Math 5364
Data Mining 2
Homework 28
Mary Barker

1. The data set cows.txt contains milk production values for 300 (hypothetical) cows, 100 from the Andrews farm, 100 from the Bailey farm, and 100 from the Carter farm.

   (a) Import the data into SAS, and find the average milk production, stratified by farm. Also, obtain a histogram and qqplot of the milk production values at each farm;

   ```
   data milkdata;
       infile '/folders/myshortcuts/sas_folder/cows.txt' dlm=',';
       input milk farm $;
       IF farm = 'Andrews' THEN Afarm = milk;
       IF farm = 'Bailey' THEN Bfarm = milk;
       IF farm = 'Carter' THEN Cfarm = milk;
   proc means data=milkdata mean;
           var Afarm Bfarm Cfarm;
   run;
   ```

   (b) Perform and ANOVA to test whether the average milk production at the three farms is the same;

   ```
   proc anova data=milkdata;
           class farm;
           model milk=farm;
   run;
   ```

   (c) Test whether the average milk production is the same using PROC GLM;

   ```
   proc glm data=milkdata;
           class farm;
           model milk=farm;
   run;
   ```

2. Let $U(a, b)$ denote a uniform distribution on the interval [a, b], and $N(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2. Let $X_{i1} \sim U(0, 100)$, $X_{i2} \sim U(30, 70)$, and $\epsilon_i \sim N(0, 1) for i = 1, ..., 1000$. Also, suppose X_{i3} is a categorical variable taking the values "A", "B", and "C" with probabilities 0.5, 0.35, and 0.15 respectively. Finally, assume that all of the random variables X_{ij} and ϵ_i are statistically independent and define

$$Y_i = 150 + 8X_{i1} + 6X_{i2} + 0.25X_{i2}^2 - 7X_{i1}X_{i2} + 5I(X_{i3} = "B") + 10I(X_{i3} = "C") + \epsilon_i$$

Recall that I is the indeicator function. e.g., $I(X_{i3} =' B') = 1$ if $X_{i3} =' B'$, and 0 otherwise;

(a) Use SAS to simulate values of all random variables described above;

```
data simdata;
call streaminit(123);
do i = 1 to 1000;
        x1 = rand('UNIFORM');
        x2 = 30 + 70 * rand('UNIFORM');
        u = rand('UNIFORM');
        if u <= 0.5 then do;
                x3 = 'A';
                I_b = 0;
                I_c = 0;
        end;
        else if u <= 0.85 then do;
                x3 = 'B';
                I_b = 1;
                I_c = 0;
        end;
        else do;
                x3 = 'C';
                I_b = 0;
                I_c = 1;
        end;
        eps = rand('NORMAL');
        y = 150. + 8 * x1 + 6 * x2 + 0.25 * x2**2 - 7 * x1 * x2 +
                        5 * I_b + 10 * I_c + eps;
        output;
end;
run;
```

(b) Verify that X_{i1}, X_{i2}, and e_i have the distributions given above by plotting histograms for these variables;

```
proc univariate data=simdata;
        histogram x1;
        histogram x2;
        histogram eps/normal;
run;
```

(c) Verify that the observed frequencies of the different levels of X_{i3} are approximately equal to those stated in the problem;

```
proc freq data=simdata;
        tables x3;
run;
```

(d) Fit the given regression equation to your simulated data, and verify that the estimated coefficients agree with those stated in the problem;

```
proc glmselect data=simdata;
      class x3;
      model y = x1 x2 x2*x2 x1*x2 x3;
run;
```

The MEANS Procedure

Variable	Mean
Afarm	6999.36
Bfarm	7683.66
Cfarm	7909.71

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
farm	3	Andrews Bailey Carter

Number of Observations Read	300
Number of Observations Used	300

The ANOVA Procedure

Dependent Variable: milk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	44936638.8	22468319.4	22.84	<.0001
Error	297	292179684.2	983770.0		
Corrected Total	299	337116323.0			

R-Square	Coeff Var	Root MSE	milk Mean
0.133297	13.17041	991.8518	7530.910

Source	DF	Anova SS	Mean Square	F Value	Pr > F
farm	2	44936638.79	22468319.40	22.84	<.0001

Distribution of milk
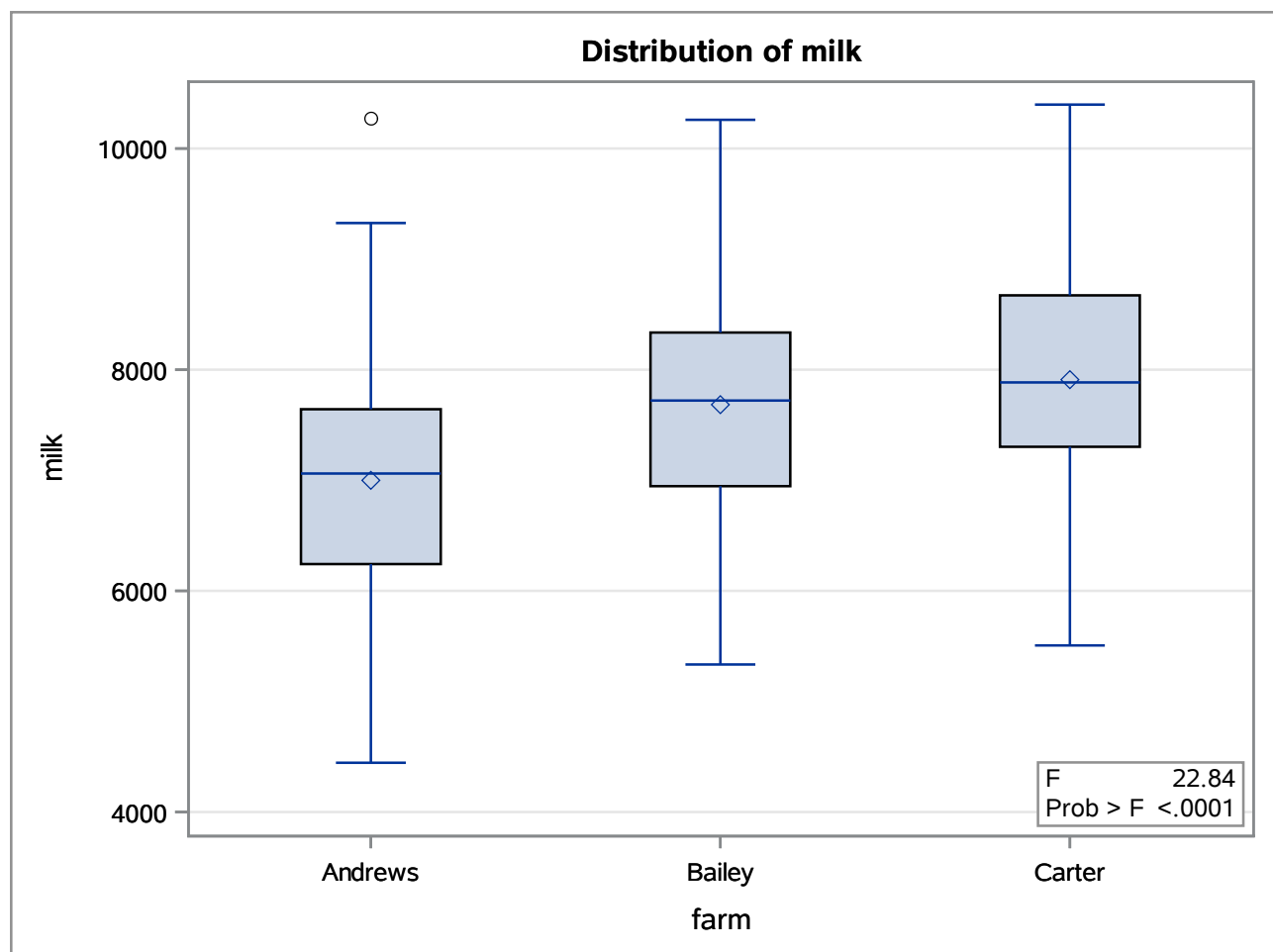
The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
farm	3	Andrews Bailey Carter

Number of Observations Read	300
Number of Observations Used	300

The GLM Procedure

Dependent Variable: milk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	44936638.8	22468319.4	22.84	<.0001
Error	297	292179684.2	983770.0		
Corrected Total	299	337116323.0			

R-Square	Coeff Var	Root MSE	milk Mean
0.133297	13.17041	991.8518	7530.910

Source	DF	Type I SS	Mean Square	F Value	Pr > F
farm	2	44936638.79	22468319.40	22.84	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
farm	2	44936638.79	22468319.40	22.84	<.0001

Distribution of milk
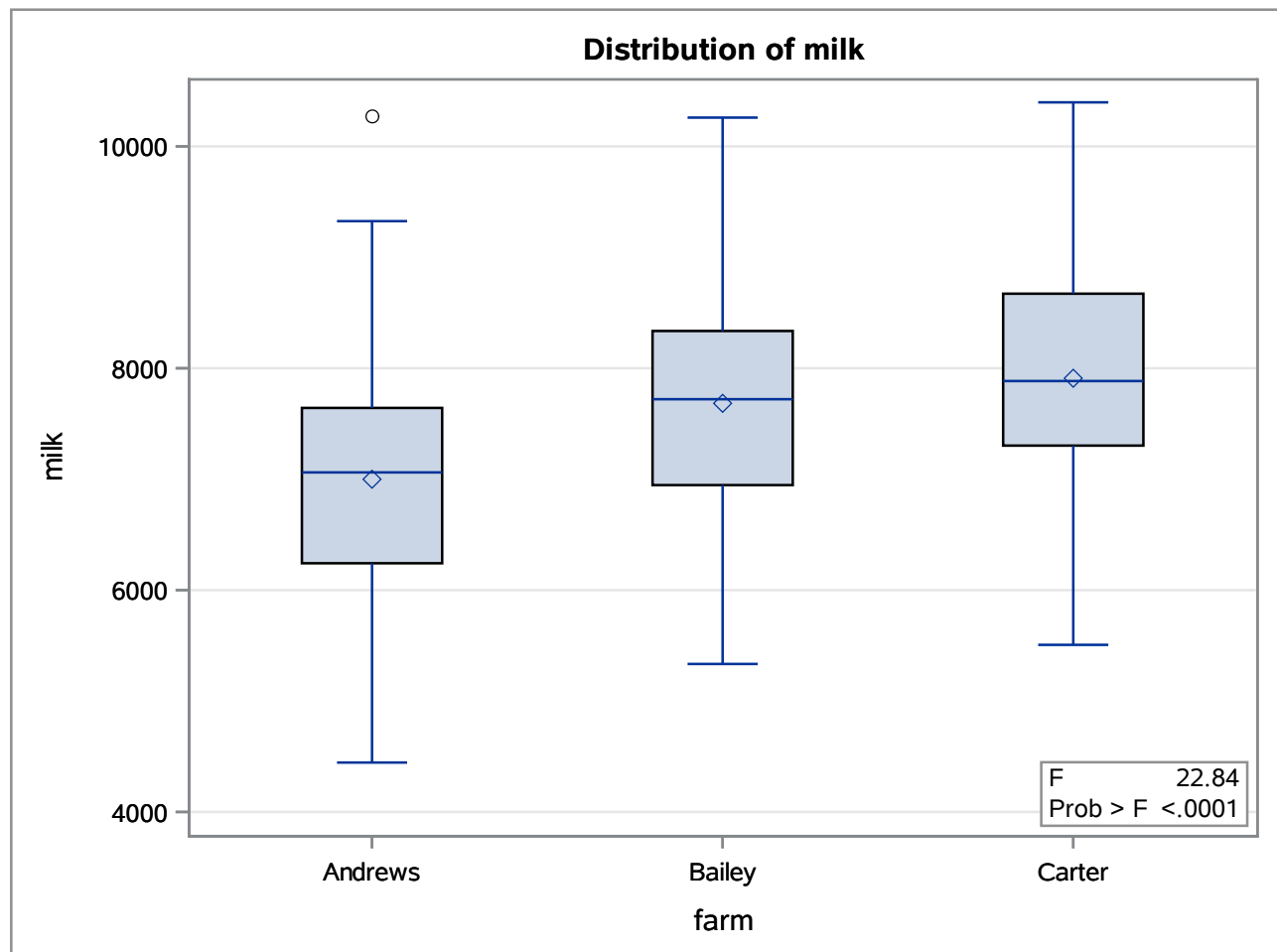
The UNIVARIATE Procedure
Variable: i

Moments			
N	1000	Sum Weights	1000
Mean	500.5	Sum Observations	500500
Std Deviation	288.819436	Variance	83416.6667
Skewness	0	Kurtosis	-1.2
Uncorrected SS	333833500	Corrected SS	83333250
Coeff Variation	57.706181	Std Error Mean	9.13327251

Basic Statistical Measures			
Location		Variability	
Mean	500.5000	Std Deviation	288.81944
Median	500.5000	Variance	83417
Mode	.	Range	999.00000
		Interquartile Range	500.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	54.79964	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1000.0
99%	990.5
95%	950.5
90%	900.5
75% Q3	750.5
50% Median	500.5
25% Q1	250.5
10%	100.5
5%	50.5
1%	10.5
0% Min	1.0

The UNIVARIATE Procedure
Variable: i

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1	1	996	996
2	2	997	997
3	3	998	998
4	4	999	999
5	5	1000	1000

The UNIVARIATE Procedure
Variable: x1

Moments			
N	1000	Sum Weights	1000
Mean	0.49771572	Sum Observations	497.715718
Std Deviation	0.29140022	Variance	0.08491409
Skewness	0.02814439	Kurtosis	-1.1867866
Uncorrected SS	332.550113	Corrected SS	84.8291766
Coeff Variation	58.5475229	Std Error Mean	0.00921488

Basic Statistical Measures			
Location		Variability	
Mean	0.497716	Std Deviation	0.29140
Median	0.491302	Variance	0.08491
Mode	.	Range	0.99898
		Interquartile Range	0.50065

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	54.01215	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.999477837
99%	0.992517911
95%	0.955291928
90%	0.908554663
75% Q3	0.755835464
50% Median	0.491302323
25% Q1	0.255183429
10%	0.086531033
5%	0.044985343
1%	0.008307807
0% Min	0.000494724

The UNIVARIATE Procedure
Variable: x1

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.000494724	828	0.996426	760
0.001378515	520	0.997883	687
0.002927703	732	0.998446	48
0.003269936	428	0.999214	158
0.006201511	86	0.999478	959

The UNIVARIATE Procedure

Distribution of x1
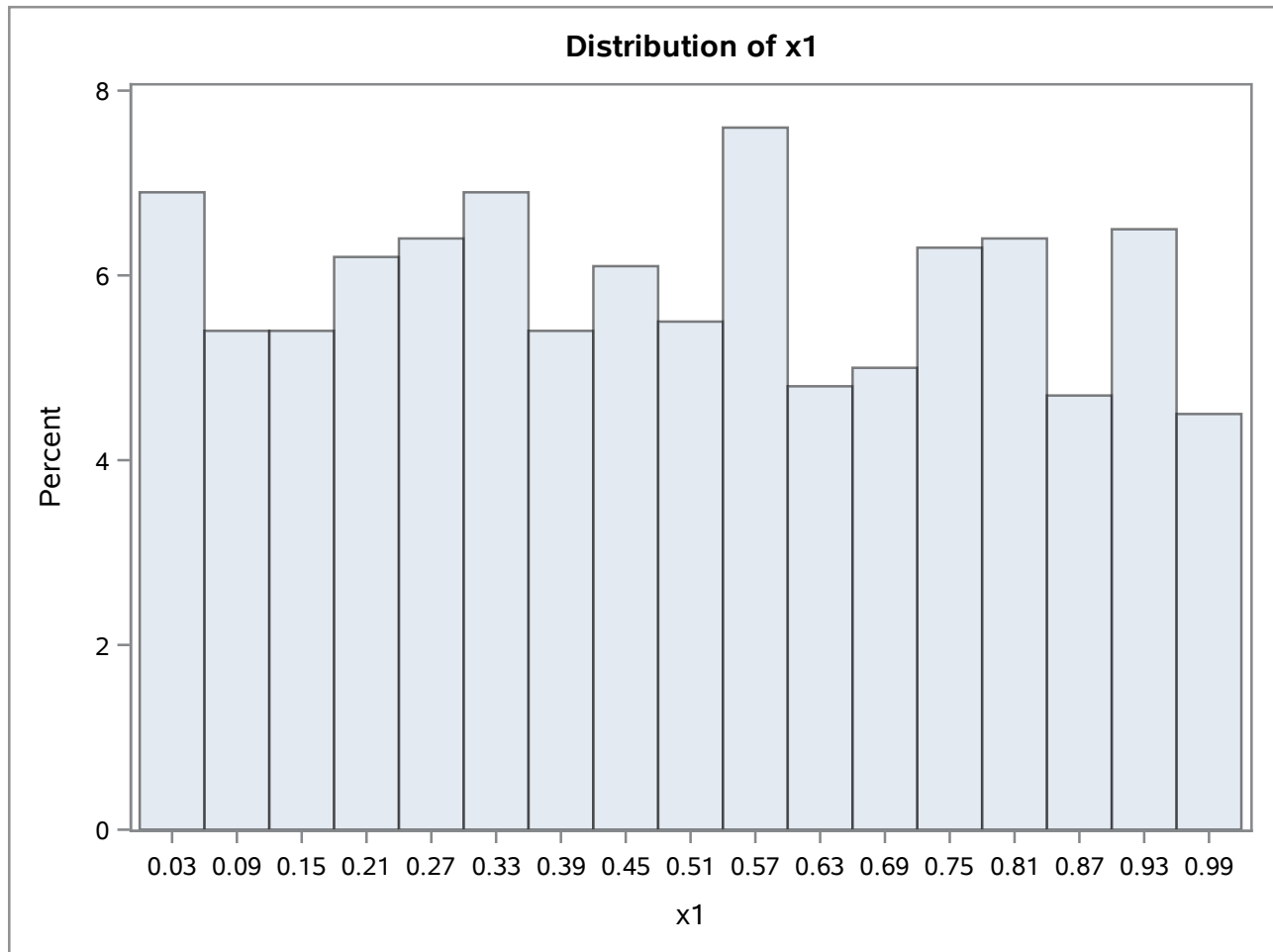
The UNIVARIATE Procedure
Variable: x2

Moments			
N	1000	Sum Weights	1000
Mean	65.5352582	Sum Observations	65535.2582
Std Deviation	20.1339301	Variance	405.375141
Skewness	-0.0542508	Kurtosis	-1.1802816
Uncorrected SS	4699839.83	Corrected SS	404969.765
Coeff Variation	30.722287	Std Error Mean	0.63669077

Basic Statistical Measures			
Location		Variability	
Mean	65.53526	Std Deviation	20.13393
Median	66.22523	Variance	405.37514
Mode	.	Range	69.89568
		Interquartile Range	34.38882

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	102.9311	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	99.9136
99%	99.2604
95%	96.8647
90%	92.7960
75% Q3	83.0793
50% Median	66.2252
25% Q1	48.6905
10%	37.2497
5%	33.3307
1%	30.8321
0% Min	30.0179

The UNIVARIATE Procedure
Variable: x2

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
30.0179	382	99.6012	215
30.1320	759	99.6436	663
30.3231	84	99.6885	198
30.5172	969	99.8700	600
30.5259	154	99.9136	704

The UNIVARIATE Procedure

Distribution of x2
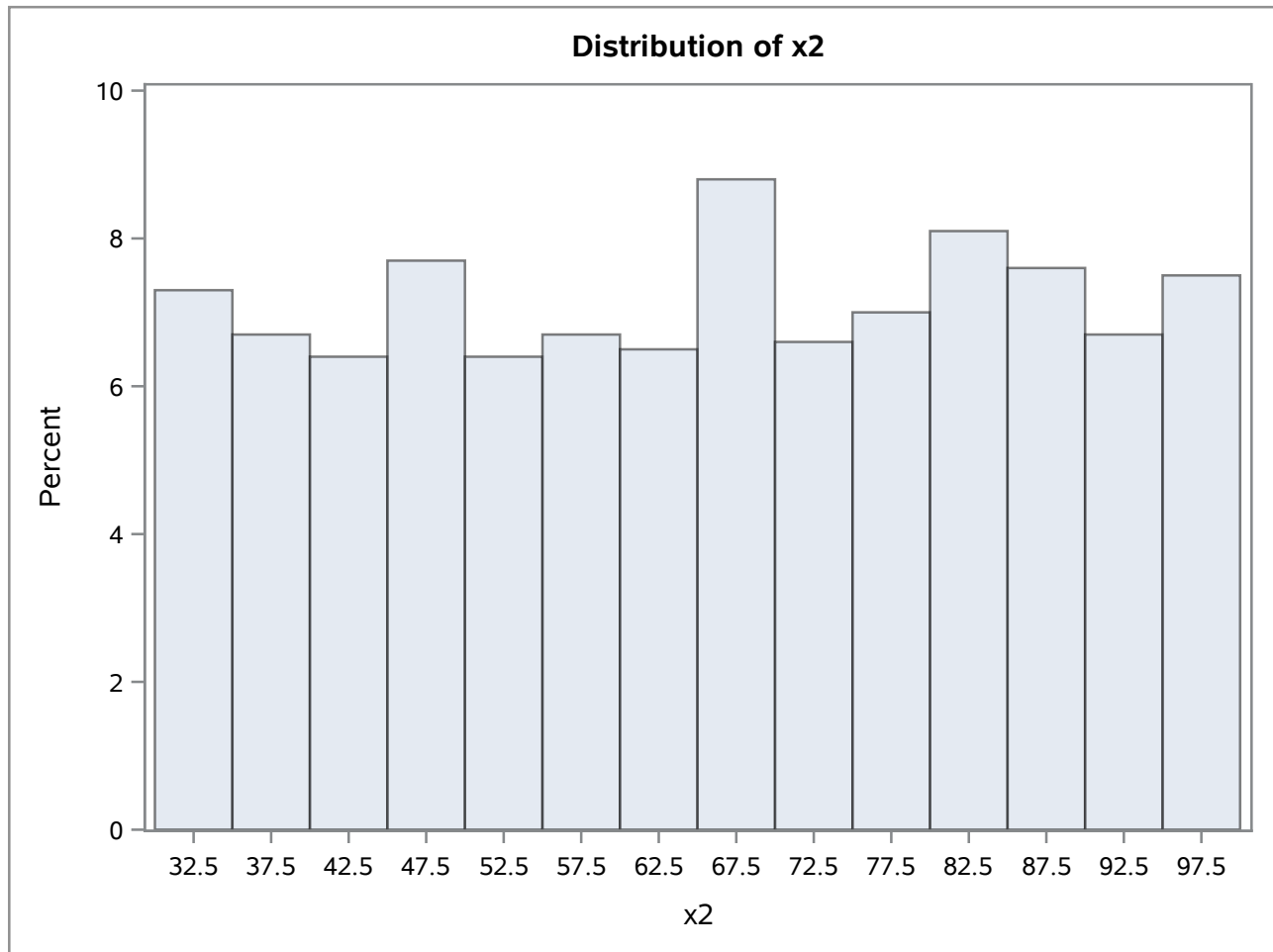
The UNIVARIATE Procedure
Variable: u

Moments			
N	1000	Sum Weights	1000
Mean	0.50601997	Sum Observations	506.019967
Std Deviation	0.28561453	Variance	0.08157566
Skewness	-0.0021395	Kurtosis	-1.2007766
Uncorrected SS	337.550292	Corrected SS	81.4940849
Coeff Variation	56.4433323	Std Error Mean	0.00903192

Basic Statistical Measures			
Location		Variability	
Mean	0.506020	Std Deviation	0.28561
Median	0.517511	Variance	0.08158
Mode	.	Range	0.99603
		Interquartile Range	0.49182

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	56.02571	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	0.99934577
99%	0.99067734
95%	0.94886888
90%	0.90434815
75% Q3	0.75370948
50% Median	0.51751077
25% Q1	0.26189177
10%	0.11289735
5%	0.06402621
1%	0.01356232
0% Min	0.00331398

The UNIVARIATE Procedure
Variable: u

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.00331398	155	0.996536	561
0.00338431	82	0.996775	811
0.00568983	710	0.996985	556
0.00661824	673	0.997022	983
0.00927767	136	0.999346	428

The UNIVARIATE Procedure
Variable: I_b

Moments			
N	1000	Sum Weights	1000
Mean	0.356	Sum Observations	356
Std Deviation	0.47905479	Variance	0.22949349
Skewness	0.60238862	Kurtosis	-1.6404128
Uncorrected SS	356	Corrected SS	229.264
Coeff Variation	134.565953	Std Error Mean	0.01514904

Basic Statistical Measures			
Location		Variability	
Mean	0.356000	Std Deviation	0.47905
Median	0.000000	Variance	0.22949
Mode	0.000000	Range	1.00000
		Interquartile Range	1.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	23.49983	Pr > |t|	<.0001
Sign	M	178	Pr >= |M|	<.0001
Signed Rank	S	31773	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1
99%	1
95%	1
90%	1
75% Q3	1
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

The UNIVARIATE Procedure
Variable: I_b

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	998	1	992
0	996	1	995
0	994	1	997
0	993	1	999
0	987	1	1000

The UNIVARIATE Procedure
Variable: I_c

Moments			
N	1000	Sum Weights	1000
Mean	0.155	Sum Observations	155
Std Deviation	0.36208577	Variance	0.13110611
Skewness	1.90944472	Kurtosis	1.64927369
Uncorrected SS	155	Corrected SS	130.975
Coeff Variation	233.603724	Std Error Mean	0.01145016

Basic Statistical Measures			
Location		Variability	
Mean	0.155000	Std Deviation	0.36209
Median	0.000000	Variance	0.13111
Mode	0.000000	Range	1.00000
		Interquartile Range	0

Tests for Location: Mu0=0					
Test	Statistic		p Value		
Student's t	t	13.53693	Pr > |t|	<.0001
Sign	M	77.5	Pr >= |M|	<.0001
Signed Rank	S	6045	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1
99%	1
95%	1
90%	1
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

The UNIVARIATE Procedure
Variable: I_c

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	1000	1	955
0	999	1	960
0	998	1	979
0	997	1	981
0	996	1	983

The UNIVARIATE Procedure
Variable: eps

Moments			
N	1000	Sum Weights	1000
Mean	0.05924359	Sum Observations	59.243589
Std Deviation	1.02486108	Variance	1.05034023
Skewness	-0.1717451	Kurtosis	0.10258259
Uncorrected SS	1052.79969	Corrected SS	1049.28989
Coeff Variation	1729.91052	Std Error Mean	0.03240895

Basic Statistical Measures			
Location		Variability	
Mean	0.059244	Std Deviation	1.02486
Median	0.043250	Variance	1.05034
Mode	.	Range	7.24362
		Interquartile Range	1.42132

Tests for Location: Mu0=0				
Test		Statistic	p Value	
Student's t	t	1.828001	Pr > |t|	0.0678
Sign	M	17	Pr >= |M|	0.2967
Signed Rank	S	19216	Pr >= |S|	0.0354

Quantiles (Definition 5)	
Level	Quantile
100% Max	3.0990968
99%	2.2041591
95%	1.7150888
90%	1.3489225
75% Q3	0.8070940
50% Median	0.0432501
25% Q1	-0.6142245
10%	-1.2260246
5%	-1.5822420
1%	-2.5658339
0% Min	-4.1445280

The UNIVARIATE Procedure
Variable: eps

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-4.14453	707	2.41116	338
-3.19043	546	2.53142	149
-2.97127	751	2.95218	375
-2.90629	371	2.96321	693
-2.81298	220	3.09910	556

The UNIVARIATE Procedure

Distribution of eps
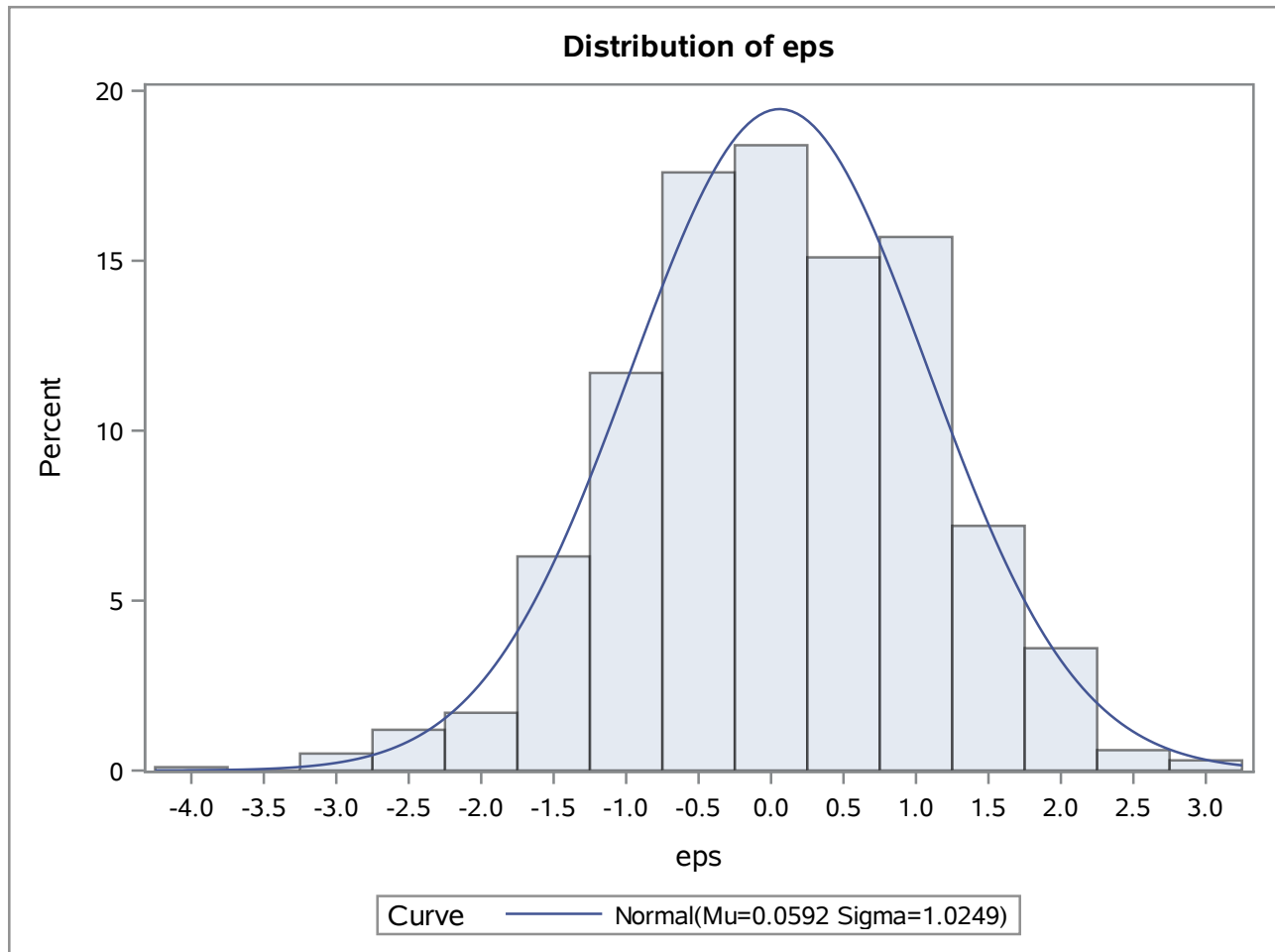
The UNIVARIATE Procedure
Fitted Normal Distribution for eps

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	0.059244
Std Dev	Sigma	1.024861

Goodness-of-Fit Tests for Normal Distribution				
Test		**Statistic**	**p Value**	
Kolmogorov-Smirnov	D	0.02530626	**Pr > D**	0.121
Cramer-von Mises	W-Sq	0.06996860	**Pr > W-Sq**	>0.250
Anderson-Darling	A-Sq	0.51795828	**Pr > A-Sq**	0.197

Quantiles for Normal Distribution		
	Quantile	
Percent	**Observed**	**Estimated**
1.0	-2.56583	-2.32494
5.0	-1.58224	-1.62650
10.0	-1.22602	-1.25417
25.0	-0.61422	-0.63201
50.0	0.04325	0.05924
75.0	0.80709	0.75050
90.0	1.34892	1.37266
95.0	1.71509	1.74499
99.0	2.20416	2.44343

The UNIVARIATE Procedure
Variable: y

Moments			
N	1000	Sum Weights	1000
Mean	1497.53375	Sum Observations	1497533.75
Std Deviation	725.12772	Variance	525810.211
Skewness	0.32608296	Kurtosis	-1.0386698
Uncorrected SS	2767891743	Corrected SS	525284400
Coeff Variation	48.4214609	Std Error Mean	22.9305519

Basic Statistical Measures			
Location		Variability	
Mean	1497.534	Std Deviation	725.12772
Median	1406.060	Variance	525810
Mode	.	Range	2771
		Interquartile Range	1221

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	65.30736	Pr > |t|	<.0001
Sign	M	500	Pr >= |M|	<.0001
Signed Rank	S	250250	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	3151.719
99%	3031.263
95%	2729.674
90%	2560.708
75% Q3	2085.413
50% Median	1406.060
25% Q1	864.897
10%	583.726
5%	516.236
1%	423.931
0% Min	380.463

The UNIVARIATE Procedure
Variable: y

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
380.463	48	3099.45	329
401.386	279	3119.22	249
404.700	759	3124.03	280
404.705	969	3130.76	372
407.615	93	3151.72	618

The FREQ Procedure

x3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	489	48.90	489	48.90
B	356	35.60	845	84.50
C	155	15.50	1000	100.00

The GLMSELECT Procedure

Data Set	WORK.SIMDATA
Dependent Variable	y
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	1000
Number of Observations Used	1000

Class Level Information		
Class	Levels	Values
x3	3	A B C

Dimensions	
Number of Effects	6
Number of Parameters	8

The GLMSELECT Procedure

	Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC
0	Intercept		1	1	13178.6029
1	x2*x2		2	2	9820.6705
2	x1*x2		3	3	5674.9457
3	x2		4	4	2752.9267
4	x3		5	6	396.2281
5	x1		6	7	80.1769*
* Optimal Value of Criterion					

Selection stopped because all effects are in the final model.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 5).

Effects:	Intercept x1 x2 x2*x2 x1*x2 x3

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	525283368	87547228	8.421E7
Error	993	1032.33434	1.03961	
Corrected Total	999	525284400		

Root MSE	1.01961
Dependent Mean	1497.53375
R-Square	1.0000
Adj R-Sq	1.0000
AIC	1047.82259
AICC	1047.96790
SBC	80.17687

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	160.634181	0.402846	398.75
x1	1	7.204582	0.370299	19.46
x2	1	5.997550	0.011841	506.53
x2*x2	1	0.249969	0.000088178	2834.83
x1*x2	1	-6.988066	0.005520	-1265.9
x3 A	1	-10.297511	0.094148	-109.38
x3 B	1	-5.322433	0.098354	-54.11
x3 C	0	0	.	.