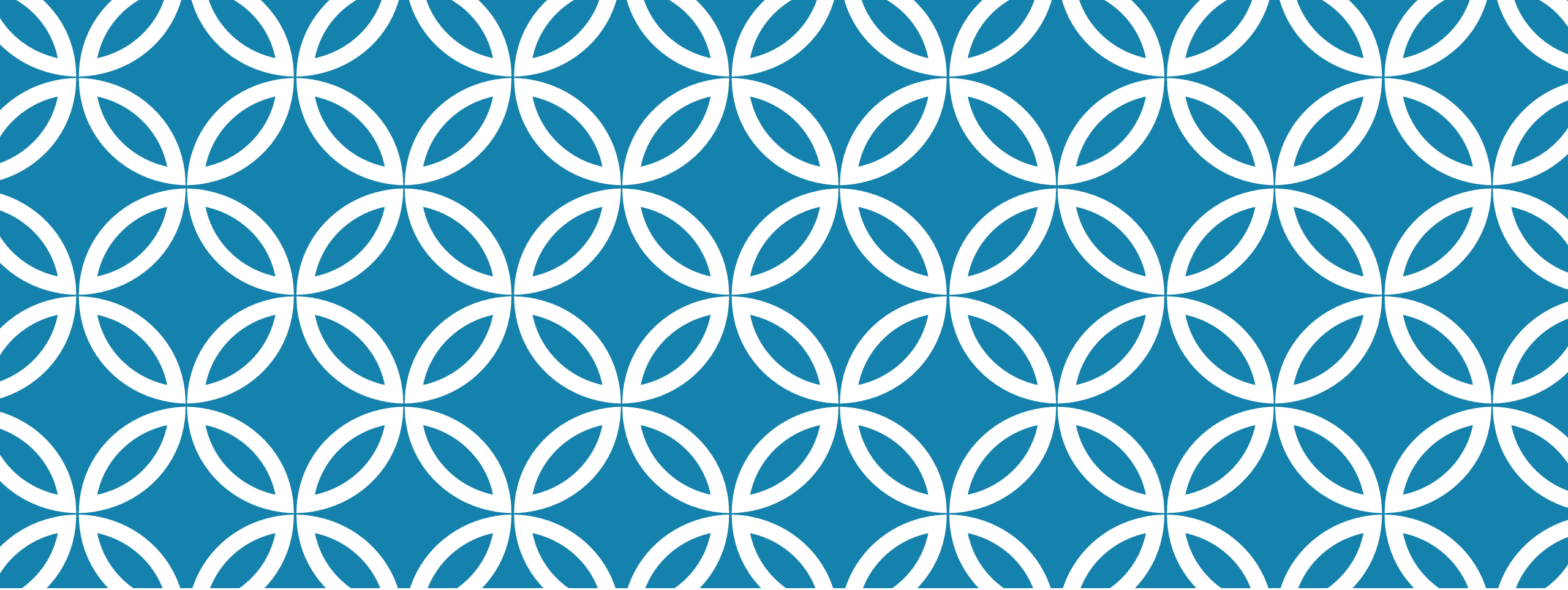


ANOVA ANALYSIS

(DESCRIPTIVE STATISTIC ANALYSIS)

Maryam Najimigoshtasb

One-Way ANOVA

One way Anova is the analytics model between one categorical independent variable and one numerical dependent variable in order to check the significant influences of independent variable on the dependent variable. In fact, it is mean to checking the variation of samples or levels around the population if the variation is less then it means the samples support the populations and have a significant influence on the population . To illustrate more, we can refer to the following **problem 1.**

Wherever we need to check the variations of samples with the population we can use this test , for example in linear regression each level of the categorical variable plays one dimension role and our dependent variable another dimension. Suppose we have one categorical variable with two levels and one dependent continues variable. Thus, we can have lots of cube coming from the all the these three vectors in 3 dimensional space, as this variable are the dummy variable then then cubes coming from each level can be categorized as the same rule of their associated level. we are looking for the best fit that can have the minimum variation with all these cubes. In fact , The best fit is average mean of all cubes is average mean of population if the variation is less our fit is perfect it means our variation among group should be high as a result the variation with in group is less which result in normality.

Business Question

What are the effects of stock changes? Does the ownership of stocks varied by age.

Hypothesis:

Null Hypothesis: There are no variation between means of the samples
Alternative Hypothesis: There is at least one mean among the samples varies from others.

$$\begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1: \exists \mu_i \neq \mu_j \quad j, i \in \{1,2,3,4\} \end{cases}$$

ASSUMPTIONS:

Before conducting the test we need to check three assumptions in the one way ANOVA.

1)-Residuals (experimental error) are approximately normally distributed
2)-Homogeneity of variances (variances are equal between groups)
3)-Observations are sampled independently from each other (no relation in observations between the groups and within the groups)

It is important to note that ANOVA is not robust to violations to the assumption of independence. This is to say, that even if you violate the assumptions of homogeneity or normality, you can conduct the test and basically trust the findings. However, the results of the ANOVA are invalid if the independence assumption is violated. In general, with violations of homogeneity the analysis is considered robust if you have equal sized groups. With violations of normality, continuing with the ANOVA is generally ok if you have a large sample size.

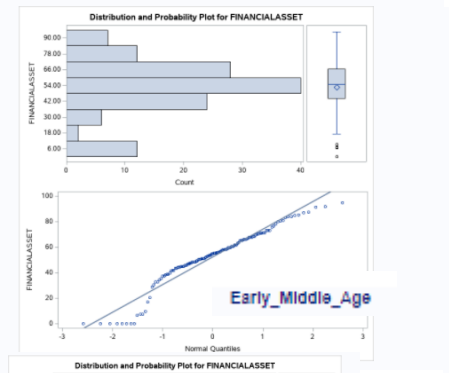
Do these data allow the analyst to determine that there are differences in stock ownership between the four age group?
To answer this question we need to check the assumptions.
Here we can just check the first two assumptions.

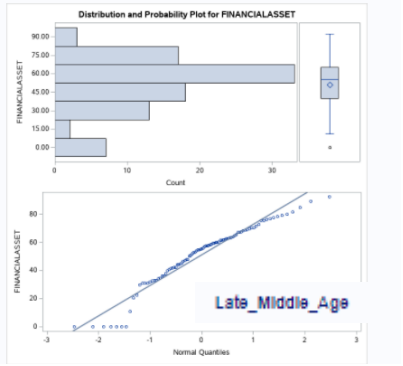
ANOVA ASSUMPTIONS:

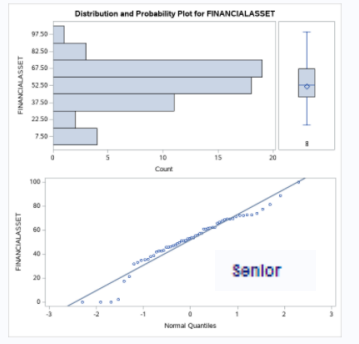
1)Normality of the residual

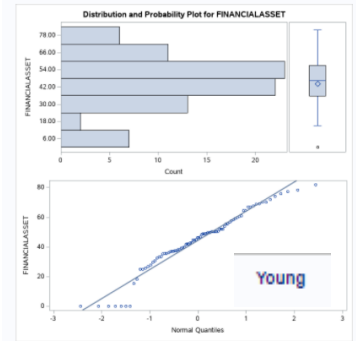
Looking at the histogram of residuals we can see the residual are normally distributes and the p-value of the tests are proving it.

All these result could be conducted by looking at the normal distribution of each groups



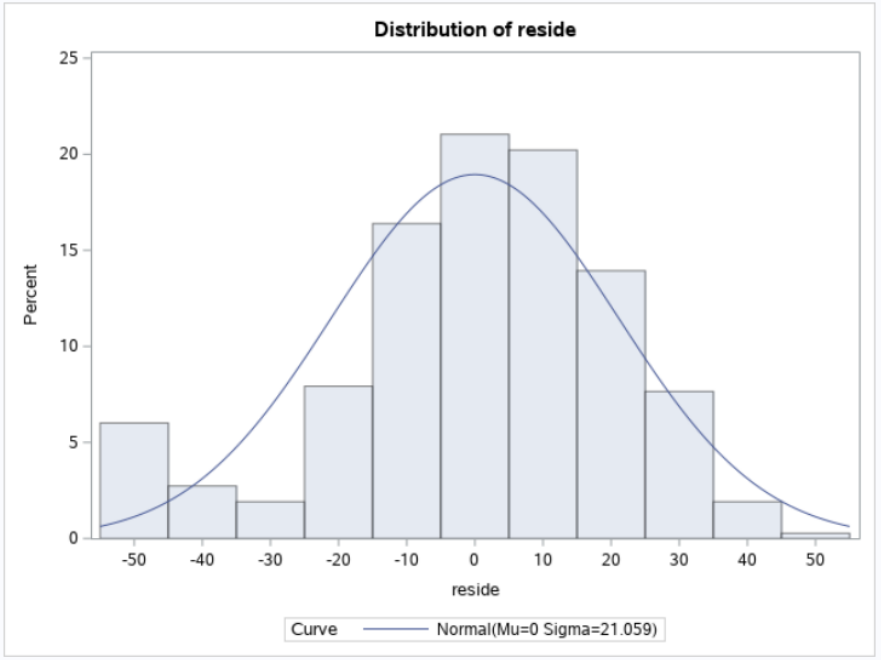







```
proc GLM data=work.stock;
class agegroup ;
model financialasset=agegroup;
output out=noresidual
 r= reside;
means agegroup /alpha=0.005;
run;

 proc univariate data=work.noresidual;
 var reside;
 histogram/normal;
 run;
```



Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.08674359	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.62666466	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4.88741730	Pr > A-Sq	<0.005

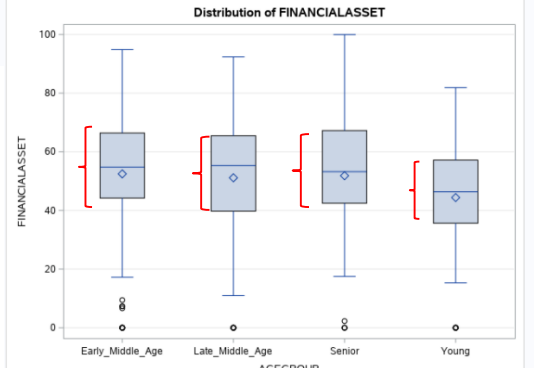
2)- check the equality of variance between and among groups

Hypothesis on variance of the samples :
Null hypothesis: there are equal variance between and among groups
Alternative Hypothesis: at least one is different.

Looking at the p-value we fail to reject the null hypothesis.
It means that variances are homogenate
Even with length of boxplot we can see all are almost equal.

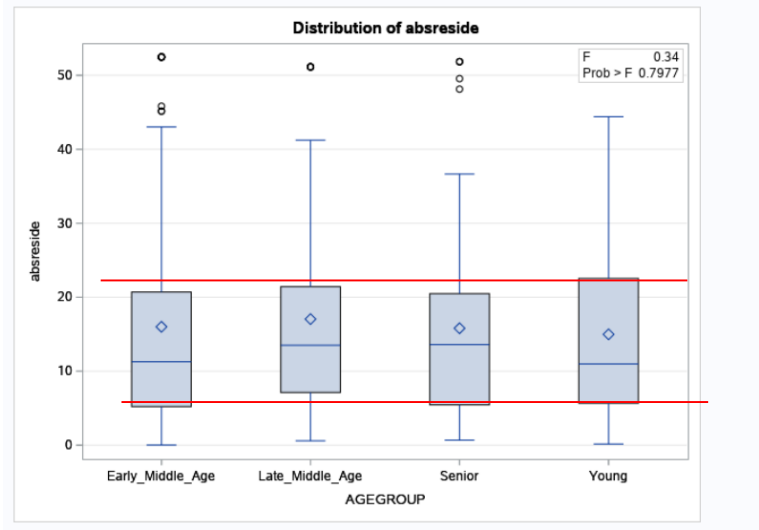
We also can conduct this from comparing the variance of the levels we can see they are almost belong the same range.

Level of AGEGROUP	N	FINANCIALASSET	
		Mean	Std Dev
Early_Middle_Age	131	52.4724427	21.6664980
Late_Middle_Age	93	51.1390323	21.7215074
Senior	58	51.8381034	21.0900334
Young	84	44.3983333	19.6607843



Dependent Variable: absreside

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	190.49067	63.49689	0.34	0.7977
Error	362	67953.65073	187.71727		
Corrected Total	365	68144.14139			




```
*check the variance equality;
data eqvar;
set work.noresidual;
absreside=abs(reside);
run;

proc glm data=work.eqvar;
class agegroup;
model absreside=agegroup;
run;
```

Do these data allow the analyst to determine that there are differences in stock ownership between the four age group?
After checking the assumptions ,except the independency the samples (which is more important that others but considering that is it true) **we can run the ANOVA test** on the data to check the differences in the sock ownership by age group.

Dependent Variable: FINANCIALASSET FINANCIALASSET

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3741.3638	1247.1212	2.79	0.0405
Error	362	161870.9817	447.1574		
Corrected Total	365	165612.3453			

ANOVA RESULT :
We know more close the P-value to zero more likely we reject the null hypothesis or more big F-test more likely we reject null hypothesis. Therefore; we can see F is not significant enough but respect to alpha=0.05 we can reject the null hypothesis, but this model is not a good model.

To sum up; at least one group has different mean from others and has more impact on stock or dominate the stock more . To see the contribution we can refer to linear regression and we can see **Early middle age** has more contribution in stock compare to others.
Comparing the means we can also conclude that.

	Coefficients
Intercept	44.39833333
Early_Middle_Age	8.074109415
Late_Middle_Age	6.740698925
Senior	7.439770115

Two-Way ANOVA

Our Analysis purpose is to measure the health of a national economy. To measure it we need to measure how quickly it creates jobs. To measure that we measure how many job has been held by individual which might be different as their educational level and their gender are different.
Therefore , our national economy is healthy if the average number of jobs held by individuals in their equational and gender category be high compare to global average number of jobs held by population; but it should be close to the average number of jobs held by individual in their group. To see that we need to check the variability of jobs in each group and and variability of jobs between groups comparing to the global average number of jobs in society.

If the variation of average number of jobs in groups be higher than average number of job in society, while the number of job held by individual have less variation in their group then it means people were able to find and change job easily in their life . So our national economy is healthy.

Here, group is a combined educational level and gender terms. Therefore, we need to check the influence of these two terms on number of jobs held by individual and check the variance of means in each group using two way ANOVA. we will check the influence of gender and educational level and their interactions on number of jobs. Thus, we will see which one has the most impact. Before that we will check the assumptions

Hypothesis one :
Null Hypothesis: Gender and educational level does not have effect on number of jobs held by individual.
Alternative Hypothesis: At least one of them has impact on number of jobs.

Hypothesis two (educational level):
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_1: $\exists \mu_i \neq \mu_j \quad j, i \in \{1,2,3,4\}$

Hypothesis three(Gender):
H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

Hypothesis four (interaction):
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_1: $\exists \mu_i \neq \mu_j \quad j, i \in \{1,2,3,4\}$

ASSUMPTIONS:

Before conducting the test we one to check three assumptions in the one way ANOVA.

1)-Residuals (experimental error) are approximately normally distributed
2)-Homogeneity of variances (variances are equal between groups)
3)-Dependent variable should be continuous – that is, measured on a scale
4)-Two independent should be in categorical, independent groups.
5)-Sample independence – that each sample has been drawn independently of the other samples

We will check the 1st and 2nd, As the last three are met.

Hypothesis on variance of the samples :

Null hypothesis: there are equal variance between and among groups
Alternative Hypothesis: at least one is different.

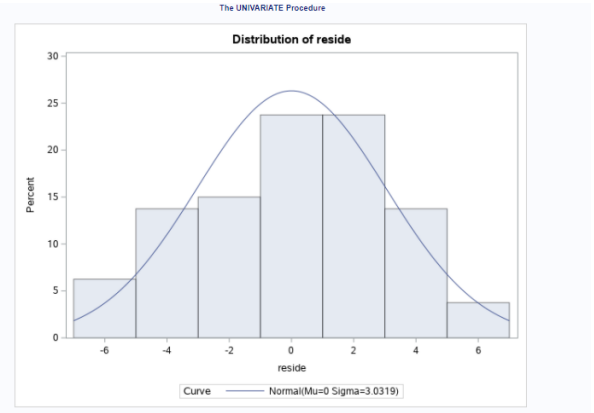
Looking at the p-value we fail to reject the null hypothesis.
It means that variances are homogenate

Dependent Variable: absreside

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	16.2580000	4.0645000	1.47	0.2189
Error	75	206.9375000	2.7591667		
Corrected Total	79	223.1955000			

Normality of the residual

Looking at the histogram of residuals we can see the residual are normally distributes and the p-value of the tests are proving it.



The result of Two-Way Anova:

However the F-statistic in not very good but respect to our threshold is good enough to reject the null hypothesis. The higher F-value the lower probability that null hypothesis be true.

So here we reject the null hypothesis and tend to say at least one of them has impact on number of jobs.

We will check other hypothesis to see what is the influence of individual terms in the model. But we can not speak about the amount of contributions in the model, but we can speak about the relative statistic significance of them.

Dependent Variable: Nu_jobs Nu_jobs

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			

Hypothesis two (educational level):

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: $\exists \mu_i \neq \mu_j \quad j, i \in \{1,2,3,4\}$

P-value is greater than our alpha we fail to reject the null hypothesis so gender compare to educational level has not significant effect on number of jobs

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

Hypothesis three(Gender):

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

P-value is greater less than our alpha we reject the null hypothesis so educational level compare gender to has a significant effect on number of jobs

Hypothesis four (interaction):

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: $\exists \mu_i \neq \mu_j \quad j, i \in \{1,2,3,4\}$

P-value is greater than our alpha we fail to reject the null hypothesis so there is not significant interaction between the gender and educational level. The interaction is absolutely low. The interaction plot also demonstrate it.