Seneca

SENECA COLLEGE OF APPLIED ARTS AND TECHNOLOGY

SENECA BUSINESS

BAN100 - Statistics for Analytics

Other **Version NA**

DATE: 6/15/2023 **TIME ALLOWED: 14 days**

PROFESSOR(S): Samaneh Gholami

Allowable Examination Aids: (check applicable boxes)

☒ Calculators (non-programmable only)	☒ Math Tables (normal distribution table)	☒ Periodic Tables
☐ Formula Sheets (attached)	☒ Textbooks	☒ Probability Tables
☒ Dictionary	☒ Notes	☐ Other

Answers to be completed on:

☐ Exam Booklet	☐ GradeMaster Card	☐ Exam Paper

TOTAL MARKS: 25 **WEIGHTED VALUE: 25**

INSTRUCTIONS:

Academic Integrity Policy. Seneca upholds a learning community that values academic integrity, honesty, fairness, trust, respect, responsibility and courage. These values enhance Seneca's commitment to students by delivering high-quality education and teaching excellence, while supporting a positive learning environment. The AI policy is always in effect. Note **Sections 2.3 and 2.4:**

*"**…2.3** Should there be a suspected violation of this policy (e.g.…cheating, falsification, impersonation or plagiarism), the academic integrity sanctions will be applied according to the severity of the offence committed. Refer to Appendix B for the academic integrity sanctions. **2.4** Should a suspected violation of this policy be a result of, or in combination with, a suspected violation of Seneca's Student Code of Conduct and/or another non-academic-related Seneca policy, the matter will be investigated and adjudicated through the process found in the Student Code of Conduct."*

TO BE COMPLETED BY STUDENT
SUBJECT SECTION NUMBER (e.g. QNM223 AA):
STUDENT NAME: Ugonna Okengwu
STUDENT NUMBER: 114939192
STUDENT SIGNATURE:

APPROVED BY: _____

Cristina Italia, Interim Chair
School of Management and Entrepreneurship

DATE: _____

Problem 1

The code below was used for the analysis.

6/30/23, 5:46 PM Code: ASSIGNMENT2.sas

```
/*Ugonna Okengwu*/

proc import Datafile = '/home/u63417899/BAN100ZBB/File_Proportion_of_Total_Assets_Invested_in_Stocks (1).xlsx'
    OUT= work.stock_asset
    DBMS=XLSX
    replace;
    GETNAMES=Yes;
run;

proc print data = Stock_Asset;
run;

Title 'Anova analysis';
proc anova data = Stock_Asset;
class Age_range;
model scale = Age_range;
run;

Title 'Q Box Plot';
proc sgplot data = stock_asset;
vbox scale / category = Age_range;
run;

Title 'Q-Q plot';
proc univariate data = stock_asset;
ppplot scale;
run;
```

The numerical variable is the only continuous variable that is the proportion of financial assets invested in the stock market. The age of head of the household has 4 groups which are young, middle age, late middle age and senior. We will be using one-way Anova. Having at least 3 different levels of categorical variable is a requirement for running a one-way anova as it is used to compare the means of more than 3 groups.

The assumptions for the one-way are:

- The populations are normally distributed.
- The population have equal variance.

The date set was converted to a long data structure by grouping their ages into a range before performing the analysis. Two variables were used for the analysis, Age range (categorical data) and Scale (Numerical data). The investigation was carried out in 3 steps:

Results: ASSIGNMENT2.sas

Obs	Young	Early_Middle_Age	Late_Middle_Age	Senior	Scale	Age_range	G	H
1	24.8	28.9	81.5	66.8	24.8	Young		
2	35.5	7.3	0.0	77.4	35.5	Young		
3	68.7	61.8	61.3	32.9	68.7	Young		
4	42.2	53.6	0.0	74.0	42.2	Young		
5	49.5	0.0	45.4	0.0	49.5	Young		
6	64.6	49.4	42.3	35.2	64.6	Young		
7	58.3	71.4	75.3	21.4	58.3	Young		
8	72.0	53.7	54.7	0.0	72.0	Young		
9	25.6	46.9	0.0	61.4	25.6	Young		
10	39.8	91.6	20.5	61.8	39.8	Young		
11	39.3	46.0	76.4	35.6	39.3	Young		
12	55.6	41.8	38.0	53.0	55.6	Young		
13	0.0	53.2	39.8	38.5	0.0	Young		
14	56.5	0.0	78.4	53.7	56.5	Young		
15	37.3	43.7	0.0	69.1	37.3	Young		
16	50.3	78.1	76.7	55.5	50.3	Young		
17	38.0	54.7	72.7	31.6	38.0	Young		
18	42.7	45.7	0.0	0.0	42.7	Young		
19	48.4	63.1	33.0	57.3	48.4	Young		
20	18.3	50.4	11.0	42.7	18.3	Young		
21	50.1	38.6	0.0	37.9	50.1	Young		
22	77.2	59.8	60.3	60.9	77.2	Young		
23	42.7	67.8	89.3	72.3	42.7	Young		
24	0.0	48.4	56.0	45.8	0.0	Young		
25	0.0	0.0	47.9	69.0	0.0	Young		
26	46.3	60.6	36.0	41.6	46.3	Young		
27	26.8	66.4	60.0	2.3	26.8	Young		
28	15.3	52.2	47.8	49.7	15.3	Young		
29	36.6	56.1	67.2	43.3	36.6	Young		
30	35.5	45.0	61.8	68.8	35.5	Young		
31	70.0	80.7	61.4	100.0	70.0	Young		
32	35.8	37.4	68.7	46.3	35.8	Young		
33	0.0	94.9	30.8	45.7	0.0	Young		
34	45.0	58.1	84.9	17.5	45.0	Young		
35	66.9	51.8	77.4	62.2	66.9	Young		
36	52.1	43.5	34.6	69.4	52.1	Young		
37	64.6	50.9	40.5	48.6	64.6	Young		

Results: ASSIGNMENT2.sas

Obs	Young	Early_Middle_Age	Late_Middle_Age	Senior	Scale	Age_range	G	H
99	.	43.8	.	.	43.7	Early_Middle_Age		
100	.	50.8	.	.	78.1	Early_Middle_Age		
101	.	59.6	.	.	54.7	Early_Middle_Age		
102	.	48.2	.	.	45.7	Early_Middle_Age		
103	.	80.1	.	.	63.1	Early_Middle_Age		
104	.	44.5	.	.	50.4	Early_Middle_Age		
105	.	57.9	.	.	38.6	Early_Middle_Age		
106	.	55.2	.	.	59.8	Early_Middle_Age		
107	.	53.8	.	.	67.8	Early_Middle_Age		
108	.	60.3	.	.	48.4	Early_Middle_Age		
109	.	64.7	.	.	0.0	Early_Middle_Age		
110	.	44.9	.	.	60.6	Early_Middle_Age		
111	.	66.0	.	.	66.4	Early_Middle_Age		
112	.	67.9	.	.	52.2	Early_Middle_Age		
113	.	30.4	.	.	56.1	Early_Middle_Age		
114	.	0.0	.	.	45.0	Early_Middle_Age		
115	.	52.5	.	.	80.7	Early_Middle_Age		
116	.	56.2	.	.	37.4	Early_Middle_Age		
117	.	20.6	.	.	94.9	Early_Middle_Age		
118	.	70.2	.	.	58.1	Early_Middle_Age		
119	.	9.5	.	.	51.8	Early_Middle_Age		
120	.	37.3	.	.	43.5	Early_Middle_Age		
121	.	71.1	.	.	50.9	Early_Middle_Age		
122	.	45.0	.	.	39.8	Early_Middle_Age		
123	.	73.2	.	.	41.6	Early_Middle_Age		
124	.	47.1	.	.	68.0	Early_Middle_Age		
125	.	6.7	.	.	58.7	Early_Middle_Age		
126	.	56.5	.	.	84.8	Early_Middle_Age		
127	.	41.2	.	.	50.1	Early_Middle_Age		
128	.	63.5	.	.	55.4	Early_Middle_Age		
129	.	71.1	.	.	47.3	Early_Middle_Age		
130	.	45.6	.	.	87.2	Early_Middle_Age		
131	.	84.0	.	.	0.0	Early_Middle_Age		
132	.	.	.	.	50.6	Early_Middle_Age		
133	.	.	.	.	44.2	Early_Middle_Age		
134	.	.	.	.	54.8	Early_Middle_Age		

1. Hypothesis.

H0: $\mu 1 = \mu^2 = \mu^3 = \mu^4$: There is no difference in the mean stock assets across the 4 age groups for American households. This is the Null hypothesis.

H0: $\mu 1 \neq \mu^2$ for one pair: At least one of the age groups has a different stock asset. This is the alternative hypothesis.

2. Sampling the distribution.

The F statistics distribution was sampled for the one-way Anova. The F value is 2.79. The P value is 0.0405 making it less than 0.05, we reject the null hypothesis and go for the alternative hypothesis.

Anova analysis

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Age_range	4	Early_Middle_Age Late_Middle_Age Senior Young

Number of Observations Read	366
Number of Observations Used	366

Anova analysis

The ANOVA Procedure

Dependent Variable: Scale Scale

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3741.3636	1247.1212	2.79	0.0405
Error	362	161870.9817	447.1574		
Corrected Total	365	165612.3453			

R-Square	Coeff Var	Root MSE	Scale Mean
0.022591	42.14046	21.14610	50.18003

Source	DF	Anova SS	Mean Square	F Value	Pr > F

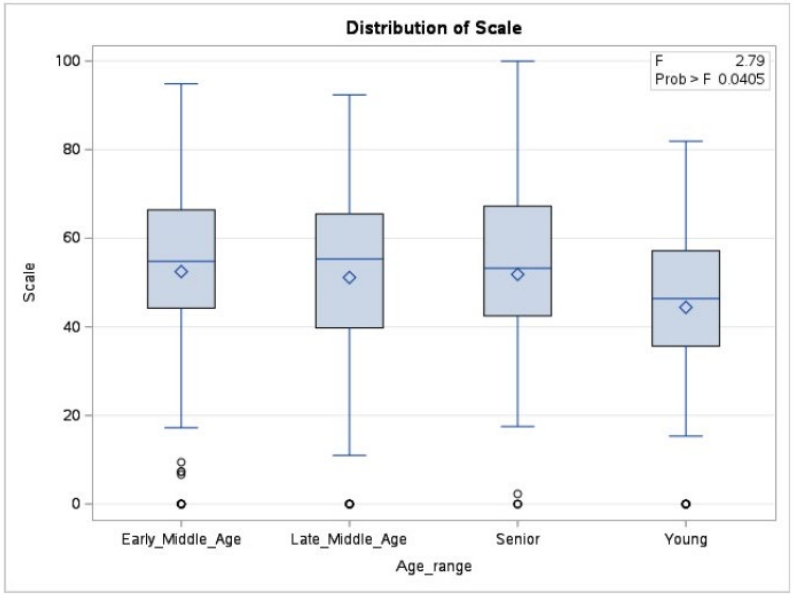
da.sas.com/SASStudio/sasexec/submissions/0161a590-3d0b-43ef-958e-44db0a0627a9/results

Results: ASSIGNMENT2.sas

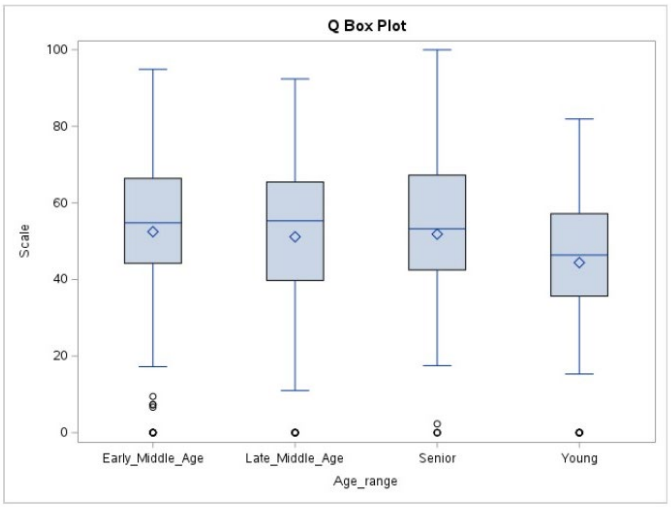
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Age_range	3	3741.363610	1247.121203	2.79	0.0405

The third table shows both the P value and the F value, it also shows the Model source and Error source within the group's variation.

The distribution of scale produced from the proc anova module is the same as the Q box plot obtained below using the proc sgplot module.

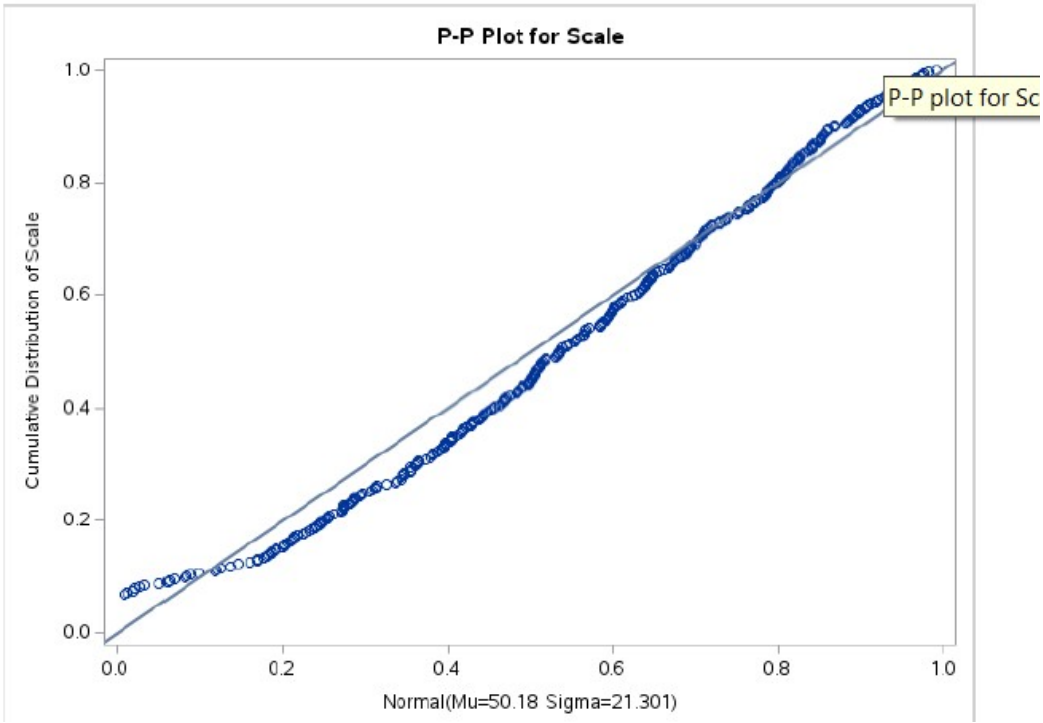


The Q box plot shows the data points are distributed for each group. The mean and median both have good spreads. From the plot, the mean for each group is almost the same. The early middle age and senior group mean stock asset are about 51 while the late middle age group is slightly lower. The outliers of each group are beyond the minimum cutoff point.



Q-Q plot

The skewness and kurtosis values are 0 which indicates a normal distribution. The standard deviation for all groups is within the range of 19.6 to 21.6 which is close. The breakdown of the mean also proves that the mean stock assets for all the age groups though close are not the same.



3. Conclusion

After much analysis it is inferred that the stock ownership varies with the age. These servers as evidence to backup the alternative hypothesis that states; the mean stock ownership of at least one of the age groups differs concluding that the age group influences the ownership of stocks.

Results: ASSIGNMENT2.sas

Variable: Scale (Scale)

Moments			
N	366	Sum Weights	366
Mean	50.1800273	Sum Observations	18365.89
Std Deviation	21.3009965	Variance	453.732453
Skewness	-0.6627725	Kurtosis	0.35760488
Uncorrected SS	1087213.21	Corrected SS	165612.345
Coeff Variation	42.4491529	Std Error Mean	1.11342092

Basic Statistical Measures			
Location		Variability	
Mean	50.18003	Std Deviation	21.30100
Median	52.04000	Variance	453.73245
Mode	0.00000	Range	99.97000
		Interquartile Range	26.12000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	45.06834	Pr > |t|	<.0001
Sign	M	170.5	Pr >= |M|	<.0001
Signed Rank	S	29155.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	99.97
99%	91.57
95%	80.73
90%	74.01
75% Q3	65.39
50% Median	52.04
25% Q1	39.27
10%	20.62
5%	0.00
1%	0.00
0% Min	0.00

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	326	91.19	145
0	316	91.57	94
0	313	92.37	290
0	293	94.87	117
0	236	99.97	339

Problem 2

Same with the first problem, One- way avon was used for conducting the analysis. I manipulated the data set on SA to achieve the long format. The data set was grouped into scale and gender to conduct the analysis effectively. See code below.

```
Title 'Q-Q plot';
proc univariate data = stock_asset;
ppplot scale;
run;


/*Problem 2*/
proc import Datafile = '/home/u63417899/BAN100ZBB/File_Comparing_the_Lifetime_of_Jobs_by_Educational_Level.xlsx'
    OUT=Job_tenure
    DBMS=XLSX
    replace;
    GETNAMES=Yes;
run;

proc print data =Job_tenure;
run;

/*Creating dataset for Male range of E1, E2, E3 and E4*/
Data Male_E1 (Drop= Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4);
set Job_tenure;
if Male_E1 = . then delete;
Rename Male_E1 = Scale;
Gender_range = 'Male_E1';
run;

Data Male_E2 (Drop= Male_E1 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4);
set Job_tenure;
if Male_E2 = . then delete;
Rename Male_E2 = Scale;
Gender_range = 'Male_E2';
run;

Data Male_E3 (Drop= Male_E2 Male_E1 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4);
set Job_tenure;
if Male_E3 = . then delete;
Rename Male_E3 = Scale;
Gender_range = 'Male_E3';
run;

Data Male_E4 (Drop= Male_E2 Male_E3 Male_E1 Female_E1 Female_E2 Female_E3 Female_E4);
set Job_tenure;
if Male_E4 = . then delete;
Rename Male_E4 = Scale;
Gender_range = 'Male_E4';
run;
```



```
/*Creating dataset for Female range of E1, E2, E3 and E48*/
Data Female_E1 (Drop= Male_E1 Male_E2 Male_E3 Male_E4 Female_E2 Female_E3 Female_E4);
set Job_tenure;
if Female_E1 = . then delete;
Rename Female_E1 = Scale;
Gender_range = 'Female_E1';
run;

Data Female_E2 (Drop= Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E3 Female_E4);
```

```
/*Creating dataset for Female range of E1, E2, E3 and E48*/
Data Female_E1 (Drop= Male_E1 Male_E2 Male_E3 Male_E4 Female_E2 Female_E3 Female_E4);
set Job_tenure;
if Female_E1 = . then delete;
Rename Female_E1 = Scale;
Gender_range = 'Female_E1';
run;

Data Female_E2 (Drop= Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E3 Female_E4);
set Job_tenure;
if Female_E2 = . then delete;
Rename Female_E2 = Scale;
Gender_range = 'Female_E2';
run;

Data Female_E3 (Drop= Male_E1 Male_E2 Male_E3 Male_E4 Female_E2 Female_E1 Female_E4);
set Job_tenure;
if Female_E3 = . then delete;
Rename Female_E3 = Scale;
Gender_range = 'Female_E3';
run;

Data Female_E4 (Drop= Male_E1 Male_E2 Male_E3 Male_E4 Female_E2 Female_E3 Female_E1);
set Job_tenure;
if Female_E4 = . then delete;
Rename Female_E4 = Scale;
Gender_range = 'Female_E4';
run;

Data Job_tenure2;
Length Scale 8 Gender_range $ 90;
set male_e1 male_e2 male_e3 male_e4 Female_E1 Female_E2 Female_E3 Female_E4 ;
run;

proc print data = Job_tenure2;
run;


/* A. Test for interaction between gender and education */
proc glm data=Job_tenure2;
class gender_range;
model scale = Gender_range;
means Gender_range;
title 'Test for interaction between Gender and Education';
run;

Title 'Q box plot';
proc sgplot data = Job_tenure2;
vbox scale/ category = Gender_range;
run;

Title 'Q-Q plot';
proc univariate data = job_tenure2;
ppplot scale;
run;
```

1. Hypothesis

Null hypothesis (H0): There is no interaction between gender and education in holding jobs.

H0: $\mu 1 = \mu 2$.

Alternative hypothesis (H1): There is an interaction between gender and education in holding jobs.

Ha: $\mu 1 \neq \mu 2$

Test for interaction between Gender and Education

The GLM Procedure

Class Level Information		
Class	Levels	Values
Gender_range	8	Female_E1 Female_E2 Female_E3 Female_E4 Male_E1 Male_E2 Male_E3 Male_E4

Number of Observations Read	80
Number of Observations Used	80

Test for interaction between Gender and Education

The GLM Procedure

Dependent Variable: Scale Male_E1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	153.3500000	21.9071429	2.17	0.0467
Error	72	726.2000000	10.0861111		
Corrected Total	79	879.5500000			

R-Square	Coeff Var	Root MSE	Scale Mean

id-usw2.oda.sas.com/SASStudio/sasexec/submissions/9f4bed06-4217-405b-b026-ce6c3ab29b3f/results

PM Results: ASSIGNMENT2.sas

R-Square	Coeff Var	Root MSE	Scale Mean
0.174351	30.46392	3.175864	10.42500

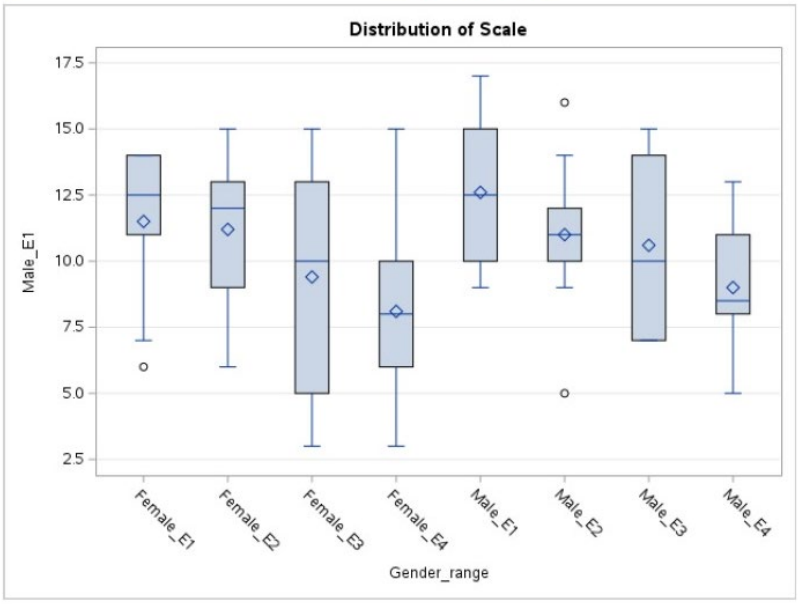
Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gender_range	7	153.3500000	21.9071429	2.17	0.0467

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gender_range	7	153.3500000	21.9071429	2.17	0.0467

2. Sampling the distribution.

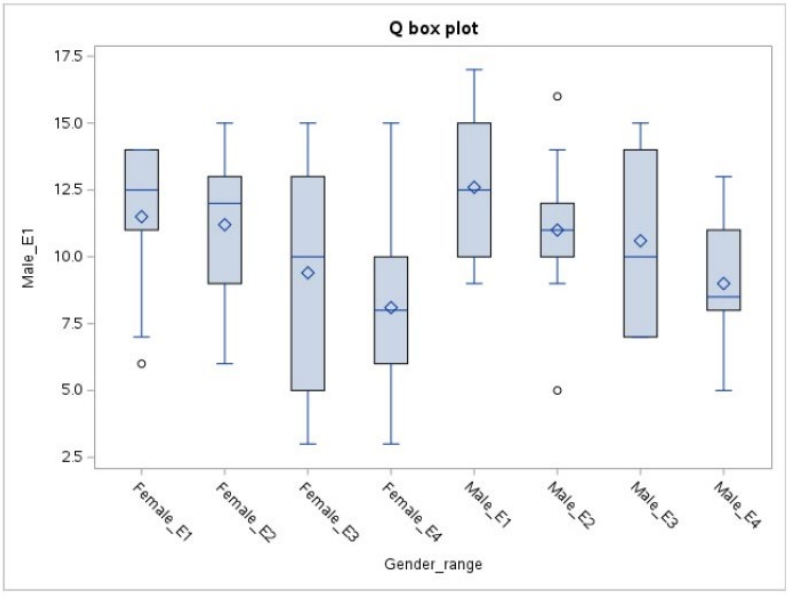
The F statistics distribution was sampled for the one-way Anova. The F value is 2.17. The P value is 0.0487 making it less than 0.05, we reject the null hypothesis and go for the alternative hypothesis.

The third table shows both the P value and the F value, it also shows the Model source and Error source within the group's variation.



Level of Gender_range	N	Scale	
		Mean	Std Dev
Female_E1	10	11.5000000	2.87711275
Female_E2	10	11.2000000	3.11982906
Female_E3	10	9.4000000	4.06065129
Female_E4	10	8.1000000	3.51030230
Male_E1	10	12.6000000	2.87518115
Male_E2	10	11.0000000	2.94392029
Male_E3	10	10.6000000	3.40587727
Male_E4	10	9.0000000	2.30940108

The distribution of scale produced from the proc anova module is the same as the Q box plot obtained below using the proc sgplot module.



Q-Q plot

The Q box plot shows the data points are distributed for each group. The mean and median are not spread evenly. From the plot, there is a slight difference amongst the mean for each group and a significant difference amongst their median. The group mean of Female_E1 and Male are the same, but their median is different. The outliers of each group are beyond the minimum cutoff point.

Q-Q plot

The UNIVARIATE Procedure
Variable: Scale (Male_E1)

Moments			
N	80	Sum Weights	80
Mean	10.425	Sum Observations	834
Std Deviation	3.33669662	Variance	11.1335443
Skewness	-0.2249118	Kurtosis	-0.6938761
Uncorrected SS	9574	Corrected SS	879.55
Coeff Variation	32.0066822	Std Error Mean	0.37305402

Basic Statistical Measures			
Location		Variability	
Mean	10.42500	Std Deviation	3.33670
Median	11.00000	Variance	11.13354
Mode	11.00000	Range	14.00000
		Interquartile Range	5.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	27.94501	Pr > |t|	<.0001
Sign	M	40	Pr >= |M|	<.0001

nid-usw2.oda.sas.com/SASStudio/sasexec/submissions/9f4bed06-4217-405b-b026-ce6c3ab29b3f/results

PM Results: ASSIGNMENT2.sas

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Signed Rank	S	1620	Pr >= |S|	<.0001

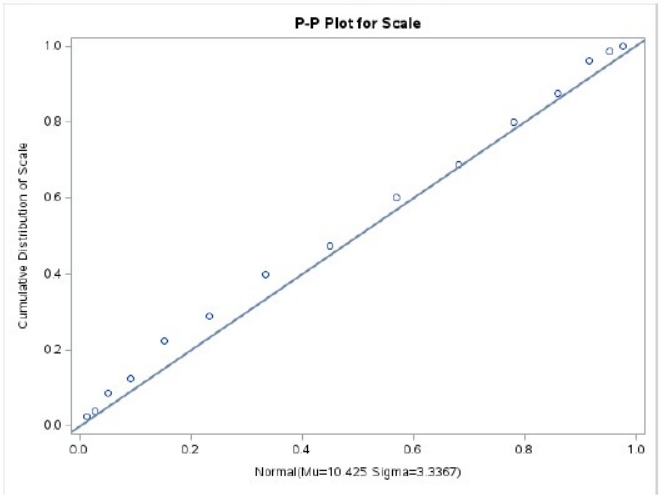
Quantiles (Definition 5)	
Level	Quantile
100% Max	17
99%	17
95%	15
90%	15
75% Q3	13
50% Median	11
25% Q1	8
10%	6
5%	5
1%	3
0% Min	3

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
3	73	15	67
3	64	15	78
4	79	16	4

The skewness and kurtosis values are 0 which indicates a normal distribution. The standard deviation for all groups is 3.34.

3. Conclusion

After much analysis it is inferred that the educational level varies with the gender. These servers as evidence to backup the alternative hypothesis that states; There is an interaction between gender and education in holding jobs.



Obs	Scale	Gender_range
1	10	Male_E1
2	9	Male_E1
3	12	Male_E1
4	16	Male_E1
5	14	Male_E1

Obs	Scale	Gender_range
6	17	Male_E1
7	13	Male_E1
8	9	Male_E1
9	11	Male_E1
10	15	Male_E1
11	12	Male_E2
12	11	Male_E2
13	9	Male_E2
14	14	Male_E2
15	12	Male_E2
16	16	Male_E2
17	10	Male_E2
18	10	Male_E2
19	5	Male_E2
20	11	Male_E2
21	15	Male_E3
22	8	Male_E3
23	7	Male_E3
24	7	Male_E3
25	7	Male_E3
26	9	Male_E3
27	14	Male_E3
28	15	Male_E3
29	11	Male_E3
30	13	Male_E3
31	8	Male_E4
32	9	Male_E4
33	5	Male_E4
34	11	Male_E4
35	13	Male_E4
36	8	Male_E4
37	7	Male_E4
38	11	Male_E4
39	10	Male_E4
40	8	Male_E4
41	7	Female_E1
42	13	Female_E1
43	14	Female_E1
44	6	Female_E1
45	11	Female_E1
46	14	Female_E1