STATISTICAL PROGRAMMING FOR BUSINESS ANALYTICS

Assignment 5

Tanay Bhalerao

U47707491

MARCH 20, 2015

UNIVERSITY OF SOUTH FLORIDA
MANAGEMENT INFORMATION SYSTEMS

Homework 5 for Chapter 6

1. A basketball player may be said to have a “hot hand” if he or she makes many consecutive baskets during a game. For example, consider the following 3 players.) indicates a missed shot; 1 indicates a good shot.

   Player A: 10001011100110101001
   Player B: 01111110000000111100
   Player C: 10101010101010101010

   All of the players made 10 of 20 shots. However, A made baskets at random, while both B and C made baskets in a nonrandom pattern. In particular, B has a “hot hand” twice during the game. A nonparametric test of randomness is based on the number of runs, or sets of equal numbers occurring in sequence. For example, Player B had 5 runs (one 0, six 1’s, seven 0’s, four 1’s, and two 0’s), as shown below:

Shot:	0	111111	0000000	1111	00
Run number:	1	2	3	4	5

   Likewise, Player A had 13 runs, and Player C had 20 runs. Notice that the number of runs is equal to (1 + (number of transitions from 0 to 1) + (number of transitions from 1 to 0)). Nonrandomness is suggested when the number of runs is either very large or very small.

   Write a SAS program which calculates the number of runs for Players A, B, and C. Your program must not be data-specific; in other words, it should work for any possible sequence of twenty 0’s and 1’s, not just those represented by Players A, B and C.

```
LIBNAME tanay "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\";

DATA tanay.baskets;
INFILE datalines DLM=":";
LENGTH BASKETS $20;
INPUT Player $ BASKETS $;
a_count     = 1;runs = 0;
DO WHILE(a_count < 20);
a_count = a_count + 1;
IF substrn(BASKETS,a_count,1) NE substrn(BASKETS,a_count-1,1)
then runs = runs+1;
END;
runs = runs + 1;
DROP a_count;

DATALINES;
Player A: 10001011100110101001
Player B: 01111110000000111100
Player C: 10101010101010101010
Player Tan: 10111011010000010110
;
```

```
RUN;
PROC PRINT DATA=tanay.baskets;
TITLE "PLAYER RUNS";
run;
```

PLAYER RUNS

Obs	BASKETS	Player	runs
1	10001011100110101001	Player A	13
2	01111110000000111100	Player B	5
3	10101010101010101010	Player C	20
4	10111011010000010110	Player T	12

2. Refer to the GRADES data. One measure of a student's performance in a class is a stanine score. Stanine is an abbreviation of "standard-nine" and is pronounced stay-nine. The stanine score is an integer from 1 to 9 which reflects how well a student performed in relation to other students in the class. A stanine score of 1 indicates the lowest level of performance; 5 average; and 9, highest. Stanines are used because they transfer well among different scales of measurement and they are somewhat easy to understand and interpret.

   Use SAS to calculate the total of each student's 13 homework grades. Then, use PROC STANDARD to calculate the Z-score for each student, where the Z-score is the total score which has been standardized to have a mean of 0 and a variance of 1. Then, calculate the stanine score according to the chart below:

Z-score	Stanine Score
Below -1.75	1
-1.75 to -1.25	2
-1.25 to -1.75	3
-.75 to -.25	4
-.25 to .25	5
.25 to .75	6
.75 to 1.25	7
1.25 to 1.75	8
1.75 or higher	9

   Print the identification number, Z-score, and stanine score for each student.

```
DATA tanay.grades;
INFILE "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\grades.txt"
DLM="\";
LENGTH Grades $13;
INPUT id Grades   $;
total_grade = 0;
a_count = 0;
DO WHILE(a_count < 13);
a_count = a_count + 1;
```

```
total_grade = total_grade + input(substrn(Grades,a_count,1),$1.);
END;
DROP a_count Grades;
RUN;

PROC STANDARD DATA=tanay.grades MEAN=0 STD=1 OUT=tanay.Z_grad;
VAR total_grade;
RUN;
DATA tanay.Z_grad;
SET tanay.Z_grad;
IF total_grade<-1.75 THEN DO;
Stanine_Score = 1;
Z_value = "Below -1.75";
END;
else if total_grade < -1.25 THEN DO;
Stanine_score = 2;
Z_value = "-1.75 to -1.25";
END;
else if total_grade < -0.75 THEN DO;
Stanine_Score = 3;
Z_value = "-1.25 to -0.75";
END;
ELSE IF total_grade < -0.25 THEN DO;
Stanine_Score = 4;
Z_value = "-0.75 to -0.25";
END;
ELSE IF total_grade < 0.25 THEN DO;
Stanine_Score = 5;
Z_value = "-0.25 to 0.25";
END;
ELSE IF total_grade< 0.75 THEN DO;
Stanine_Score = 6;
Z_value = "0.25 to 0.75";
END;
ELSE IF total_grade< 1.25 THEN DO;
Stanine_Score = 7;
Z_value = "0.75 to 1.25";
END;
ELSE IF total_grade < 1.75 THEN DO;
Z_value = "1.25 to 1.75";
Stanine_Score = 8;
END;
ELSE DO;
Z_value = "1.75 or higher";
Stanine_Score = 9;
END;
DROP total_grade;
RUN;

PROC PRINT DATA=tanay.Z_grad;
title "STANINE SCORE";
ID id;
var Z_value Stanine_score;
RUN;
```

STANINE SCORE

id	Z_value	Stanine_Score
1105	0.25 to 0.7	6
1294	-0.25 to 0.	5
2009	-1.25 to -0	3
2341	-1.25 to -0	3
2354	-1.25 to -0	3
2761	0.75 to 1.2	7
3345	-0.75 to -0	4
3585	0.25 to 0.7	6
3622	-0.25 to 0.	5
3785	0.75 to 1.2	7
3800	-0.75 to -0	4
4232	-0.75 to -0	4
5235	0.25 to 0.7	6
5464	0.75 to 1.2	7
6584	1.25 to 1.7	8
6801	1.25 to 1.7	8
6844	-1.75 to -1	2
7054	-0.25 to 0.	5
7655	0.25 to 0.7	6
8043	-0.75 to -0	4
8553	Below -1.75	1
8744	0.75 to 1.2	7
9086	0.25 to 0.7	6

3. Refer to the IRIS data. Perform a Kruskal-Wallis test to see if the median length varies significantly among the three iris species.

```
DATA tanay.iris;
INFILE "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\iris.txt"
firstobs=2 LRECL=200;
INPUT Class $ 1-10 SL 16-18 SW 24-26 PL 32-34 PW 40-42;
RUN;
PROC SORT DATA = tanay.iris;
BY SL PL;
RUN;
PROC RANK DATA = tanay.iris OUT = tanay.rank_length GROUPS = 151;
VAR SL PL;
RANKS rank_group;
RUN;
PROC PRINT DATA=tanay.rank_length;
RUN;
```

```
PROC NPAR1WAY WILCOXON DATA = tanay.rank_length;
CLASS CLASS;
VAR SL PL;
TITLE "Kruskal-Wallis Test";
      RUN;
```

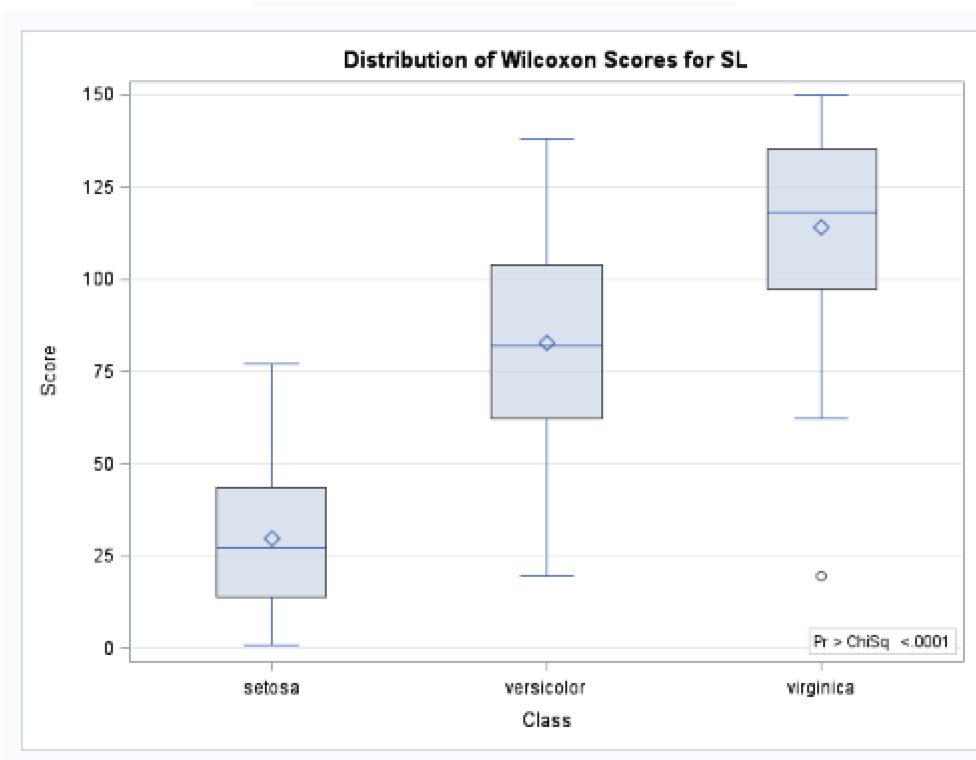
Kruskal-Wallis Test is used when we have one nominal variable and one ranked variable. Basically it checks for the similarity in the mean ranks in all the groups.
In our case we have a nominal variable as class and 4 measurement variables: SL,SW,PL,PW.
Kruskal-Wallis is an one-way ANOVA Test. We will perform the Kruskal-Wallis test on length-SL and PL

Kruskal-Wallis Test

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable SL Classified by Variable Class					
Class	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
setosa	50	1482.00	3775.0	250.603522	29.640
versicolor	50	4132.50	3775.0	250.603522	82.650
virginica	50	5710.50	3775.0	250.603522	114.210
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	96.9374
DF	2
Pr > Chi-Square	<.0001



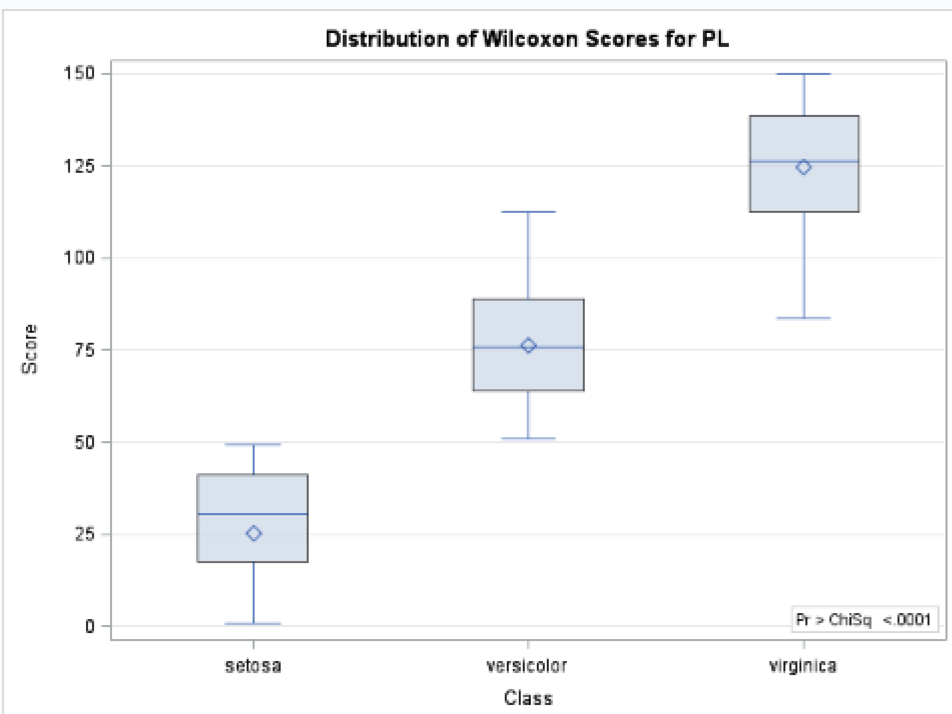
The table Wilcoxon Score rank sums for variable SL. The virginica has a mean score of 114.21 which is higher than the mean scores of both setosa and versicolor. The test statistic of 96.937 indicates that there is a significant difference in class levels across SL (the *p*-value is less than 0.0001).

Kruskal-Wallis Test

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable PL Classified by Variable Class					
Class	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
setosa	50	1275.00	3775.0	250.583124	25.500
versicolor	50	3819.50	3775.0	250.583124	76.390
virginica	50	6230.50	3775.0	250.583124	124.610
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	130.4141
DF	2
Pr > Chi-Square	<.0001



The table Wilcoxon Score rank sums for variable PL. The virginica has a mean score of 124.61 which is higher than the mean scores of both setosa and versicolor. The test statistic of 130.4141 indicates that there is a significant difference in class levels across PL (the *p*-value is less than 0.0001).

4. Refer to the DOGS data. Create a variable which indicates whether each dog's eosinophil count in Week 2 was above or below the median of all 25 eosinophil counts in Week 2. Then, prepare a frequency table showing the cross-classification of drug concentration with the indicator for the median (a 3x2 table), and perform Fisher's exact test on this table.

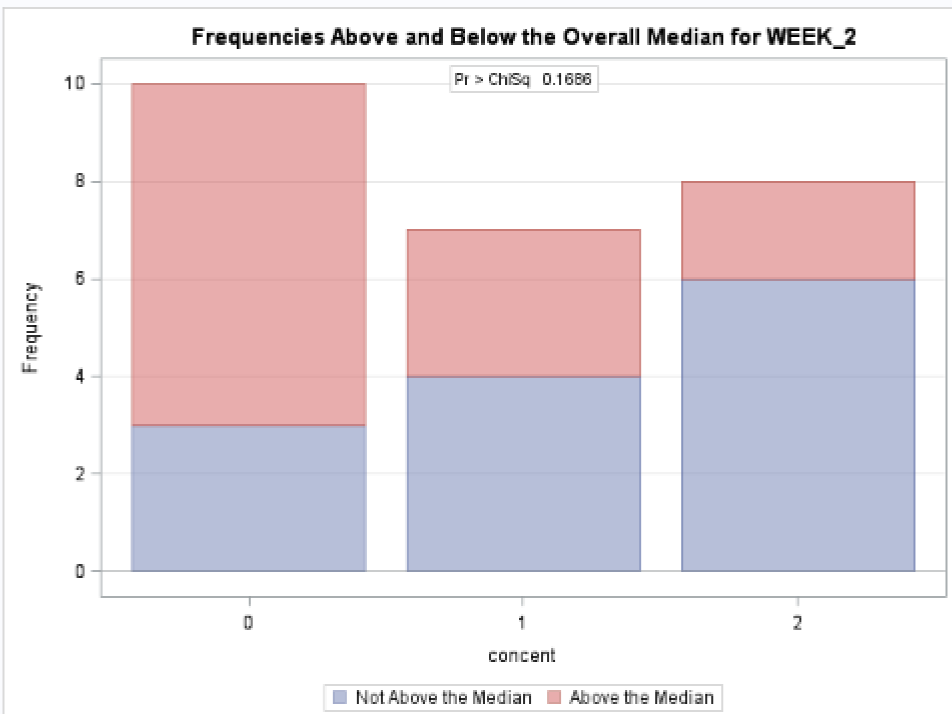
```
DATA tanay.DOGS3;
INFILE "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\DOGS3.TXT"
FIRSTOBS = 3;
INPUT name_dog $ 1-8 WEEK_0 9-16 WEEK_2 17-24 WEEK_4 25-32;
RUN;

PROC MEANS DATA= tanay.DOGS3 median;
var WEEK_2;
RUN;

DATA infile_dogs1;
infile "\\Client\C$\Users\Tanay\Documents\Sem2\BusinessAnalytics\dogs1.txt"
LRECL= 200 firstobs=2;
input name_dog $ 1-8 concent 16 sex $ 17 age 31-32 haircoat $ 33-37 weight
45-48;
RUN;
PROC PRINT DATA=infile_dogs1;
RUN;

proc sort Data=tanay.DOGS3;
by name_dog;
run;

proc sort Data=infile_dogs1;
by name_dog;
run;
Data doggy;
      merge tanay.DOGS3 infile_dogs1;
      by name_dog;
      drop sex age haircoat weight;
run;
Data Dogs_f;
set doggy;
if WEEK_2 >375 then IND = 'ABOVE';
else if WEEK_2 <375 THEN IND = 'BELOW';
run;
proc npar1way data=Dogs_f median;
class concent;
var WEEK_2;
title "Dogs";
run;
proc freq Data=Dogs_f;
title "Fisher's Test";
tables concent*IND / fisher;
run;
```

Dogs

The NPAR1WAY Procedure

Median Scores (Number of Points Above Median) for Variable WEEK_2 Classified by Variable concent					
concent	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
0	10	7.0	4.800	1.249000	0.700000
1	7	3.0	3.360	1.144727	0.428571
2	8	2.0	3.840	1.189285	0.250000
Average scores were used for ties.					

Median One-Way Analysis	
Chi-Square	3.5604
DF	2
Pr > Chi-Square	0.1686



Fisher's Test

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of concent by IND		
	IND		
concent	ABOVE	BELOW	Total
0	7 29.17 77.78 58.33	2 8.33 22.22 16.67	9 37.50
1	3 12.50 42.86 25.00	4 16.67 57.14 33.33	7 29.17
2	2 8.33 25.00 16.67	6 25.00 75.00 50.00	8 33.33
Total	12 50.00	12 50.00	24 100.00
Frequency Missing = 1			

Statistics for Table of concent by IND

Statistic	DF	Value	Prob
Chi-Square	2	4.9206	0.0854
Likelihood Ratio Chi-Square	2	5.1783	0.0751
Mantel-Haenszel Chi-Square	1	4.5774	0.0324
Phi Coefficient		0.4528	
Contingency Coefficient		0.4125	
Cramer's V		0.4528	
WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Table Probability (P)	0.0130
Pr <= P	0.1194

Effective Sample Size = 24
Frequency Missing = 1

Fisher’s exact test of independence when we have two nominal variables and we want to see if the proportions of one variable are different depending on the value of the other variable. The p-value for the Fisher's exact test is 0.1194 which is high. So the drug concentration is statistically independent of the median of frequency of Week 2. We do not reject the null hypothesis.

5. Chapter 6: 6.2 and 6.4

6.2

```
DATA tanay.Reading_prog;
INPUT Prog_grp $ Score @@;

DATALINES;

CODY 500 CODY 450 CODY 505 CODY 404 CODY 555 CODY 567 CODY 588 CODY 577
CODY 566 CODY 644 CODY 511 CODY 522 CODY 543 CODY 578
SMITH 355 SMITH 388 SMITH 440 SMITH 600 SMITH 510 SMITH 501 SMITH 502
SMITH 489 SMITH 499 SMITH 489 SMITH 515 SMITH 520 SMITH 520 SMITH 480
;
RUN;

Proc TTEST Data = tanay.Reading_prog;
Class Prog_grp;
TITLE "T-TEST";
Var Score;
Run;

Proc NPAR1WAY Data = tanay.Reading_prog;
CLASS Prog_grp;
VAR Score;
EXACT WILCOXON;
TITLE "EXACT Wilcoxon Test";
Run;
```

T-TE ST

The TTEST Procedure

Variable: Score

Prog_grp	N	Mean	Std Dev	Std Err	Minimum	Maximum
CODY	14	536.4	60.7513	16.2365	404.0	644.0
SMITH	14	486.3	59.6843	15.9513	355.0	600.0
Diff (1-2)		50.1429	60.2202	22.7611		

Prog_grp	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
CODY		536.4	501.4	571.5	60.7513	44.0419	97.8730
SMITH		486.3	451.8	520.7	59.6843	43.2684	96.1540
Diff (1-2)	Pooled	50.1429	3.3568	96.9290	60.2202	47.4244	82.5277
Diff (1-2)	Satterthwaite	50.1429	3.3560	96.9297			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	26	2.20	0.0367
Satterthwaite	Unequal	25.992	2.20	0.0367

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	13	13	1.04	0.9500

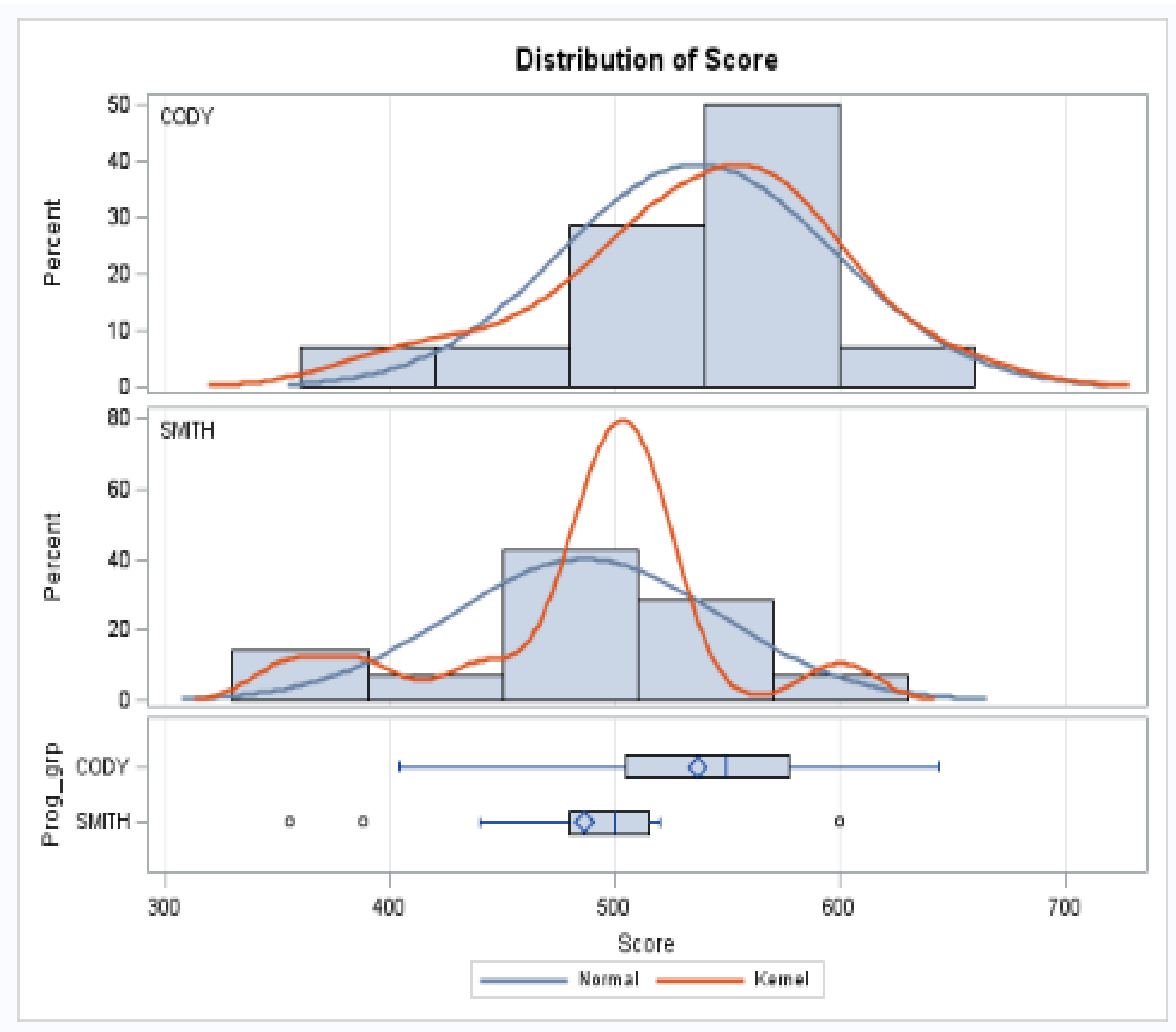
For the T-Test

p-value for the Student's t–test(Pooled):

Pr > |t| =0.0367

p-value for the Welsh's t-test(Satterthwaite):

Pr>|t|=0.0367



The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Score Classified by Variable Prog_grp					
Prog_grp	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
CODY	14	254.0	203.0	21.757927	18.142857
SMITH	14	152.0	203.0	21.757927	10.857143
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic (S)	254.0000
Normal Approximation	
Z	2.3210
One-Sided Pr > Z	0.0101
Two-Sided Pr > |Z|	0.0203
t Approximation	
One-Sided Pr > Z	0.0140
Two-Sided Pr > |Z|	0.0281
Exact Test	
One-Sided Pr >= S	0.0090
Two-Sided Pr >= |S - Mean|	0.0179
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test	
Chi-Square	5.4942
DF	1
Pr > Chi-Square	0.0191

For Wilcoxon Two-Sample Test:

Normal approximation, p-value:

Two-sided Pr>|z| = 0.0203

Exact Test, p-value:

Two-Sided Pr>=|S-Mean| = 0.0179

Wilcoxon test is based on the median value while T-Test is based on the mean value.If data is not symmetric both tests will give different p-values. In the above case, it is observed that the distribution is different and the p-values are not that different to be called significant for carrying out two tests.

6.4

```
DATA QUES6_4;
DO GROUP = 'A','B','C';
DO I = 1 TO 10;
X = ROUND(RANNOR(135)*10 + 300 +5*(GROUP EQ 'A') - 7*(GROUP EQ 'C'));

Y = ROUND(RANUNI(135)*100 + X);
OUTPUT;
END;
END;
DROP I;
RUN;

Proc TTEST Data = QUES6_4;
TITLE "T-TEST -Q 6.4";
CLASS Group;
VAR X Y;
WHERE Group = 'A' or Group ='C';
      Run;
```

T-TEST -Q 6.4

The TTEST Procedure

Variable: X

GROUP	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	10	305.1	11.6376	3.6801	281.0	321.0
C	10	288.4	7.5011	2.3721	276.0	299.0
Diff (1-2)		16.7000	9.7903	4.3784		

GROUP	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		305.1	296.8	313.4	11.6376	8.0047	21.2457
C		288.4	283.0	293.8	7.5011	5.1595	13.6941
Diff (1-2)	Pooled	16.7000	7.5014	25.8986	9.7903	7.3977	14.4781
Diff (1-2)	Satterthwaite	16.7000	7.3877	26.0123			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	3.81	0.0013
Satterthwaite	Unequal	15.377	3.81	0.0016

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	2.41	0.2068

Variable: Y

GROUP	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	10	373.7	23.9214	7.5646	326.0	405.0
C	10	338.1	38.2723	12.1028	291.0	387.0
Diff (1-2)		35.6000	31.9139	14.2724		

GROUP	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		373.7	356.6	390.8	23.9214	16.4540	43.6711
C		338.1	310.7	365.5	38.2723	26.3250	69.8702
Diff (1-2)	Pooled	35.6000	5.6149	65.5851	31.9139	24.1146	47.1951
Diff (1-2)	Satterthwaite	35.6000	5.1969	66.0031			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	2.49	0.0226
Satterthwaite	Unequal	15.101	2.49	0.0247

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	2.56	0.1777

6. Chapter 7: 7.4 and 7.10

7.4

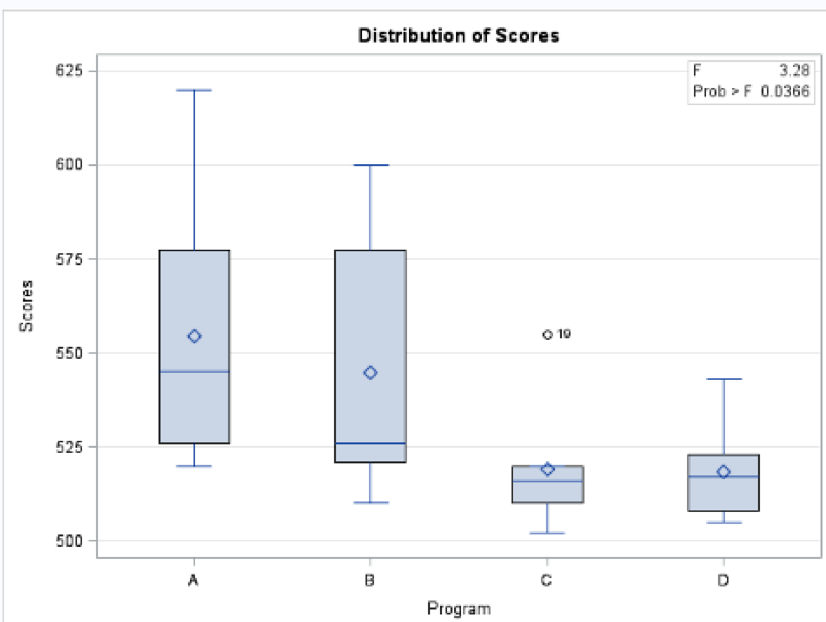
```
Data Entrance_scores;
Input Program $ Scores @@;
Datalines;
A 560 A 520 A 530 A 525 A 575 A 527 A 580 A 620
B 565 B 522 B 520 B 530 B 510 B 522 B 600 B 590
C 512 C 518 C 555 C 502 C 510 C 520 C 516
D 505 D 508 D 512 D 520 D 543 D 523 D 517
;
Run;

PROC GLM DATA= Entrance_scores;
CLASS Program;
Model Scores = Program;
CONTRAST "A and B - C and D" Program 1 1 -1 -1;
CONTRAST "D - A B C" Program 1 1 1 -3;
TITLE "Preparation Methods";
RUN;
```

Preparation Methods

The GLM Procedure

Dependent Variable: Scores

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
A and B - C and D	1	7225.152381	7225.152381	9.36	0.0051
D - A B C	1	2413.012158	2413.012158	3.13	0.0888



7.10

```
DATA RAT_MAZE;
INPUT AGE $ STRAIN $ SPEED @@;
DATALINES;
3Mo. A 12 3Mo. A 14 3Mo. A 9 3Mo. A 17 3Mo. A 10 3Mo. A 11 3Mo. A 9 3Mo. A 10
3Mo. B 24 3Mo. B 17 3Mo. B 22 3Mo. B 16 3Mo. B 18 6Mo. A 22 6Mo. A 20 6Mo. A
12 6Mo. A 12 6Mo. A 17 6Mo. A 14 6Mo. A 17 6Mo. B 23 6Mo. B 26 6Mo. B 34 6Mo.
B 20 9Mo. A 14 9Mo. A 14 9Mo. A 10 9Mo. A 15 9Mo. A 17 9Mo. A 12 9Mo. A 19
9Mo. B 27 9Mo. B 29 9Mo. B 27 9Mo. B 23
;
RUN;

PROC ANOVA DATA = RAT_MAZE;
TITLE "TWO-WAY ANOVA-SPEED";
CLASS AGE STRAIN;
MODEL SPEED = AGE | STRAIN;
MEANS AGE|STRAIN / SNK;
RUN;
```

TWO-WAY ANOVA-SPEED

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
AGE	3	3Mo. 6Mo. 9Mo.
STRAIN	2	A B

Number of Observations Read	35
Number of Observations Used	35

TWO-WAY ANOVA-SPEED

The ANOVA Procedure

Dependent Variable: SPEED

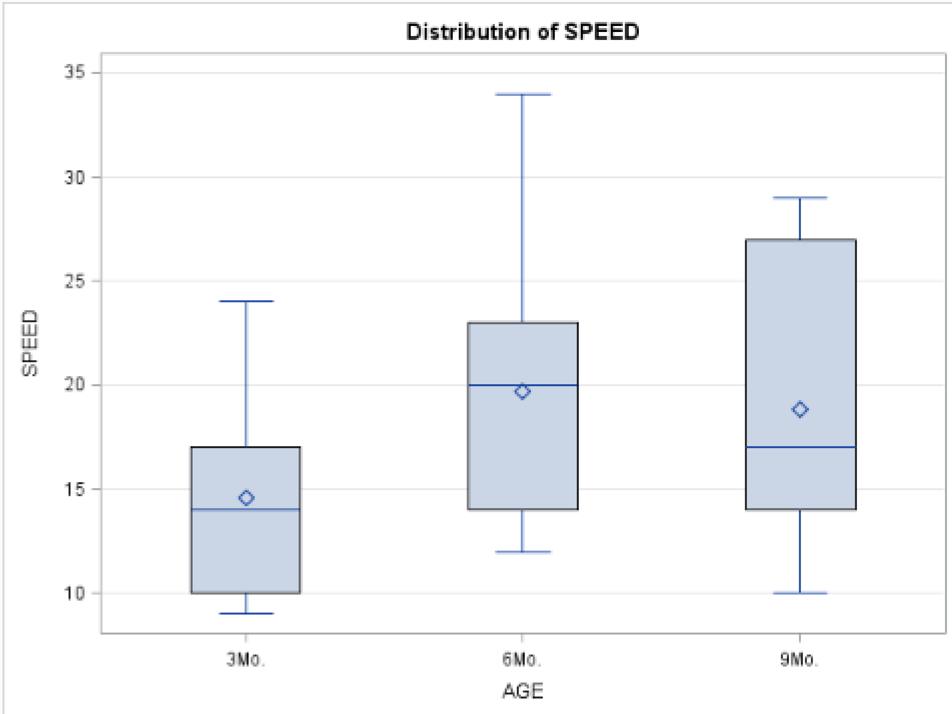
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	978.650000	195.730000	15.25	<.0001
Error	29	372.092857	12.830788		
Corrected Total	34	1350.742857			

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

Source	DF	Anova SS	Mean Square	F Value	Pr > F
AGE	2	187.6939061	93.8469530	7.31	0.0027
STRAIN	1	750.5575425	750.5575425	58.50	<.0001
AGE*STRAIN	2	40.3985514	20.1992757	1.57	0.2244

TWO-WAY ANOVA-SPEED

The ANOVA Procedure



TWO-WAY ANOVA-SPEED

The ANOVA Procedure

Student-Newman-Keuls Test for SPEED

Note: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha	0.05
Error Degrees of Freedom	29
Error Mean Square	12.83079
Harmonic Mean of Cell Sizes	11.59459

Note: Cell sizes are not equal.

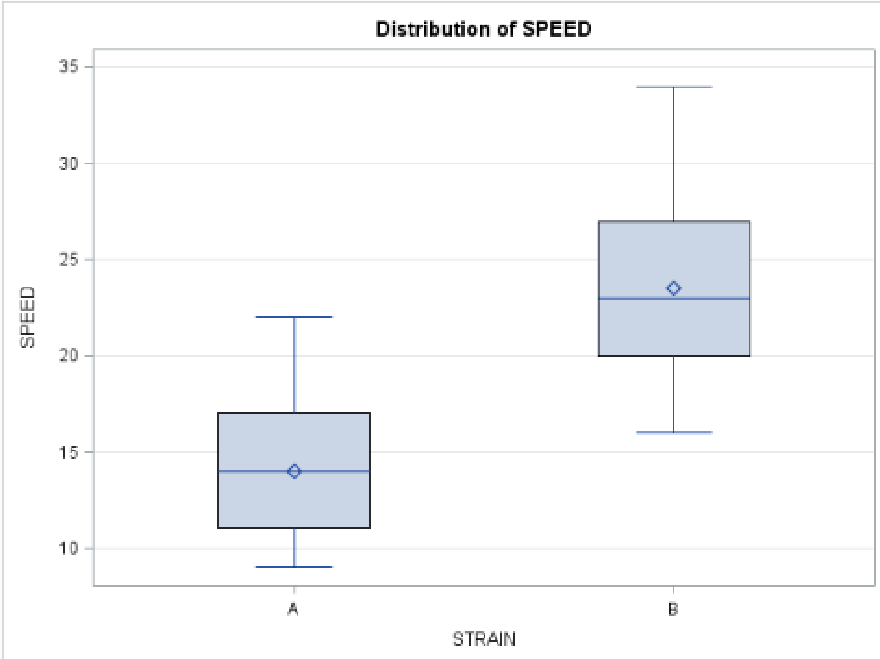
Number of Means	2	3
Critical Range	3.0426538	3.6740289

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	AGE
A	19.727	11	6Mo.
A			
A	18.818	11	9Mo.
B	14.538	13	3Mo.

TWO-WAY ANOVA-SPEED

The ANOVA Procedure



TWO-WAY ANOVA-SPEED

The ANOVA Procedure

Student-Newman-Keuls Test for SPEED

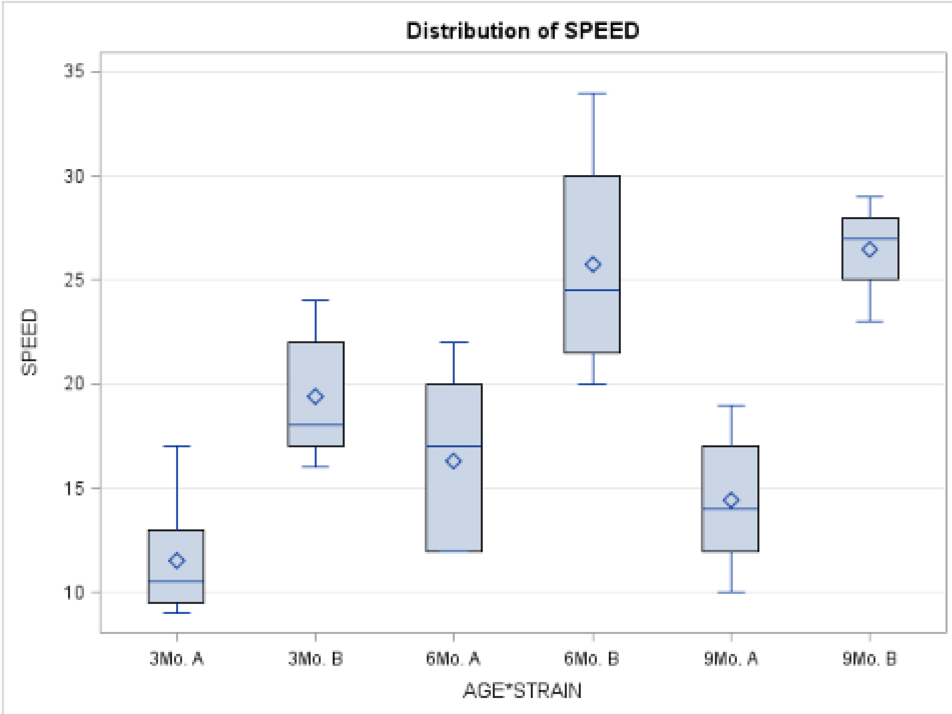
Note: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha	0.05
Error Degrees of Freedom	29
Error Mean Square	12.83079
Harmonic Mean of Cell Sizes	16.34286

Note: Cell sizes are not equal.

Number of Means	2
Critical Range	2.5628095

Means with the same letter are not significantly different.			
SNK Grouping	Mean	N	STRAIN
A	23.538	13	B
B	13.955	22	A



Level of AGE	Level of STRAIN	N	SPEED	
			Mean	Std Dev
3Mo.	A	8	11.5000000	2.77746030
3Mo.	B	5	19.4000000	3.43511281
6Mo.	A	7	16.2857143	3.86066858
6Mo.	B	4	25.7500000	6.02079729
9Mo.	A	7	14.4285714	2.99205297
9Mo.	B	4	26.5000000	2.51661148