Statistics for Analytics (BAN 100)

Assignment 3

by : Aaron Gonsalves

(161288196)

PROBLEM 1

HYPOTHESIS

Determining whether ownership of stocks varied by age.

Null Hypothesis(H_0):	**There is no difference in mean of stock ownership between the four age groups.**
Alternative Hypothesis(H_a):	**There is a change in at least one of the four age groups regarding stock ownership.**

CODES

```
data stocks;
set work.stocks;
run;
proc print data=stocks;
run;

data agecategory1;
set stocks(obs=84);
agecatg = "Young (under 35)";
stockvalue = young;
drop Young Early_Middle_Age Late_Middle_Age Senior E F G H;
run;
*proc print data=agecategory1;
*run;

data agecategory2;
set stocks(obs=131);
agecatg = "Early middle age (35 to 40)";
stockvalue = Early_Middle_Age;
drop Young Early_Middle_Age Late_Middle_Age Senior E F G H;
run;
*proc print data=agecategory2;
*run;
```

```
data agecategory3;
set stocks(obs=93);
agecatg = "Late middle age (50 to 65)";
stockvalue = Late_Middle_Age;
drop Young Early_Middle_Age Late_Middle_Age Senior E F G H;
run;
*proc print data=agecategory3;
*run;

data agecategory4;
set stocks(obs=58);
agecatg = "Senior (over 65)";
stockvalue = Senior;
drop Young Early_Middle_Age Late_Middle_Age Senior E F G H;
run;
*proc print data=agecategory4;
*run;


data stocksbyage;
set agecategory1 agecategory2 agecategory3 agecategory4;
run;
*proc print data=stocksbyage;
*run;

title 'Anova calculations :';
proc anova data=stocksbyage;
class agecatg;
model stockvalue = agecatg;
run;
```

Anova calculations :

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
agecatg	4	Early middle age Late middle age Senior (over 65) Young (under 35)

Number of Observations Read	366
Number of Observations Used	366

Anova calculations :

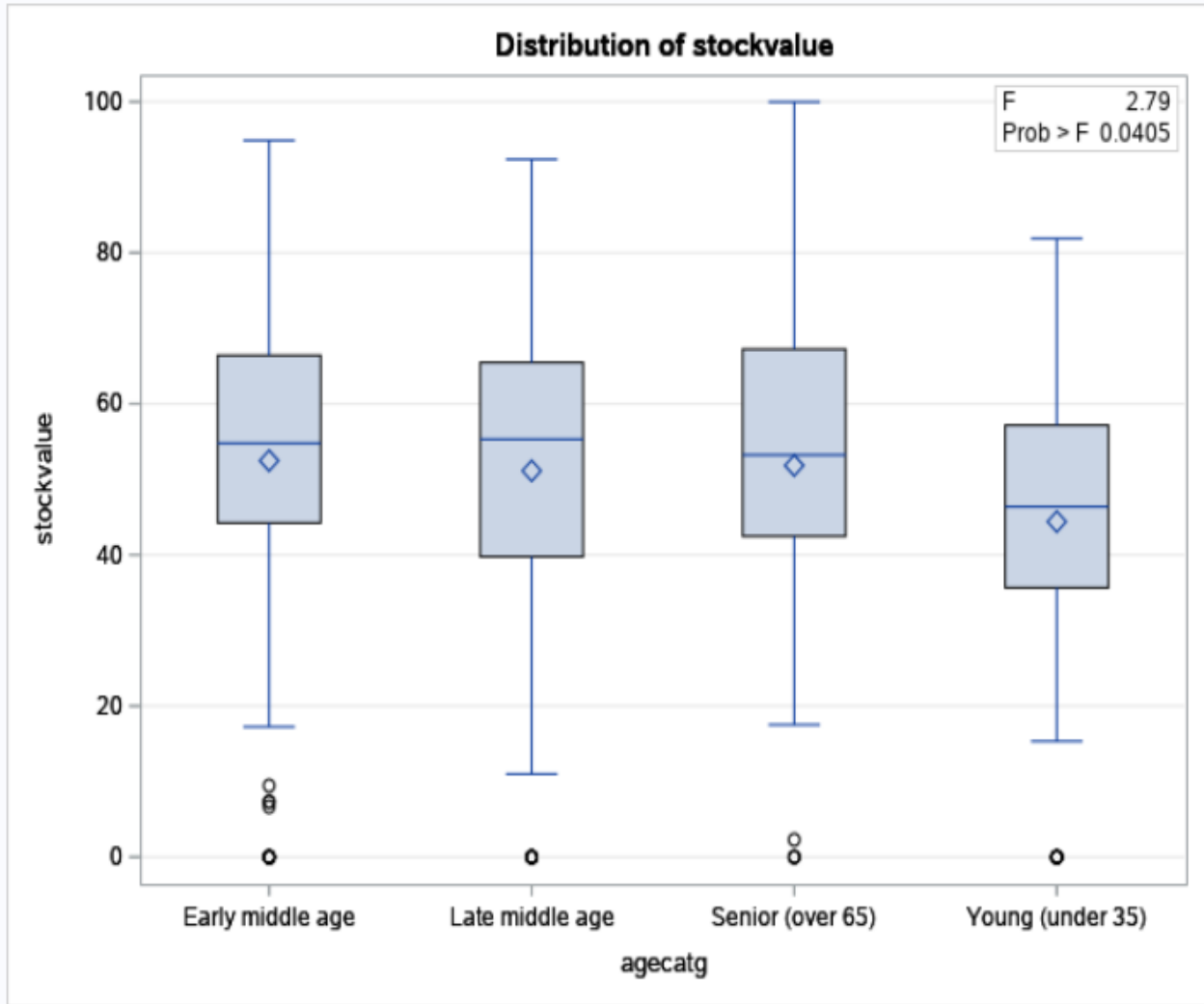
The ANOVA Procedure

Dependent Variable: stockvalue

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	stockvalue Mean
0.022591	42.14046	21.14610	50.18003

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



Interpretation

- There is a 5% significance level, and P-value from Anova test is 0.0405.

- As P-value<0.05, Null hypothesis (H_0) is rejected i.e. no difference between mean stock value between different age groups, and there is much evidence in support of Alternative Hypothesis(H_a).

Conclusion

- As Null hypothesis (H_0) is rejected from our interpretation, the conclusion is in support of Alternative Hypothesis(H_a).

- This deduces that there is a change in at least one of the four age groups regarding stock ownership.

PROBLEM 2

CODES TO CONVERT DATA INTO PROPER FORMAT

```
data jobs;
set work.jobs;
run;
proc print data=jobs;
run;
```

CODES TO CONVERT DATA INTO PROPER FORMAT

```
data jobs_m1;
set jobs(obs=10);
edu_level = "Less than high school (E1)";
jobs_num = Male_E1;
Gender = "Male";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;

data jobs_m2;
set jobs(obs=10);
edu_level = "High school (E2)";
jobs_num = Male_E2;
Gender = "Male";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;

data jobs_m3;
set jobs(obs=10);
edu_level = "Some college/university but not degree (E3)";
jobs_num = Male_E3;
Gender = "Male";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;

data jobs_m4;
set jobs(obs=10);
edu_level = "At least one university (E4)";
jobs_num = Male_E4;
Gender = "Male";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;
```

```
data jobs_f1;
set jobs(obs=10);
edu_level = "Less than high school (E1)";
jobs_num = Female_E1;
Gender = "Female";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;

data jobs_f2;
set jobs(obs=10);
edu_level = "High school (E2)";
jobs_num = Female_E2;
Gender = "Female";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;

data jobs_f3;
set jobs(obs=10);
edu_level = "Some college/university but not degree (E3)";
jobs_num = Female_E3;
Gender = "Female";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;

data jobs_f4;
set jobs(obs=10);
edu_level = "At least one university (E4)";
jobs_num = Female_E4;
Gender = "Female";
drop Male_E1 Male_E2 Male_E3 Male_E4 Female_E1 Female_E2 Female_E3 Female_E4;
run;


title 'Education by gender:';
data educationbygender;
set jobs_m1 jobs_m2 jobs_m3 jobs_m4 jobs_f1 jobs_f2 jobs_f3 jobs_f4;
proc print data=educationbygender;
run;
```

HYPOTHESIS 1

A. Test to determine whether there is interaction between gender and education in holding jobs.

Null Hypothesis(H_0):	**There is no interaction between gender and education in holding jobs.**
Alternative Hypothesis(H_a):	**There is interaction between gender and education in holding jobs.**

2- WAY ANOVA STATISTICS

```
proc glm data=educationbygender;
class edu_level Gender;
model jobs_num = edu_level | Gender;
run;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
edu_level	4	At least one university (E High school (E2) Less than high school (E1) Some college/university bu
Gender	2	Fema Male

Number of Observations Read	80
Number of Observations Used	80

The GLM Procedure

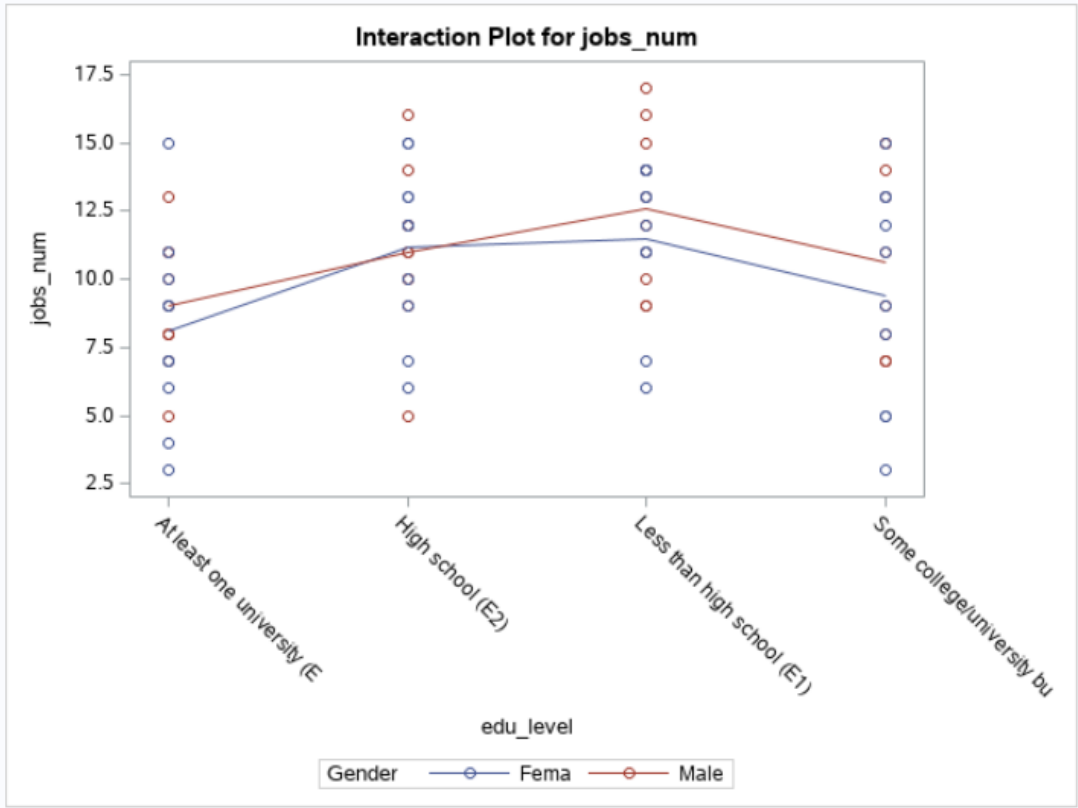
Dependent Variable: jobs_num

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	jobs_num Mean
0.174351	30.46392	3.175864	10.42500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
edu_level	3	135.8500000	45.2833333	4.49	0.0060
Gender	1	11.2500000	11.2500000	1.12	0.2944
edu_level*Gender	3	6.2500000	2.0833333	0.21	0.8915

Source	DF	Type III SS	Mean Square	F Value	Pr > F
edu_level	3	135.8500000	45.2833333	4.49	0.0060
Gender	1	11.2500000	11.2500000	1.12	0.2944
edu_level*Gender	3	6.2500000	2.0833333	0.21	0.8915



Interpretation

- From the 4th table of the GLM procedure, we check the p value for edu_level*Gender.
- The level of significance is 5%.
- The p value is 0.8915, which is significant as p-value>0.05, implying that null hypothesis (H_0) should not be rejected.

Conclusion

- Based on the above interpretation our conclusion is in support of Null hypothesis (H_0).

- This deduces that there is no interaction between gender and education in holding jobs.

HYPOTHESIS 2

B. Test to determine whether there are differences in holding jobs between men and women.

Null Hypothesis(H_0):	**The means of men and women are equal in holding jobs.**
Alternative Hypothesis(H_a):	**The means of men and women are not equal in holding jobs.**

2- WAY ANOVA STATISTICS

```
proc glm data=educationbygender;
class edu_level Gender;
model jobs_num = edu_level | Gender;
run;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
edu_level	4	At least one university (E High school (E2) Less than high school (E1) Some college/university bu
Gender	2	Fema Male

Number of Observations Read	80
Number of Observations Used	80

The GLM Procedure

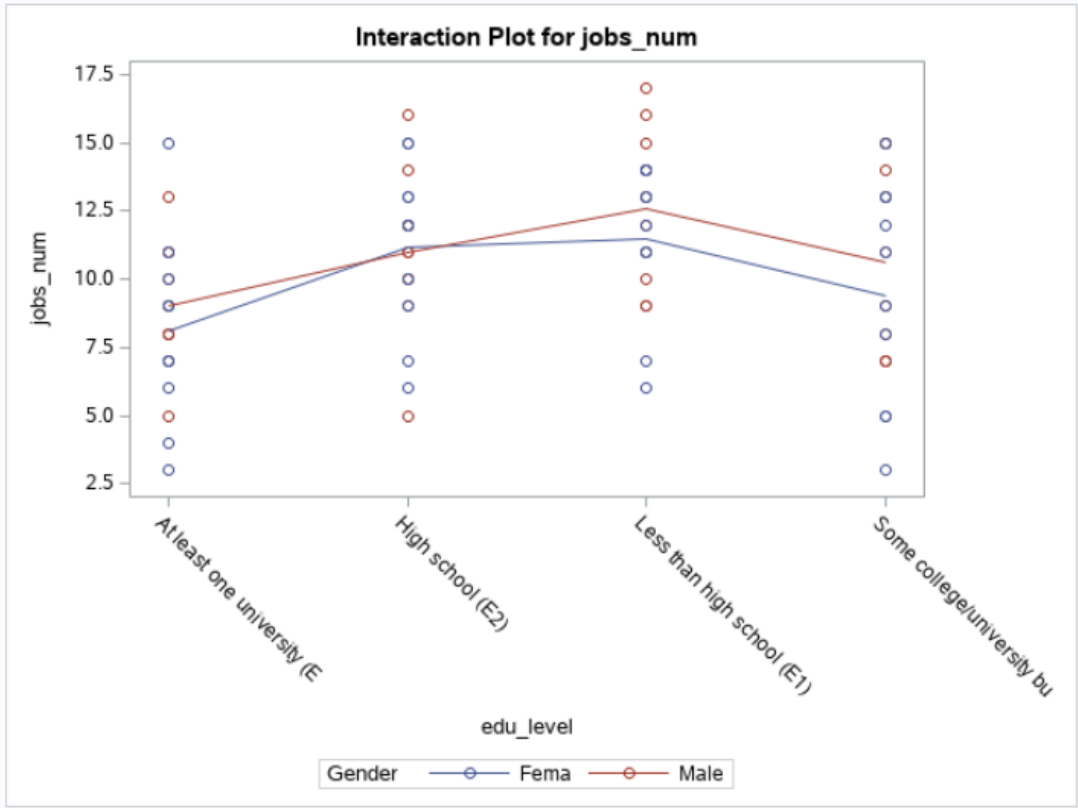
Dependent Variable: jobs_num

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	jobs_num Mean
0.174351	30.46392	3.175864	10.42500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
edu_level	3	135.8500000	45.2833333	4.49	0.0060
Gender	1	11.2500000	11.2500000	1.12	0.2944
edu_level*Gender	3	6.2500000	2.0833333	0.21	0.8915

Source	DF	Type III SS	Mean Square	F Value	Pr > F
edu_level	3	135.8500000	45.2833333	4.49	0.0060
Gender	1	11.2500000	11.2500000	1.12	0.2944
edu_level*Gender	3	6.2500000	2.0833333	0.21	0.8915



Interpretation

- From the 4th table of the GLM procedure, we check the p value for Gender.
- The level of significance is 5%.
- The p value is 0.2944, which is significant as p-value>0.05, implying that null hypothesis (H_0) should not be rejected.

Conclusion

- Based on the above interpretation our conclusion is in support of Null hypothesis (H_0).

- This deduces that there are no differences in holding jobs between men and women as gender does not matter.

HYPOTHESIS 3

C. Test to determine whether there are differences in holding jobs between the educational levels.

Null Hypothesis(H_0):	**The means of different educational levels are equal in holding jobs.**
Alternative Hypothesis(H_a):	**The mean of at least one educational level is different in holding jobs.**

2- WAY ANOVA STATISTICS

```
proc glm data=educationbygender;
class edu_level Gender;
model jobs_num = edu_level | Gender;
run;
```

The GLM Procedure

Class Level Information		
Class	Levels	Values
edu_level	4	At least one university (E High school (E2) Less than high school (E1) Some college/university bu
Gender	2	Fema Male

Number of Observations Read	80
Number of Observations Used	80

The GLM Procedure

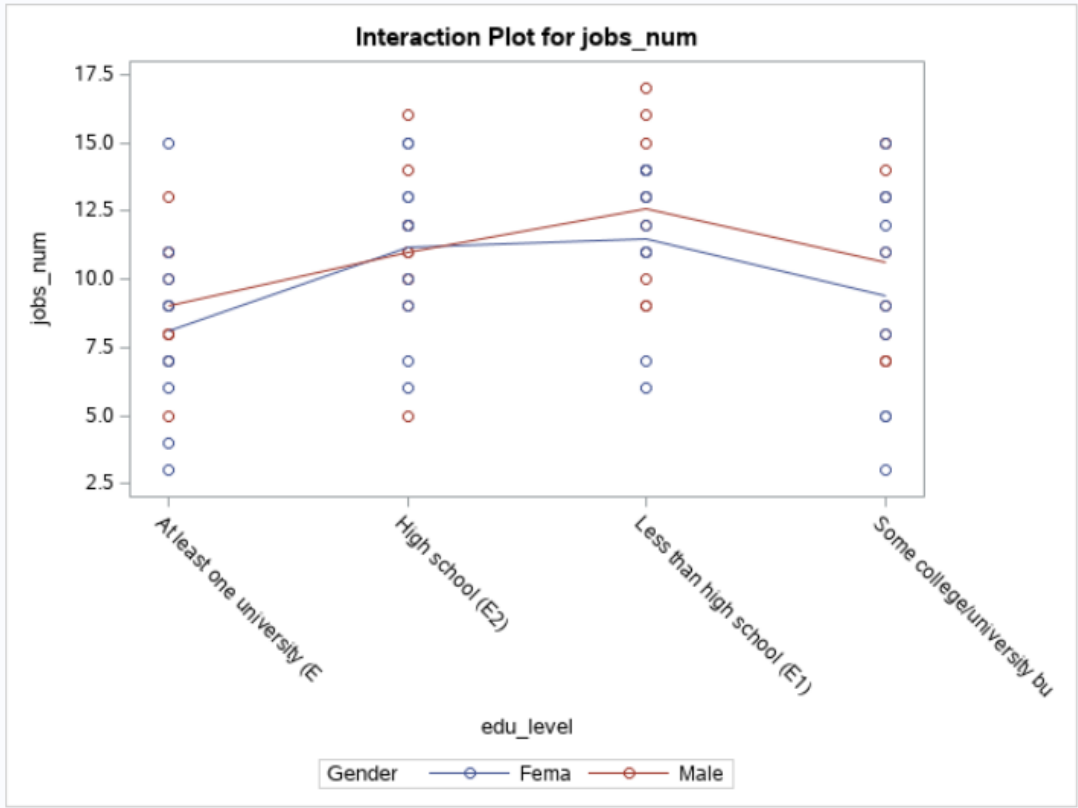
Dependent Variable: jobs_num

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	jobs_num Mean
0.174351	30.46392	3.175864	10.42500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
edu_level	3	135.8500000	45.2833333	4.49	0.0060
Gender	1	11.2500000	11.2500000	1.12	0.2944
edu_level*Gender	3	6.2500000	2.0833333	0.21	0.8915

Source	DF	Type III SS	Mean Square	F Value	Pr > F
edu_level	3	135.8500000	45.2833333	4.49	0.0060
Gender	1	11.2500000	11.2500000	1.12	0.2944
edu_level*Gender	3	6.2500000	2.0833333	0.21	0.8915



Interpretation

- From the 4th table of the GLM procedure, we check the p value for edu_level.
- The level of significance is 5%.
- The p value is 0.0060, which is not significant as p-value<0.05, implying that null hypothesis (H_0) should be rejected.
- Alternative hypothesis should be considered.

Conclusion

- Based on the above interpretation our conclusion is in support of alternative hypothesis (H_a).

- This deduces that there are differences in holding jobs between the education levels.

THANK YOU