
37	37	beef	low	78	beef	low
38	38	beef	high	111	beef	high
39	39	cereal	low	58	cereal	low
40	40	cereal	high	92	cereal	high

		weightgain		
		Mean	Std	N
source	amount			
beef	high	100.00	15.14	10.00
	low	79.20	13.89	10.00
cereal	high	85.90	15.02	10.00
	low	83.90	15.71	10.00

The ANOVA Procedure

Dependent Variable: weightgain

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2404.10000	801.36667	3.58	0.0230
Error	36	8049.40000	223.59444		
Corrected Total	39	10453.50000			

R-Square	Coeff Var	Root MSE	weightgain Mean
0.229980	17.13819	14.95307	87.25000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
cell	3	2404.100000	801.366667	3.58	0.0230

The ANOVA Procedure

Levene's Test for Homogeneity of weightgain Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
cell	3	12289.3	4096.4	0.07	0.9775
Error	36	2230875	61968.7		

The ANOVA Procedure

Dependent Variable: weightgain

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2404.10000	801.36667	3.58	0.0230
Error	36	8049.40000	223.59444		
Corrected Total	39	10453.50000			

R-Square	Coeff Var	Root MSE	weightgain Mean
0.229980	17.13819	14.95307	87.25000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
source	1	220.900000	220.900000	0.99	0.3269
amount	1	1299.600000	1299.600000	5.81	0.0211
source*amount	1	883.600000	883.600000	3.95	0.0545

The ANOVA Procedure

Level of source	Level of amount	N	weightgain Mean	weightgain Std Dev
beef	high	10	100.000000	15.1364167
beef	low	10	79.200000	13.8868443
cereal	high	10	85.900000	15.0218360
cereal	low	10	83.900000	15.7088086

Solutions to Sample Test #1

1a)
```
data pollution;
    input mutation $ air $ n;
cards;
    Y SM 30
    Y  R 23
    N SM 66
    N  R 127;
proc freq order = data;
    tables mutation*air/chisq cellchi;
    weight n;
run;
```

b) Sum of cell chi-squares, chi-square table

c) Yes, because p-value=0.0031, we reject 'no relation'

2a)
```
data cscore;
    input age gender gpa cscore;
    cards;
    18 1 3.7 650
    18 0 3.3 490
     .
     .
    21 1 3.5 640;    run;
```

b)
```
data cscore;
    infile 'C:\cscore';
    input age gender gpa cscore;
run;
```

c)
```
proc means;
    var gpa cscore;
run;
```

d) `proc reg; model cscore=age gpa gender/vif; run;`

e) None of the variables seem to be important because of high p-values

3a) 561.93 (Problem should be "Give the estimated coefficient of age.")

b) Reject claim because of low p-value

c) i) Almost a straight line so normal ii) normal shape

d) y=-763.9 + 561.9*(5.5) = 2326.6

e) Shaded area is 95% CI for regression line and dotted line is 95% prediction limits

4a)
```
data rats;
    infile 'C:/rats.txt';
    input id source $ amount $ weightgain ;
run;
```

b)
```
proc tabulate;
    class source amount;
    var weightgain;
    table source*amount, weightgain*(mean std n);
```

c) 2X2-1=3 d) p-value < .05, appears to be a good model

e) Since high p-value, don't reject equality of variances (i.e., probably equal variances)

f) 'amount' appears to be important for model, but source and 'interaction' do not