STATISTICAL PROGRAMMING ASSIGNMENT IN (SAS)

By: Eric Brown

Table of Contents

Problem 1.)

Program Code:

```
/***************************************************************************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p1.sas
Written by: Eric Brown
Date: December 13, 2019

This program reads in data from a data file that contains info
regarding speed with which rats can negotiate a maze. The rats
are also grouped into three age groups (3,6, and 9) and two
genetic strains (A and B). Then the program conducts a two-way analysis
of variance with age and strain as the predictor and speed as the response
variable. Then the program creates an interaction plot with the average
speed and age.

Input: ratmaze.dat - data file
Output: two-way anova analysis and interaction plot
produced via PROC GLM, PPROC MEANS, & PROC GPLOT
***************************************************************************************************/
DATA ratmaze; *temporary dataset containing rat maze data;
  infile 'C:\Users\esbro\Desktop\STAT 482\ratmaze.dat';
  input age strain $ speed @@;
RUN;
OPTIONS ps=94 ls=98 nodate nonumber;
PROC PRINT data=ratmaze; *making sure dataset stored and displayed correctly since design is unbalanced;
run;

PROC GLM data=ratmaze; *can not use proc anova since design is not balanced;
  TITLE 'Two-way Analysis of Variance - Unbalanced Design';
  CLASS age strain;
  MODEL speed= age | strain / ss3; *producing only type III sum of squares;
  LSMEANS age | strain / PDIFF ADJUST=TUKEY; *produce least-square, adjusted means for main effects;
   *computes probabilities for all pairwise differences and adjustment for multiple comparisons;
RUN;

*restrict the output data set while getting cell means;
PROC MEANS data=ratmaze NWAY NOPRINT;
  CLASS age strain;
  VAR speed;
  OUTPUT OUT=MEANS MEAN=average_speed;
RUN;
TITLE;

SYMBOL1 V=SQUARE COLOR=BLUE I=JOIN;
SYMBOL2 V=CIRCLE COLOR=BLACK I=JOIN;
PROC GPLOT DATA=MEANS;
  TITLE 'Interaction Plot';
  PLOT average_speed * age = strain; *age is x-axis variable & average_speed y-axis variable;
RUN;
TITLE;
```

Problem 1.) SAS Log Window:

```
1     /*************************************************************************************
1   ! **************************************
2     Filename: C:\Users\Eric\Desktop\STAT 482\final_p1.sas
3     Written by: Eric Brown
4     Date: December 13, 2019
5
6     This program reads in data from a data file that contains info
7     regarding speed with which rats can negotiate a maze. The rats
8     are also grouped into three age groups (3,6, and 9) and two
9     genetic strains (A and B). Then the program conducts a two-way analysis
10    of variance with age and strain as the predictor and speed as the response
11    variable. Then the program creates an interaction plot with the average
12    speed and age.
13
14    Input: ratmaze.dat - data file
15    Output: two-way anova analysis and interaction plot
16    produced via PROC GLM, PPROC MEANS, & PROC GPLOT
17    *************************************************************************************
17  ! **************************************/
18    DATA ratmaze; *temporary dataset containing rat maze data;
19      infile 'C:\Users\esbro\Desktop\STAT 482\ratmaze.dat';
20      input age strain $ speed @@;
21    RUN;

NOTE: The infile 'C:\Users\esbro\Desktop\STAT 482\ratmaze.dat' is:
      Filename=C:\Users\esbro\Desktop\STAT 482\ratmaze.dat,
      RECFM=V,LRECL=32767,File Size (bytes)=262,
      Last Modified=12Dec2019:20:15:54,
      Create Time=12Dec2019:20:15:53

NOTE: 6 records were read from the infile 'C:\Users\esbro\Desktop\STAT 482\ratmaze.dat'.
      The minimum record length was 27.
      The maximum record length was 59.
NOTE: SAS went to a new line when INPUT statement reached past the end of a line.
NOTE: The data set WORK.RATMAZE has 35 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds


22    OPTIONS ps=94 ls=98 nodate nonumber;
23    PROC PRINT data=ratmaze; *making sure dataset stored and displayed correctly since design is
23  ! unbalanced;
24    run;

NOTE: There were 35 observations read from the data set WORK.RATMAZE.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds


25
26    PROC GLM data=ratmaze; *can not use proc anova since design is not balanced;
27      TITLE 'Two-way Analysis of Variance - Unbalanced Design';
28      CLASS age strain;
29      MODEL speed= age | strain / ss3; *producing only type III sum of squares;
30      LSMEANS age | strain / PDIFF ADJUST=TUKEY; *produce least-square, adjusted means for main
30  ! effects;
31        *computes probabilities for all pairwise differences and adjustment for multiple
31  ! comparisons;
32    RUN;

33
34    *restrict the output data set while getting cell means;

NOTE: PROCEDURE GLM used (Total process time):
      real time           0.07 seconds
      cpu time            0.03 seconds


35    PROC MEANS data=ratmaze NWAY NOPRINT;
36      CLASS age strain;
37      VAR speed;
38      OUTPUT OUT=MEANS MEAN=average_speed;
39    RUN;

NOTE: There were 35 observations read from the data set WORK.RATMAZE.
NOTE: The data set WORK.MEANS has 6 observations and 5 variables.
NOTE: PROCEDURE MEANS used (Total process time):
      real time           0.04 seconds
      cpu time            0.00 seconds


40    TITLE;
41
42    SYMBOL1 V=SQUARE COLOR=BLUE I=JOIN;
43    SYMBOL2 V=CIRCLE COLOR=BLACK I=JOIN;
44    PROC GPLOT DATA=MEANS;
45      TITLE 'Interaction Plot';
46      PLOT average_speed * age = strain; *age is x-axis variable & average_speed y-axis variable;
47    RUN;
```

Problem 1.) Program Output:

The SAS System

Obs	age	strain	speed
1	3	A	12
2	3	A	14
3	3	A	9
4	3	A	17
5	3	A	10
6	3	A	11
7	3	A	9
8	3	A	10
9	3	B	24
10	3	B	17
11	3	B	22
12	3	B	16
13	3	B	18
14	6	A	22
15	6	A	20
16	6	A	12
17	6	A	12
18	6	A	17
19	6	A	14
20	6	A	17
21	6	B	23
22	6	B	26
23	6	B	34
24	6	B	20
25	9	A	14
26	9	A	14
27	9	A	10
28	9	A	15
29	9	A	17
30	9	A	12
31	9	A	19
32	9	B	27
33	9	B	29
34	9	B	27
35	9	B	23

Two-way Analysis of Variance - Unbalanced Design

The GLM Procedure

Class Level Information

Class	Levels	Values
age	3	3 6 9
strain	2	A B

Number of Observations Read 35
Number of Observations Used 35

Two-way Analysis of Variance - Unbalanced Design

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

age	speed LSMEAN	LSMEAN Number
3	15.4500000	1
6	21.0178571	2
9	20.4642857	3

Least Squares Means for effect age
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: speed

i/j	1	2	3
1		0.0027	0.0070
2	0.0027		0.9353
3	0.0070	0.9353	

Two-way Analysis of Variance - Unbalanced Design

The GLM Procedure

Dependent Variable: speed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	978.650000	195.730000	15.25	<.0001
Error	29	372.092857	12.830788		
Corrected Total	34	1350.742857			

R-Square	Coeff Var	Root MSE	speed Mean
0.724527	20.45193	3.582009	17.51429

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	2	216.3195806	108.1597903	8.43	0.0013
strain	1	780.0937529	780.0937529	60.80	<.0001
age*strain	2	24.4241883	12.2120942	0.95	0.3978

Problem 1.) Program Output – Continued:

Two-way Analysis of Variance - Unbalanced Design

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

strain	speed LSMEAN	H0:LSMean1= LSMean2 Pr > |t|
A	14.0714286	<.0001
B	23.8833333	

Two-way Analysis of Variance - Unbalanced Design

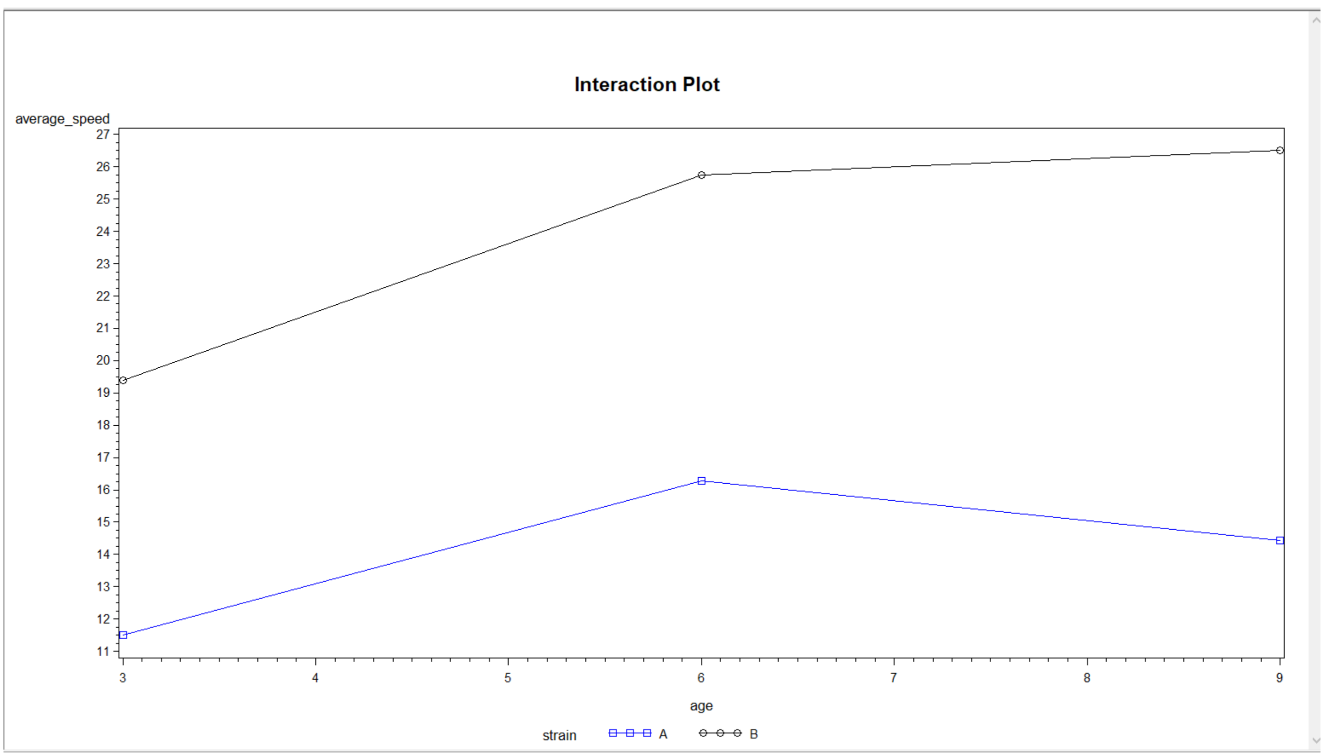
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

age	strain	speed LSMEAN	LSMEAN Number
3	A	11.5000000	1
3	B	19.4000000	2
6	A	16.2857143	3
6	B	25.7500000	4
9	A	14.4285714	5
9	B	26.5000000	6

Least Squares Means for effect age*strain
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: speed

i/j	1	2	3	4	5	6
1		0.0068	0.1343	<.0001	0.6177	<.0001
2	0.0068		0.6762	0.1189	0.1999	0.0616
3	0.1343	0.6762		0.0028	0.9237	0.0011
4	<.0001	0.1189	0.0028		0.0003	0.9997
5	0.6177	0.1999	0.9237	0.0003		0.0001
6	<.0001	0.0616	0.0011	0.9997	0.0001	



Explanation: Age was significant in this model (p=.0013). And strain was significant in this model (p<.0001). However, the interaction between age and strain was not significant (p=.3978 not less than 0.05). The interaction plot even confirms this statement and shows that there is no interaction between age and strain. Mainly, since the lines of both strain types never cross each other in the graph plot. The graph does show that rats with a genetic strain of B have a higher average speed through the maze than rats with a genetic strain of A.

Problem 2.)

Program Code:

```
/*****************************************************************************************************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p2.sas
Written by: Eric Brown
Date: December 13, 2019

This program gathers data regarding northern flicker birds. Pertaining
to their tail feathers. Some of the birds had one odd feather that
was different in length and/or color from the rest of their tail feathers.
So this program compares the yellowness of one typical feather against
the one odd feather of the same bird. Then concludes whether the
mean yellowness of the odd feather differs from the typical feather.

Input: birds.dat - data file
Output: difference of means analysis
produced via PROC TTEST
*****************************************************************************************************************************/

DATA birds; *gathering in data from input data file;
  infile 'C:\Users\esbro\Desktop\STAT 482\birds.dat';
  input birdLetter $ birdType $ featherLength;
RUN;

DATA analysis;
  SET birds;
    by birdLetter; *have to sort birds by bird type for paired ttest;
    *storing the feather length of each type of bird;
      if birdType= 'Typical' then typical_Len=featherLength;
      else if birdType= 'Odd' then odd_Len=featherLength;
    if last.birdLetter then output;
    retain typical_Len odd_Len;
    drop birdType featherLength; *no longer need orignal feather data;
RUN;
OPTIONS ls=98 ps=95 nodate nonumber;
PROC PRINT data=analysis; *making sure data is displayed correctly;

PROC TTEST data=analysis;
  TITLE 'Paired T-test of of Northern Flicker''s Feathers';
  PAIRED typical_Len * odd_Len; *comparing typical and odd feather lengths;
RUN;
TITLE;

```

Problem 2.) SAS Log Window:

```
1     /*****************************************************************************************
1   ! ***************************************
2     Filename: C:\Users\Eric\Desktop\STAT 482\final_p2.sas
3     Written by: Eric Brown
4     Date: December 13, 2019
5
6     This program gathers data regarding northern flicker birds. Pertaining
7     to their tail feathers. Some of the birds had one odd feather that
8     was different in length and/or color from the rest of their tail feathers.
9     So this program compares the yellowness of one typical feather against
10    the one odd feather of the same bird. Then concludes whether the
11    mean yellowness of the odd feather differs from the typical feather.
12
13    Input: birds.dat - data file
14    Output: difference of means analysis
15    produced via PROC TTEST
16    *****************************************************************************************
16  ! ***************************************/
17
18    DATA birds; *gathering in data from input data file;
19      infile 'C:\Users\esbro\Desktop\STAT 482\birds.dat';
20      input birdLetter $ birdType $ featherLength;
21    RUN;

NOTE: The infile 'C:\Users\esbro\Desktop\STAT 482\birds.dat' is:
      Filename=C:\Users\esbro\Desktop\STAT 482\birds.dat,
      RECFM=V,LRECL=32767,File Size (bytes)=734,
      Last Modified=13Dec2019:00:01:13,
      Create Time=13Dec2019:00:01:12

NOTE: 32 records were read from the infile 'C:\Users\esbro\Desktop\STAT 482\birds.dat'.
      The minimum record length was 21.
      The maximum record length was 21.
NOTE: The data set WORK.BIRDS has 32 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      cpu time            0.01 seconds


22
23    DATA analysis;
24      SET birds;
25        by birdLetter; *have to sort birds by bird type for paired ttest;
26        *storing the feather length of each type of bird;
27          if birdType= 'Typical' then typical_Len=featherLength;
28          else if birdType= 'Odd' then odd_Len=featherLength;
29        if last.birdLetter then output;
30        retain typical_Len odd_Len;
31        drop birdType featherLength; *no longer need orignal feather data;
32    RUN;

NOTE: There were 32 observations read from the data set WORK.BIRDS.
NOTE: The data set WORK.ANALYSIS has 16 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds


33    OPTIONS ls=98 ps=95 nodate nonumber;
34    PROC PRINT data=analysis; *making sure data is displayed correctly;
35    RUN;

NOTE: There were 16 observations read from the data set WORK.ANALYSIS.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds


36
37    PROC TTEST data=analysis;
38      TITLE 'Paired T-test of of Northern Flicker''s Feathers';
39      PAIRED typical_Len * odd_Len; *comparing typical and odd feather lengths;
40    RUN;

NOTE: PROCEDURE TTEST used (Total process time):
      real time           0.03 seconds
      cpu time            0.00 seconds


41    TITLE;
```

Problem 2.) Program Output:

Obs	bird Letter	typical_ Len	odd_Len
1	A	-0.255	-0.324
2	B	-0.213	-0.185
3	C	-0.190	-0.299
4	D	-0.185	-0.144
5	E	-0.045	-0.027
6	F	-0.025	-0.039
7	G	-0.015	-0.264
8	H	0.003	-0.077
9	I	0.015	-0.017
10	J	0.020	-0.169
11	K	0.023	-0.096
12	L	0.040	-0.330
13	M	0.040	-0.346
14	N	0.050	-0.191
15	O	0.055	-0.128
16	P	0.058	-0.182

Paired T-test of of Northern Flicker's Feathers

The TTEST Procedure

Difference: typical_Len - odd_Len

N	Mean	Std Dev	Std Err	Minimum	Maximum
16	0.1371	0.1349	0.0337	-0.0410	0.3860

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.1371	0.0652	0.2090	0.1349	0.0997	0.2089

DF	t Value	Pr > |t|
15	4.06	0.0010

Explanation: In this problem, the mean difference (typical_Len - odd_Len) is positive (tail length increased) and equal to 0.1371. Then, the probability of the difference happened by chance was 0.0010. As a result, we can conclude that the mean yellowness of the odd and typical feathers differs. Since the probability is statistically significant at the .05 significance level (p=0.0010< .05).

Problem 3.)

Program Code:

```
/*******************************************************************************************************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p3.sas
Written by: Eric Brown
Date: December 15, 2019

This program determines what the kappa coefficient is when the cutoff values
are set at 0.4, 0.5, and 0.6. Then it suggests what level of agreement exists
between the two rates. Then the program does it all over again for cutoff
values of 0.2, 0.5, 0.8.

Input: created own data set based from problem description
Output: kappa coefficient caculation
produced via PROC FREQ
*******************************************************************************************************************************/

%Let cutoff1=0.4;
%Let cutoff2=0.5;
%Let cutoff3=0.6;
DATA agree;
  y=RANUNI(456);

  Do subj= 1 to 100; *100 observations
  *using seed of 456 in RANUNI function;
    If RANUNI(456) lt &cutoff1 then do;
    *two character variables to calculate the Kappa coef. between them;
      rater1='Yes';
      rater2='Yes';
    End;
    *second cutoff value least 0.4, but less than 0.5, ;
    Else if RANUNI(456) ge &cutoff1 and RANUNI(456) lt &cutoff2 then do;
      rater1='Yes';
      rater2='No';
    End;
    * least 0.5, but less than 0.6;
    Else if RANUNI(456) ge &cutoff2 and RANUNI(456) lt &cutoff3 then do;
      rater1='No';
      rater2='Yes';
    End;
    *greater than 0.6;
    Else if RANUNI(456) ge &cutoff3  then do;
      rater1='No';
      rater2='No';
    End;
    Output;
  End;
RUN;

PROC PRINT data=agree;
Run;

PROC FREQ data=agree;
  TITLE 'Computing Coefficient Kappa for Two Raters';
  Tables rater1 * rater2 / AGREE; *computing kappa coef.;
RUN;
```

Problem 3.) Program Code – Continued:

```
/**************************************************************************************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p3.sas
Written by: Eric Brown
Date: December 15, 2019

This program determines what the kappa coefficient is when the cutoff values
are set at 0.4, 0.5, and 0.6. Then it suggests what level of agreement exists
between the two rates. Then the program does it all over again for cutoff
values of 0.2, 0.5, 0.8.

Input: created own data set based from problem description
Output: kappa coefficient caculation
produced via PROC FREQ
**************************************************************************************************************/

%Let cutoff1=0.2;
%Let cutoff2=0.5;
%Let cutoff3=0.8;
DATA agree;
  y=RANUNI(456);

  Do subj= 1 to 100; *100 observations
  *using seed of 456 in RANUNI function;
    If RANUNI(456) lt &cutoff1 then do;
    *two character variables to calculate the Kappa coef. between them;
      rater1='Yes';
      rater2='Yes';
    End;
    *second cutoff value least 0.4, but less than 0.5, ;
    Else if RANUNI(456) ge &cutoff1 and RANUNI(456) lt &cutoff2 then do;
      rater1='Yes';
      rater2='No';
    End;
    * least 0.5, but less than 0.6;
    Else if RANUNI(456) ge &cutoff2 and RANUNI(456) lt &cutoff3 then do;
      rater1='No';
      rater2='Yes';
    End;
    *greater than 0.6;
    Else if RANUNI(456) ge &cutoff3  then do;
      rater1='No';
      rater2='No';
    End;
    Output;
  End;
RUN;

PROC PRINT data=agree;
Run;

PROC FREQ data=agree;
  TITLE 'Computing Coefficient Kappa for Two Raters';
  Tables rater1 * rater2 / AGREE; *computing kappa coef.;
RUN;


*%createDATA(cutoff1=0.2 , cutoff2=0.5 , cutoff3=0.8);
```

Problem 3.) SAS Log Window:

```
/*************************************************************************************
! **************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p3.sas
Written by: Eric Brown
Date: December 15, 2019

This program determines what the kappa coefficient is when the cutoff values
are set at 0.4, 0.5, and 0.6. Then it suggests what level of agreement exists
between the two rates. Then the program does it all over again for cutoff
values of 0.2, 0.5, 0.8.

Input: created own data set based from problem description
Output: kappa coefficient caculation
produced via PROC FREQ
*************************************************************************************
! **************************************/

%Let cutoff1=0.4;
%Let cutoff2=0.5;
%Let cutoff3=0.6;
DATA agree;
  y=RANUNI(456);

  Do subj= 1 to 100; *100 observations
  *using seed of 456 in RANUNI function;
    If RANUNI(456) lt &cutoff1 then do;
    *two character variables to calculate the Kappa coef. between them;
      rater1='Yes';
      rater2='Yes';
    End;
    *second cutoff value least 0.4, but less than 0.5, ;
    Else if RANUNI(456) ge &cutoff1 and RANUNI(456) lt &cutoff2 then do;
      rater1='Yes';
      rater2='No';
    End;
    * least 0.5, but less than 0.6;
    Else if RANUNI(456) ge &cutoff2 and RANUNI(456) lt &cutoff3 then do;
      rater1='No';
      rater2='Yes';
    End;
    *greater than 0.6;
    Else if RANUNI(456) ge &cutoff3  then do;
      rater1='No';
      rater2='No';
    End;
    Output;
  End;
RUN;

NOTE: The data set WORK.AGREE has 100 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time           0.04 seconds
      cpu time            0.03 seconds


PROC PRINT data=agree;
Run;

NOTE: There were 100 observations read from the data set WORK.AGREE.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.03 seconds
      cpu time            0.01 seconds


PROC FREQ data=agree;
  TITLE 'Computing Coefficient Kappa for Two Raters';
  Tables rater1 * rater2 / AGREE; *computing kappa coef.;
RUN;

NOTE: There were 100 observations read from the data set WORK.AGREE.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds
```

Problem 3.) SAS Log Window – Continued:

```
/**************************************************************************************
****************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p3.sas
Written by: Eric Brown
Date: December 15, 2019

This program determines what the kappa coefficient is when the cutoff values
are set at 0.4, 0.5, and 0.6. Then it suggests what level of agreement exists
between the two rates. Then the program does it all over again for cutoff
values of 0.2, 0.5, 0.8.

Input: created own data set based from problem description
Output: kappa coefficient caculation
produced via PROC FREQ
**************************************************************************************
***************************************/

%Let cutoff1=0.2;
%Let cutoff2=0.5;
%Let cutoff3=0.8;
DATA agree;
  y=RANUNI(456);

  Do subj= 1 to 100; *100 observations
  *using seed of 456 in RANUNI function;
    If RANUNI(456) lt &cutoff1 then do;
    *two character variables to calculate the Kappa coef. between them;
      rater1='Yes';
      rater2='Yes';
    End;
    *second cutoff value least 0.4, but less than 0.5, ;
    Else if RANUNI(456) ge &cutoff1 and RANUNI(456) lt &cutoff2 then do;
      rater1='Yes';
      rater2='No';
    End;
    * least 0.5, but less than 0.6;
    Else if RANUNI(456) ge &cutoff2 and RANUNI(456) lt &cutoff3 then do;
      rater1='No';
      rater2='Yes';
    End;
    *greater than 0.6;
    Else if RANUNI(456) ge &cutoff3  then do;
      rater1='No';
      rater2='No';
    End;
    Output;
  End;
RUN;

NOTE: The data set WORK.AGREE has 100 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds


PROC PRINT data=agree;
Run;

NOTE: There were 100 observations read from the data set WORK.AGREE.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


PROC FREQ data=agree;
  TITLE 'Computing Coefficient Kappa for Two Raters';
  Tables rater1 * rater2 / AGREE; *computing kappa coef.;
RUN;

NOTE: There were 100 observations read from the data set WORK.AGREE.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

Problem 3.) Program Output:

The FREQ Procedure

Table of rater1 by rater2

rater1 / rater2 (Frequency, Percent, Row Pct, Col Pct)	No	Yes	Total
No	15 15.00 51.72 42.86	14 14.00 48.28 21.54	29 29.00
Yes	20 20.00 28.17 57.14	51 51.00 71.83 78.46	71 71.00
Total	35 35.00	65 65.00	100 100.00

Statistics for Table of rater1 by rater2

McNemar's Test

Chi-Square	DF	Pr > ChiSq
1.0588	1	0.3035

Simple Kappa Coefficient

Estimate	Standard Error	95% Confidence Limits	
0.2220	0.1011	0.0239	0.4200

Sample Size = 100

The FREQ Procedure

Table of rater1 by rater2

rater1 / rater2 (Frequency, Percent, Row Pct, Col Pct)	No	Yes	Total
No	6 6.00 20.00 13.95	24 24.00 80.00 42.11	30 30.00
Yes	37 37.00 52.86 86.05	33 33.00 47.14 57.89	70 70.00
Total	43 43.00	57 57.00	100 100.00

Statistics for Table of rater1 by rater2

McNemar's Test

Chi-Square	DF	Pr > ChiSq
2.7705	1	0.0960

Simple Kappa Coefficient

Estimate	Standard Error	95% Confidence Limits	
-0.2924	0.0862	-0.4612	-0.1235

Sample Size = 100

Explanation: The interpretation of kappa coefficients according to J.L. Fleiss (1981) are that values which exceed .75 have a strong agreement above chance, values in the range of .40 to .75 indicate fair levels of agreement above chance, and values below .40 indicate poor levels of agreement above chance. In this problem, we have a kappa coefficient of 0.2220, which indicates a poor level of agreement between the two raters when the cutoff values are set at 0.4, 0.5, and 0.6. However, when the cutoff values are set at 0.2, 0.5, and 0.8, the kappa coefficient we receive is -0.2924. This kappa coefficient indicates not only a poor level of agreement, but a substantial level of disagreement between the two raters.

Problem 4.)

Program Code:

```
/***********************************************************************************************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p4.sas
Written by: Eric Brown
Date: December 14, 2019

This program gathers data regarding Donner Party emigrants survival
rate when traveling through the Sierra Nevada during the 1840s. The
input data file contains information about the age and gender of the
emigrants as well. Once the data is read into the program, it uses the
logistic procedure to predict the survival of a Party member based on their
age and gender.

Input: donner.dat - data file
Output: Logistic Regression analysis
produced via PROC LOGISTIC
***********************************************************************************************************************/

PROC FORMAT; *formatting data to show if the person survived or not and their gender;
  VALUE survival_fmt 0='No'
                     1='Yes';
  VALUE gender_fmt 0='Male'
                   1='Female';
RUN;

DATA donner; *reading in data after the column heading from the data file;
  infile 'C:\Users\esbro\Desktop\STAT 482\donner.dat' firstobs=2;
  input survival age gender;
  format survival survival_fmt. gender gender_fmt.;
RUN;

OPTIONS ls=94 ps=90 nodate nonumber;
PROC PRINT data=donner; *making sure data is displayed properly;

PROC LOGISTIC data=donner DESCENDING;
  title 'Predicting Odds of Survival Using Logistic Regression';
  *creating a dummy variable for gender using male sex as the reference level;
  class gender (PARAM=REF REF='Male');
  model survival= age gender; *predicting survival based on age and gender;
RUN;
title;
QUIT;
```

Problem 4.) SAS Log Window:

```
/************************************************************************************
*************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p4.sas
Written by: Eric Brown
Date: December 14, 2019

This program gathers data regarding Donner Party emigrants survival
rate when traveling through the Sierra Nevada during the 1840s. The
input data file contains information about the age and gender of the
emigrants as well. Once the data is read into the program, it uses the
logistic procedure to predict the survival of a Party member based on their
age and gender.

Input: donner.dat - data file
Output: Logistic Regression analysis
produced via PROC LOGISTIC
************************************************************************************
*************************************/

PROC FORMAT;
             *formatting data to show if the person survived or not and their gender;
  VALUE survival_fmt 0='No'
                     1='Yes';
NOTE: Format SURVIVAL_FMT has been output.
  VALUE gender_fmt 0='Male'
                   1='Female';
NOTE: Format GENDER_FMT has been output.
RUN;

NOTE: PROCEDURE FORMAT used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds


DATA donner; *reading in data after the column heading from the data file;
  infile 'C:\Users\esbro\Desktop\STAT 482\donner.dat' firstobs=2;
  input survival age gender;
  format survival survival_fmt. gender gender_fmt.;
RUN;

NOTE: The infile 'C:\Users\esbro\Desktop\STAT 482\donner.dat' is:
      Filename=C:\Users\esbro\Desktop\STAT 482\donner.dat,
      RECFM=V,LRECL=32767,File Size (bytes)=1148,
      Last Modified=13Dec2019:01:51:54,
      Create Time=13Dec2019:01:51:54

NOTE: 45 records were read from the infile 'C:\Users\esbro\Desktop\STAT 482\donner.dat'.
      The minimum record length was 23.
      The maximum record length was 23.
NOTE: The data set WORK.DONNER has 45 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds


OPTIONS ls=94 ps=90 nodate nonumber;
PROC PRINT data=donner; *making sure data is displayed properly;
RUN;

NOTE: There were 45 observations read from the data set WORK.DONNER.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds


PROC LOGISTIC data=donner DESCENDING;
  title 'Predicting Odds of Survival Using Logistic Regression';
  *creating a dummy variable for gender using male sex as the reference level;
  class gender (PARAM=REF REF='Male');
  model survival= age gender; *predicting survival based on age and gender;
RUN;

NOTE: PROC LOGISTIC is modeling the probability that survival='Yes'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 45 observations read from the data set WORK.DONNER.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.09 seconds
      cpu time            0.04 seconds


title;
QUIT;
```

Problem 4.) Program Output:

Obs	survival	age	gender
1	No	23	Male
2	Yes	40	Female
3	Yes	40	Male
4	No	30	Male
5	No	28	Male
6	No	40	Male
7	No	45	Female
8	No	62	Male
9	No	65	Male
10	No	45	Female
11	No	25	Female
12	Yes	28	Male
13	No	28	Male
14	No	23	Male
15	Yes	22	Female
16	Yes	23	Female
17	Yes	28	Male
18	Yes	15	Female
19	No	47	Female
20	No	57	Male
21	Yes	20	Female
22	Yes	18	Male
23	No	25	Male
24	No	60	Male
25	Yes	25	Male
26	Yes	20	Male
27	Yes	32	Male
28	Yes	32	Female
29	Yes	24	Female
30	Yes	30	Male
31	No	15	Male
32	No	50	Female
33	Yes	21	Female
34	No	25	Male
35	Yes	46	Male
36	Yes	32	Female
37	No	30	Male
38	No	25	Male
39	No	25	Male
40	No	25	Male
41	No	30	Male
42	No	35	Male
43	Yes	23	Male
44	No	24	Male
45	Yes	25	Female

Predicting Odds of Survival Using Logistic Regression

The LOGISTIC Procedure

Model Information

Data Set	WORK.DONNER
Response Variable	survival
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	45
Number of Observations Used	45

Response Profile

Ordered Value	survival	Total Frequency
1	Yes	20
2	No	25

Probability modeled is survival='Yes'.

Class Level Information

Class	Value	Design Variables
gender	Female	1
	Male	0

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	63.827	57.256
SC	65.633	62.676
-2 Log L	61.827	51.256

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.5703	2	0.0051
Score	9.0965	2	0.0106
Wald	6.8627	2	0.0323

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	1	4.3988	0.0360
gender	1	4.4699	0.0345

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.6331	1.1102	2.1637	0.1413
age		1	-0.0782	0.0373	4.3988	0.0360
gender	Female	1	1.5973	0.7555	4.4699	0.0345

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	0.925	0.860	0.995
gender Female vs Male	4.940	1.124	21.716

Problem 4.) Program Output – Continued:

Predicting Odds of Survival Using Logistic Regression

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Percent Concordant	73.0	Somers' D	0.492
Percent Discordant	23.8	Gamma	0.508
Percent Tied	3.2	Tau-a	0.248
Pairs	500	c	0.746

The odds ratio for gender is 4.940 and that was for female versus male, since the males were the reference level. So in this model, the point estimate for being a female was 4.940 and the 95% CI goes from 1.124 to 21.716 and consequently not significant.

Problem 5.)

Program Code:

```
/***************************************************************************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p5.sas
Written by: Eric Brown
Date: December 12, 2019

This program uses data stored in the permanent dataset called qul. To if
a patient's qul score changes depending on how many office visits they
attend through the study. Then the program tests if there is an association
between years (change) and qul score.

Input: qul.sas7bdat permanent dataset
Output: frequency and assocation test
produced via PROC FREQ
***************************************************************************************************/

Libname stat482 'C:\Users\esbro\Desktop\STAT 482'; *using libref to reference permanent dataset later on;
DATA studyData (keep=subj v_date first_visit last_visit years first_qul last_qul change score);
  length score $ 8; *making sure the formated score value is displayed correctly;
  set stat482.qul;
  by subj; *sorting data by subject;
  retain first_visit last_visit first_qul last_qul; *have to retain the values to calculate correctly;
  if first.subj and not missing(qul_1)then do;
    first_visit=v_date; *this is first office visit;
    first_qul=qul_1; *first qul_1 score;
  end;
  if last.subj and not missing(qul_1) then do;
    last_visit=v_date; *last office visit;
    last_qul=qul_1; *last qul_1 score;
    years=ROUND(YRDIF(first_visit, last_visit,'ACTUAL')); *getting the number of years between office visits;
    change=first_qul - last_qul; *getting the change in qul_1 score;
    *displaying score correctly via character string value;
      if (change<0) then score='Better';
      else if (change>0) then score='Worse';
      else if (change=0) then score='NoChange';
    output;

PROC PRINT DATA=studyData;
format first_visit last_visit mmddyy10.;
RUN;

PROC FREQ DATA=studyData; *do not need cumulative values;
  Tables years*score/ nopercent nocol chisq; *need chi-square statisitic for association level;
RUN;
```

Problem 5.) SAS Log Window:

```
/*******************************************************************************
***************************************
Filename: C:\Users\Eric\Desktop\STAT 482\final_p5.sas
Written by: Eric Brown
Date: December 12, 2019

This program uses data stored in the permanent dataset called qul. To if
a patient's qul score changes depending on how many office visits they
attend through the study. Then the program tests if there is an association
between years (change) and qul score.

Input: qul.sas7bdat permanent dataset
Output: frequency and assocation test
produced via PROC FREQ
*******************************************************************************
**************************************/

Libname stat482 'C:\Users\esbro\Desktop\STAT 482';
NOTE: Libref STAT482 was successfully assigned as follows:
      Engine:        V9
      Physical Name: C:\Users\esbro\Desktop\STAT 482
                                                         *using libref to reference
permanent dataset later on;
DATA studyData (keep=subj v_date first_visit last_visit years first_qul last_qul
change score);
  length score $ 8; *making sure the formated score value is displayed correctly;
  set stat482.qul;
  by subj; *sorting data by subject;
  retain first_visit last_visit first_qul last_qul; *have to retain the values to
calculate correctly;
  if first.subj and not missing(qul_1)then do;
    first_visit=v_date; *this is first office visit;
    first_qul=qul_1; *first qul_1 score;
  end;
  if last.subj and not missing(qul_1) then do;
    last_visit=v_date; *last office visit;
    last_qul=qul_1; *last qul_1 score;
    years=ROUND(YRDIF(first_visit, last_visit,'ACTUAL')); *getting the number of
years between office visits;
    change=first_qul - last_qul; *getting the change in qul_1 score;
    *displaying score correctly via character string value;
      if (change<0) then score='Better';
      else if (change>0) then score='Worse';
      else if (change=0) then score='NoChange';
    output;
  end;
RUN;

NOTE: There were 2650 observations read from the data set STAT482.QUL.
NOTE: The data set WORK.STUDYDATA has 637 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.04 seconds
      cpu time            0.03 seconds


PROC PRINT DATA=studyData;
format first_visit last_visit mmddyy10.;
RUN;

NOTE: There were 637 observations read from the data set WORK.STUDYDATA.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.03 seconds
      cpu time            0.01 seconds


PROC FREQ DATA=studyData; *do not need cumulative values;
  Tables years*score/ nopercent nocol chisq; *need chi-square statisitic for
association level;
RUN;

NOTE: There were 637 observations read from the data set WORK.STUDYDATA.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      cpu time            0.01 seconds
```

Problem 5.) Program Output:

The FREQ Procedure

Table of years by score

years / score Frequency Row Pct	Better	NoChange	Worse	Total
0	4 3.57	106 94.64	2 1.79	112
1	56 26.54	106 50.24	49 23.22	211
2	34 27.87	56 45.90	32 26.23	122
3	47 38.84	50 41.32	24 19.83	121
4	19 26.76	36 50.70	16 22.54	71
Total	160	354	123	637

Statistics for Table of years by score

Statistic	DF	Value	Prob
Chi-Square	8	91.9228	<.0001
Likelihood Ratio Chi-Square	8	109.2440	<.0001
Mantel-Haenszel Chi-Square	1	1.7436	0.1867
Phi Coefficient		0.3799	
Contingency Coefficient		0.3551	
Cramer's V		0.2686	

Sample Size = 637

Explanation: There is sufficient enough evidence to conclude, there is an association between years and score variables. When generating a Chi-square test, one can see that the p-value which equals less than .0001 indicates the association between the two variables is statistically significant at the 0.05 alpha level. The row percentages show what types of association percentages were present between patients visit years and qul_1 values. One can notice the largest difference between percentages in patients' qul_1 score changes, occurred when patients score did not change with less than one year of scheduled visits compared to when patients had three years of scheduled visits. Patients with less than one year of scheduled visits were more likely to have no change (94.64%) in their qul_1 score than patients who had three years of scheduled visits (41.32%).