Assignment 3 (ANOVA)

Problem 1 - Use 5% as a significance level

In the last decade, stockbrokers have drastically changed the way they do business. Internet trading has become quite common and online trades can cost as little as $7. It is now easier and cheaper to invest in the stock market than ever before. What are the effects of these changes? To help answer this question, a financial analyst randomly sampled 366 American households and asked each to report the age of the head of the household and the proportion of their financial assets that are invested in the stock market. The age categories are:

Young (under 35)
Early middle age (35 to 40)
Late middle age (50 to 65)
Senior (over 65)

The analyst was particularly interested in determining whether the ownership of stocks varied by age. Do these data allow the analyst to determine that there are differences in stock ownership between the four age groups? Check the required conditions.

Problem 2

One measure of the health of a national economy is how quickly it creates jobs. One aspect of this issue is the number of jobs individuals hold. As part of a study on job tenure, a survey was conducted wherein Americans aged between 17 and 45 were asked how many jobs they have held in their lifetimes. Also recorded were gender and educational attainment. The categories are:

Less than high school (E1)
High school (E2)
Some college/university but not degree (E3)
At least one university (E4)

a. Test to determine whether there is an interaction between gender and education in holding jobs.
b. Test to determine whether there are differences in holding jobs between men and women.
c. Test to determine whether there are differences in holding jobs between the educational levels.

Problem 1

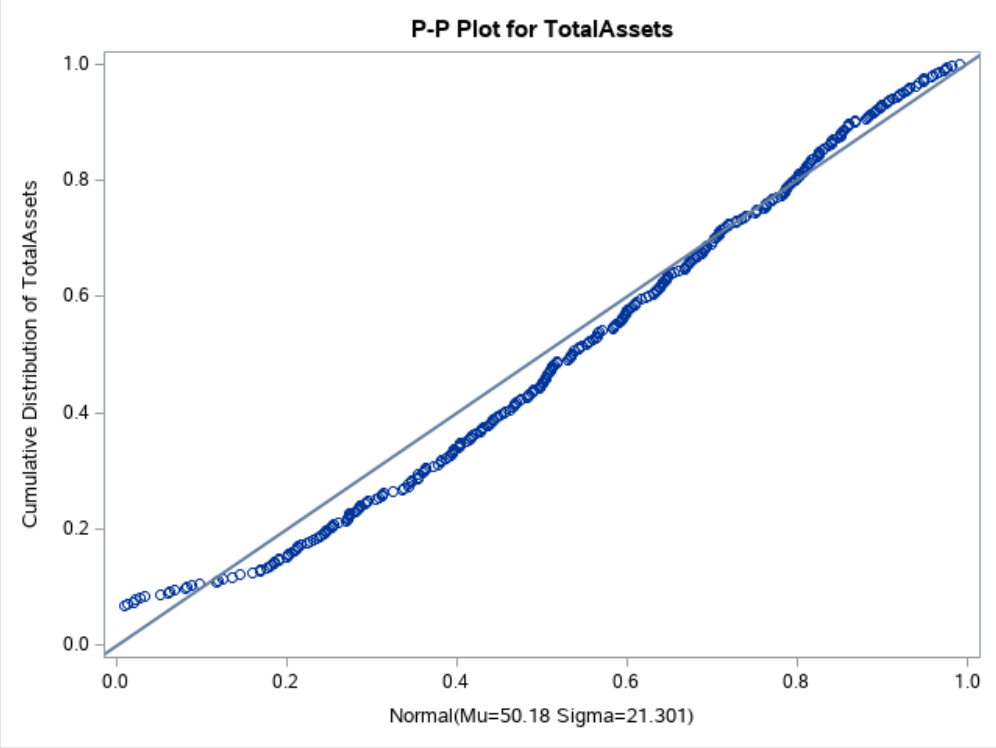
In this analysis, we will look into the **Total Assets Invested** dataset that contains **366 records** of both the age of the household head and the proportion of their financial assets that are invested in the stock market. age was distributed into 4 main categories:

Young (under 35)
Early middle age (35 to 40)
Late middle age (50 to 65)
Senior (over 65)

This Analysis aims to determine whether stock ownership varied by age and to answer that question the following hypothesis was tested at a 95% significance level.

H_0 : $\mu_{Young} = \mu_{Early\ Middle\ Age} = \mu_{Late\ Middle\ Age} = \mu_{Senior}$
H_a : Not all means are equal

First, the normal probability plot showed that the data was distributed fairly in a straight line which meant it is normally distributed with a mean of 50.2, a Standard deviation of 21.3, and a variance of 453.7



Second, an ANOVA test was conducted on the dataset and the following tables were generated.

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	TotalAssets Mean
0.022591	42.14046	21.14610	50.18003

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

The Model (SSA) had 3 degrees of freedom while the Error (SSW) had 362 and the F value of 2.79 at that degree of freedom returns a P value of 0.0405 which is less than α (0.05) therefore the null hypothesis is rejected and the decision is that there is a statistical difference in the ownership of stocks based on the age group of the head of the household.

AgeGroup	TotalAssets LSMEAN	LSMEAN Number
Early_Middle_Age	52.4724427	1
Late_Middle_Age	51.1390323	2
Senior	51.8381034	3
Young	44.3983333	4

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

Dependent Variable: TotalAssets

i/j	1	2	3	4
1		0.9666	0.9976	0.0333
2	0.9666		0.9973	0.1494
3	0.9976	0.9973		0.1681
4	0.0333	0.1494	0.1681	

After the null hypothesis was rejected, we look at the means In order to determine which age group is different from the others and it seems like it is the age group Young, but in order to prove that statistically, we use post hoc test for Tukey, Where it shows that the only significant difference in the means is between Young age group (number 4) and Early Middle age (number 1). Where P value is 0.0333 which is less than α (0.05) and we reject the null hypothesis

$H_0 : \mu_{Young} = \mu_{Early\ Middle\ Age}$

$H_a : \mu_{Young} \neq \mu_{Early\ Middle\ Age}$

Problem 2

In this report, we will be analyzing the Lifetime of Jobs by Educational level dataset. It contains 80 records of the Gender, Educational level, and the number of jobs held in the participant's lifetime, the report aims to answer three questions.

Is there an interaction between gender and education in holding jobs?

H_0 : An interaction is absent
H_a : An interaction is present

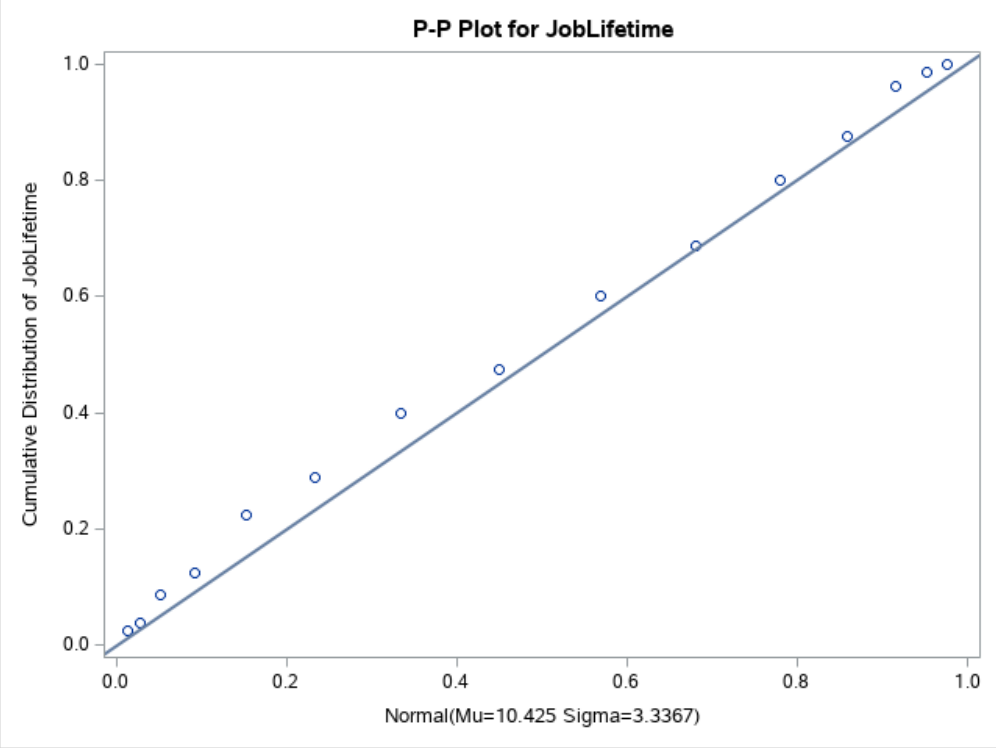
Are there any differences in holding jobs between men and women?

$H_0 : \mu_{Men} = \mu_{Women}$
$H_a : \mu_{Men} \neq \mu_{Women}$

Are there any differences in holding jobs between the educational levels?

$H_0 : \mu_{E1} = \mu_{E2} = \mu_{E3} = \mu_{E4}$
H_a : Not all means are equal

First, we used the P-P plot to determine if the data is normally distributed and the plot showed that the data followed the theoretical normal distribution straight line fairly which means we can assume a normal distribution.

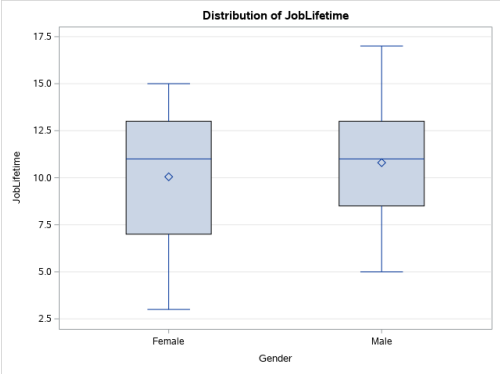


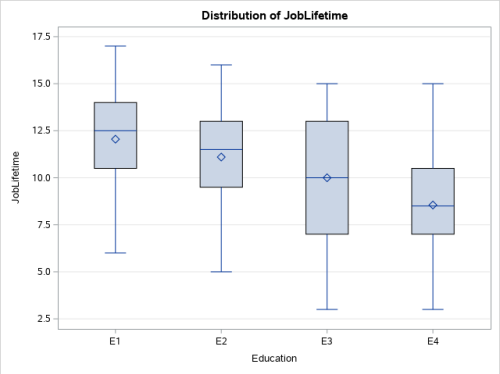
Level of Gender	N	JobLifetime	
		Mean	Std Dev
Female	40	10.0500000	3.57304725
Male	40	10.8000000	3.08179102

The means of both genders are close to each other as well as the standard deviation.

Level of Education	N	JobLifetime	
		Mean	Std Dev
E1	20	12.0500000	2.85574214
E2	20	11.1000000	2.95403382
E3	20	10.0000000	3.69921756
E4	20	8.5500000	2.92853475

For the Education level on the other hand it is clear that there are some differences between the means and at least E3 has a higher standard deviation than the rest of the education levels.





To further analyze these findings we need to conduct an ANOVA test, and the test generated the following results, the model had 7 degrees of freedom and the Error had 72 degrees of freedom.

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	JobLifetime Mean
0.174351	30.46392	3.175864	10.42500

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

For the first question, the P value is 0.8915 which is higher than α (0.05) therefore we fail to reject the null hypothesis.

In the second question, the P value is 0.2944 which is also higher than α (0.05) therefore we fail to reject the null hypothesis.

For both previous variables, there is no statistical difference between the two genders or if there is statistical evidence of an interaction between the variables.

As for the last question, the P value is 0.0060 which is lower than α (0.05) therefore we reject the null hypothesis and decide that there is a statistical difference between education levels in terms of jobs held in a lifetime.

After rejecting the null hypothesis, a deeper analysis of Education levels was needed to determine which levels had unequal means to the others, thus we used the Tukey method.

Means with the same letter are not significantly different.

Tukey Grouping		Mean	N	Education
	A	12.050	20	E1
	A			
B	A	11.100	20	E2
B	A			
B	A	10.000	20	E3
B				
B		8.550	20	E4

Education	JobLifetime LSMEAN	LSMEAN Number
E1	12.0500000	1
E2	11.1000000	2
E3	10.0000000	3
E4	8.5500000	4

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

Dependent Variable: JobLifetime

i/j	1	2	3	4
1		0.7722	0.1715	0.0038
2	0.7722		0.6833	0.0564
3	0.1715	0.6833		0.4630
4	0.0038	0.0564	0.4630	

E1 and E4 were grouped in different groups while E2 and E3 once were grouped with E1 in group A and E4 in group B. This means that E1 and E4 have different means. And that difference is shown in the Tukey test where the P value was 0.0038 which is less than α (0.05) and we reject the null hypothesis

$H_0 : \mu_{E1} = \mu_{E4}$

$H_a : \mu_{E1} \neq \mu_{E4}$

This means that the mean of jobs held in a lifetime in Education level 4 is different from the mean of other education levels.