Chapter V

One-Way ANOVA

Solutions

Part I:

1.

Step 1 (Formulation of the Hypotheses):

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$; $H_1: \mu_i \neq \mu_j$, for at least one $i \neq j$, $i, j = 1,2,3,4$.

Step 2

Working Table

i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	$(x_{i1} - \bar{x}_1)^2$	$(x_{i2} - \bar{x}_2)^2$	$(x_{i3} - \bar{x}_3)^2$	$(x_{i4} - \bar{x}_4)^2$
1	75	59	65	76	30.25	361	196	210.25
2	83	75	70	60	182.25	9	81	2.25
3	68	100	97	52	2.25	484	324	90.25
4	52		90	58	306.25		121	12.25
5			73				36	
Summe	278	234	395	246	521	854	758	315
$\bar{x}_j$	69.5	78.0	79.0	61.5				
s_j^2	173.67	427.00	189.50	105.00				

$$\bar{\bar{x}} = \frac{69.5 + 48.0 + 79.0 + 61.5}{4} = 72$$

$$SSTR = 4 \cdot (69.5 - 72)^2 + 3 \cdot (78.0 - 72)^2 + 5 \cdot (79.0 - 72)^2 + 4 \cdot (61.5 - 72)^2 = 819.0$$

$$MSTR = \frac{819}{4 - 1} = 273.$$

$$SSE = (4 - 1) \cdot 173.67 + (3 - 1) \cdot 427.00 + (5 - 1) \cdot 189.50 + (4 - 1) \cdot 105.00 = 2448.00$$

$$MSE = \frac{2448}{16 - 4} = 204$$

$$F = \frac{273}{204} = 1.3382 .$$

Step 3:

$\alpha = 0.01.$

Numerator degree of freedom: $k - 1 = 4 - 1,$
Denominator degree of freedom: $n_T - k = 16 - 4 = 12 .$

$$F_{0.01}(df_1 = 3; df_2 = 12) = 5.95 .$$

Because of

$$F = 1.3382 < 5.95$$

the null hypothesis will not be rejected.

2.

Step 1:

$H_0 : \mu_1 = \mu_2 = \mu_3,$ $H_1 : \mu_i \neq \mu_j$ for at least one $i \neq j,\ i, j = 1, 2, 3$

Step 2:

i	x_{i1}	x_{i2}	x_{i3}	$(x_{i1} - \bar{x}_1)^2$	$(x_{i2} - \bar{x}_2)^2$	$(x_{i3} - \bar{x}_3)^2$
1	643	469	484	560.27	21.81	1344.69
2	655	427	456	136.19	2178.09	75.17
3	702	525	402	1248.21	2634.77	2054.81
Summe	2000	1421	1342	1944.67	4834.67	3474.67
$\bar{x}_j$	666.67	473.67	447.33			
s_j^2	972.33	2417.33	1737.33			

$$\bar{\bar{x}} = \frac{666.67 + 473.67 + 447.33}{3} = 529.22$$

$$SSTR = 3 \cdot (666.67 - 529.22)^2 + 3 \cdot (473.67 - 529.22)^2 + 3 \cdot (447.33 - 529.22)^2 = 86052.83$$

$$MSTR = \frac{86052.83}{3 - 1} = 43026.415.$$

$$SSE = (3-1)\cdot 972.33 + (3-1)\cdot 2417.33 + (3-1)\cdot 1737.33 = 10253.98$$

$$MSE = \frac{10253.98}{9-3} = 1709.00$$

$$F = \frac{43026.415}{1709.00} \approx 25.176.$$

Step 3:

$\alpha = 0.05$

Numerator degree of freedom: $k - 1 = 3 - 1 = 2,$
Denominator degree of freedom: $n_T - k = 9 - 3 = 6.$

$$F_{0.05,crit}(df_1 = 2; df_2 = 6) = 5.14.$$

$$F \approx 25.175 > 5.14 = F_{crit}.$$

Thus, we reject the null hypothesis.

Pairwise Comparison (LSD):

1. Compact cars vs. midsize cars

Step 1:

$$H_0: \mu_1 = \mu_2, \qquad H_1: \mu_1 \neq \mu_2$$

Step 2:

$$t_{stat} = \frac{666.67 - 473.67}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} = 5.75$$

Step 3:

$$t_{stat} = 5.75 > 2.447 = t_{0.05;6}$$

Thus, we reject the null hypothesis.

2. Compact cars vs. full-size cars

Step 1:

$$H_0: \mu_1 = \mu_3, \qquad H_1: \mu_1 \neq \mu_3$$

Step 2:

$$t_{stat} = \frac{666.67 - 447.33}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} \approx 6.50$$

Step 3:

$$t_{stat} = 6.50 > 2.447 = t_{0.05;6}$$

Thus, we reject the null hypothesis.

3. Midsize cars vs. full-size cars

Step 1:

$$H_0: \mu_2 = \mu_3, \qquad H_1: \mu_2 \neq \mu_3$$

Step 2:

$$t_{stat} = \frac{473.67 - 447.33}{\sqrt{1709 \cdot \left(\frac{1}{3} + \frac{1}{3}\right)}} \approx 0.780$$

Step 3:

$$t_{stat} = 0.789 < 2.447 = t_{0.05;6}$$

Thus, we do not reject the null hypothesis.

Part II: *SPSS*

1.

1.
See the data file *weight.sav*

2.

- ***Analyze -> Compare Means -> One-Way ANOVA...***
- Place *weight* in the ***Dependent List*** box and *method* in the ***Factor*** box.

  Select ***Options...***

  - Choose ***Homogeneity of variance test***, ***Means plot***

    Continue.
- Select ***Post Hoc...***

  Choose ***LSD.***

  Continue.
- ***OK.***
- Select

  Analyze -> Descriptive Statistics -> Explore...

  Move *weight* to ***Dependent List*** and *method* to ***Factor List.***

  Click on the ***Plots...***
- Choose ***Normality plots with tests***, ***Untransformed***

  Continue
- ***OK.***

Output and Interpretation:

ANOVA

weight

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	160,133	2	80,067	9,168	,004
Within Groups	104,800	12	8,733		
Total	264,933	14			

$$p-value = 0.004 < 0.05 = \alpha$$

Therefore, the null hypothesis that all population means are equal will be rejected. Apart from that the table summarises some calculations to perform the test.

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	method	Statistic	df	Sig.	Statistic	df	Sig.
weight	Method 1	,175	5	,200*	,974	5	,899
	Method 2	,199	5	,200*	,967	5	,858
	Method 3	,191	5	,200*	,958	5	,794

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

The results of two tests examining the assumption that the populations from which the samples were obtained must be (approximately) normally distributed are given in the table "Test of Normality". The one to use is the "Shapiro-Welk" test.
The *p-values* given in the last column for *weight*, for each of the methods are large enough to conclude that the assumption of (approximate) normality of the three populations should not be rejected

H_0: The population random variable is normally distributed.

H_1: The population random variable is not normally distributed.

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
weight	Based on Mean	,384	2	12	,689
	Based on Median	,205	2	12	,818
	Based on Median and with adjusted df	,205	2	10,977	,818
	Based on trimmed mean	,383	2	12	,690

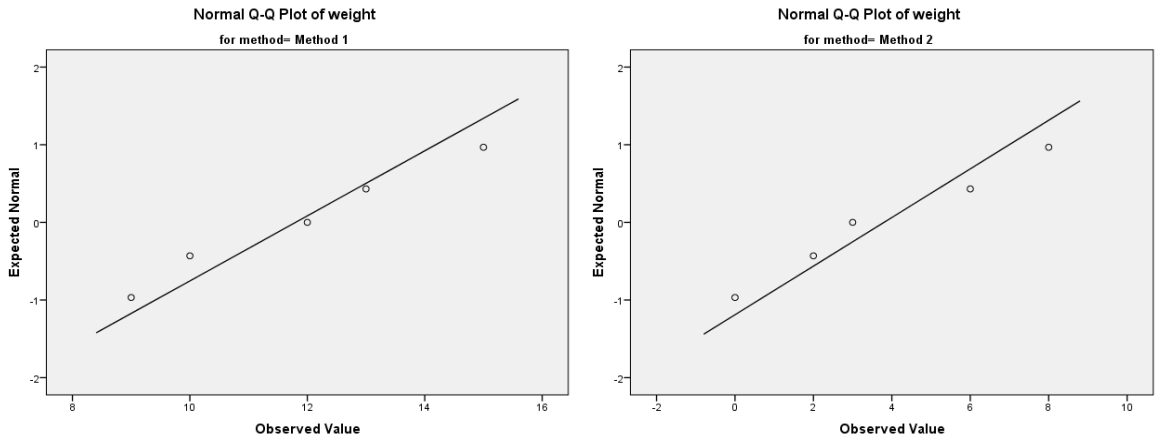
The table "Test of Homogeneity of Variance" tests the following hypotheses:

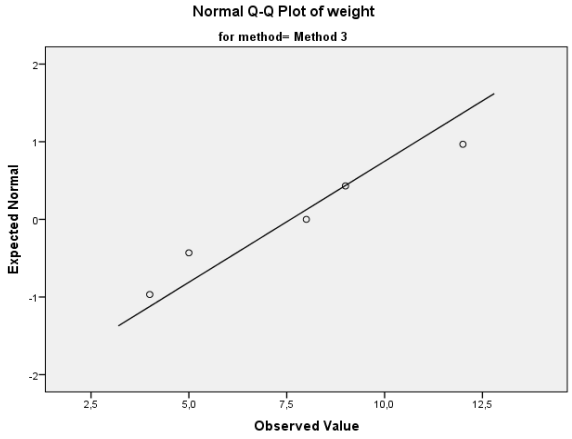
H_0 : The population variances are equal
H_1 : The population variances are not equal.

The test to use here is the one that is "Based on Median".
Here too, the *p-value* given in the last column is sufficiently large to conclude that the assumption of constant variances should not be rejected.

In addition to the above tables, several scatter plots appear in the output; of these the three "***Normal Q-Q Plots***" can be useful:

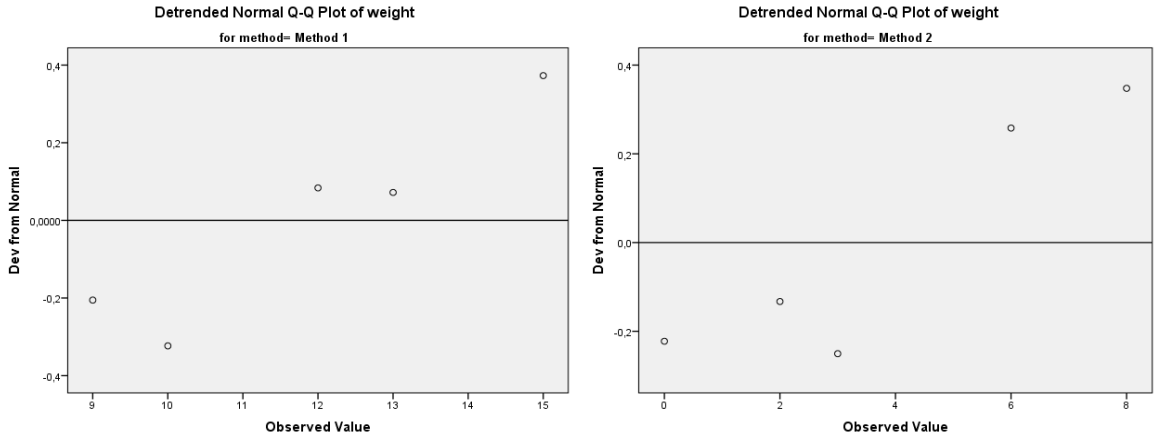
(A Q-Q plot charts observed values against a known distribution, in this case a normal distribution. If our distribution is normal, the plot would have observations distributed closely around the straight line.)

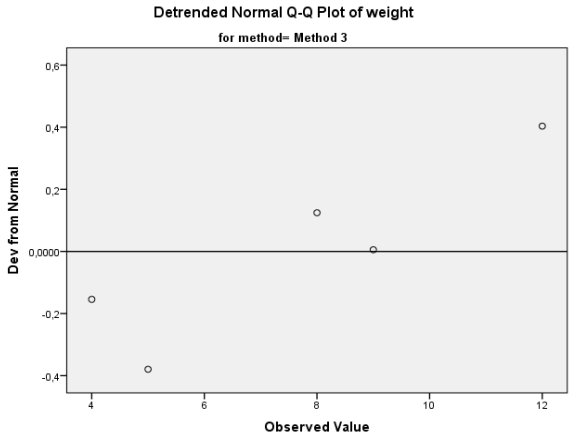




The ***Detrended normal Q-Q plots*** confirm the normality of the populations from which the samples have been chosen:

(They show the differences between the observed and expected values of a normal distribution. If the distribution is normal, the points should cluster in a horizontal band around zero with no pattern.)





However, in this case the data samples are small and the plots are not very informative.

A graphical assessment of the constant variance assumption may be performed using the following procedure:

- Select

  Graphs -> Legacy Dialogs -> Error Bar…

- Choose ***Simple***, ***Summaries for groups of cases***

  Define

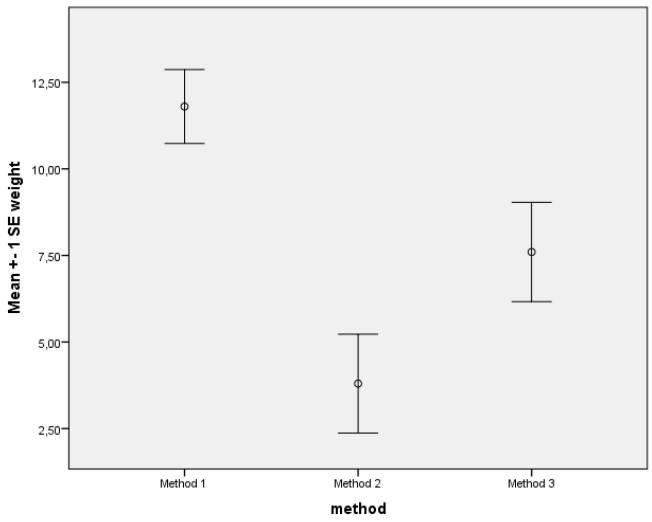
- Move *weight* to ***Variable*** and *method* to ***Category Axis***

  Set the ***Bars Represent*** option to be ***Standard error of mean***

  Use a ***Multiplier*** of 1

  OK

Output and Interpretation:



If the error bars are close to each other in length, as appears to be the case here, one might expect the constant variance assumption to be approximately valid.

The validity of the independence of samples is difficult to assess in the process of setting up for the one-way ANOVA *F* test.

The results of multiple comparisons are summarised in the following table:

Multiple Comparisons

Dependent Variable: weight

LSD

(I) method	(J) method	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Method 1	Method 2	8,00000*	1,86905	,001	3,9277	12,0723
	Method 3	4,20000*	1,86905	,044	,1277	8,2723
Method 2	Method 1	-8,00000*	1,86905	,001	-12,0723	-3,9277
	Method 3	-3,80000	1,86905	,065	-7,8723	,2723
Method 3	Method 1	-4,20000*	1,86905	,044	-8,2723	-,1277
	Method 2	3,80000	1,86905	,065	-,2723	7,8723

*. The mean difference is significant at the 0.05 level.

2.

1.
Share of unemployed youth to total unemployed (per cent for males)" [*ythunemm*] is quantitative, satisfying the level of measurement requirement for the dependent variable. "Degree of urbanization" [*urbaniz*] is categorical, satisfying the level of measurement requirement for the independent variable. Therefore, the statement is true.

2.
While SPSS has a procedure for one-way analysis of variance, we will use the General Linear Model procedure, since that is what we will have to use when we next move to two-factor analysis of variance.

- ***Analyze -> General Linear Model -> Univariate***

- Move the dependent variable, *ythunemm*, to the ***Dependent Variable*** text box and the independent variable, *urbaniz*, to the ***Fixed Factor(s)*** list box.

  Click on the ***Options*** button to request basic statistics.

- Mark the checkboxes for: ***Descriptive statistics***, ***Homogeneity tests***, and ***Residual plot***.

  Move the dependent variable *urbaniz* to the ***Display Means for*** list box.
  Continue.

- Click on the ***Save*** button to instruct SPSS to compute the standardized residuals.
  Click on the ***Standardized*** check box to create standardized residuals in the SPSS Data Editor.
  Continue.

- Click on the ***Post Hoc*** button to instruct SPSS to compute the ***Post Hoc Tests for:***
  Click un the check box for the ***Bonferroni*** post hoc test.
  Continue.

- ***OK.***

Output and Interpretation:

Between-Subjects Factors

		Value Label	N
Degree of urbanization	1	bottom third	15
	2	middle third	36
	3	top third	41

2.

The number of cases with valid data to analyze the relationship between "degree of urbanization" and "share of unemployed youth to total unemployed (per cent for males)" is 15 + 36 + 41 = 92.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Share of unemployed youth to total unemployed (per cent for males)

F	df1	df2	Sig.
,629	2	89	,535

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

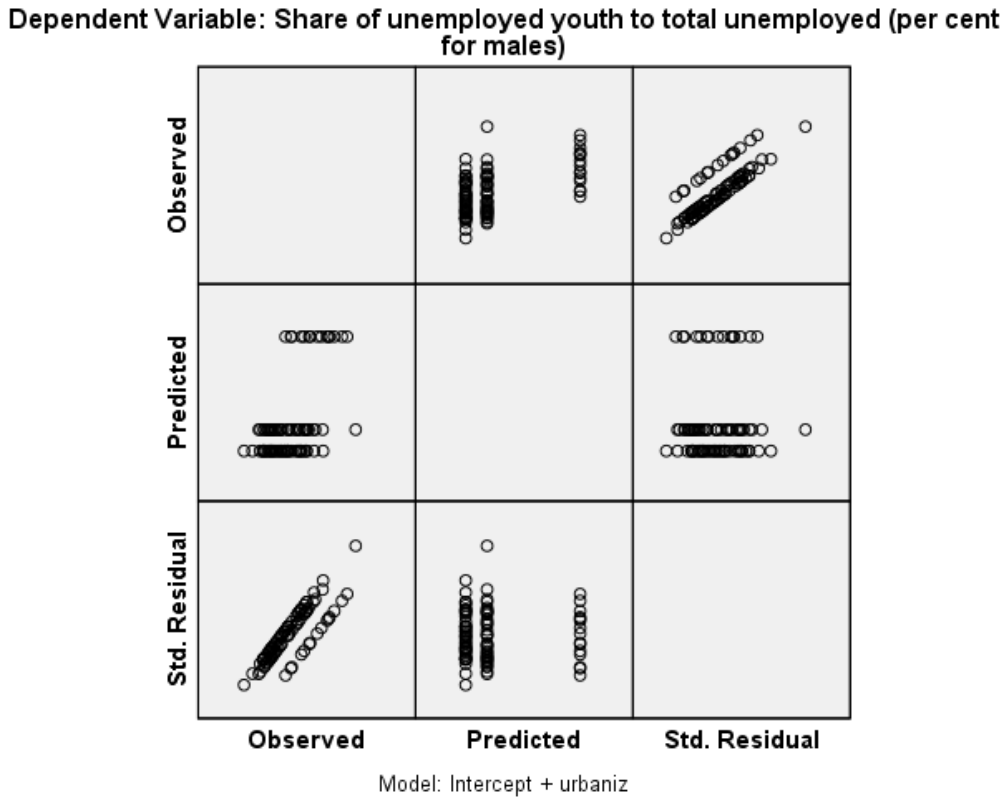
a. Design: Intercept + urbaniz

Because of

$$p-value = 0.535 \geq 0.05 = \alpha$$

we do not reject the null hypothesis that the assumption of equal variance is <u>not</u> rejected.

The residual plot indicates that the spreads (heights) of the points for each group are similar, reinforcing the interpretation that the equal variance condition is satisfied:



- ***Analyze -> Descriptive Statistics -> Explore***

- Move the dependent variable, *ythunem,* and the variable for standardized residuals (*ZRE_1*) to the ***Dependent List***.

  (While we are interested in the normality test only for the standardized residuals, we include the dependent variable *ythunemm* so that we have its skewness value in case we have to transform the variable.)

  Click on the ***Plots*** button to request the normality test.

- Mark the check box for ***Normality plots with tests***.

  Clear the ***Stem-and-leaf*** check box and mark the ***Histogram***.

  Continue.

- Mark the ***Option*** button to ***Exclude cases pairwise***.

  Continue.

- ***OK.***

Output and Interpretation:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Share of unemployed youth to total unemployed (per cent for males)	,079	92	,200*	,967	92	,021
Standardized Residual for ythunemm	,086	92	,088	,970	92	,033

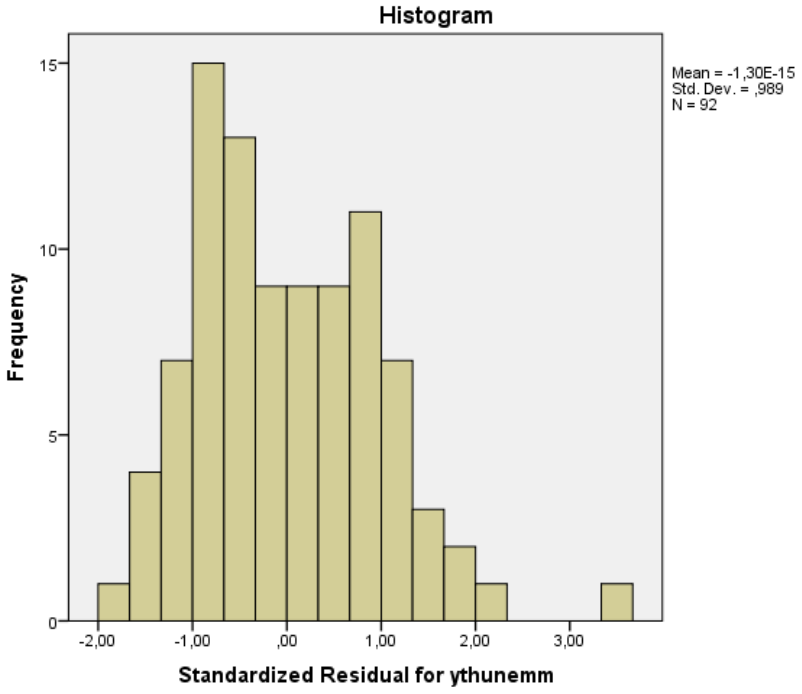
*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

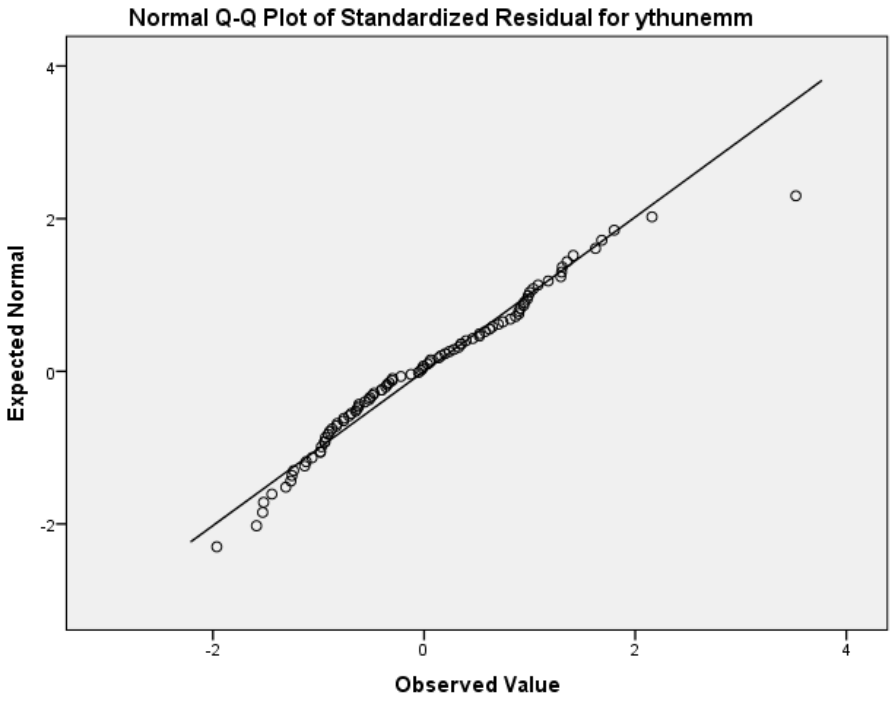
Because of $p-value = 0.033 < 0.05 = \alpha$

we reject the null hypothesis that the residuals are normally distributed.

The histogram does not show a major departure from normality:



There is one value that appears to be an outlier:



Therefore:
The Shapiro-Wilk test is very sensitive to departures from normality.
The normality plot also does not show a serious departure from normality.
The outlier is more evident.

When the raw data does not satisfy the conditions of equal variance and normality, we examine the skewness of the variable to identify skewing that might be corrected with re-expression:

Descriptives

			Statistic	Std. Error
Share of unemployed youth to total unemployed (per cent for males)	Mean		38,799	1,5385
	95% Confidence Interval for Mean	Lower Bound	35,743	
		Upper Bound	41,855	
	5% Trimmed Mean		38,147	
	Median		37,550	
	Variance		217,769	
	Std. Deviation		14,7570	
	Minimum		8,7	
	Maximum		82,1	
	Range		73,4	
	Interquartile Range		22,1	
	Skewness		,608	,251
	Kurtosis		,067	,498
Standardized Residual for ythunemm	Mean		,0000	,10311
	95% Confidence Interval for Mean	Lower Bound	-,2048	
		Upper Bound	,2048	
	5% Trimmed Mean		-,0323	
	Median		-,0415	
	Variance		,978	
	Std. Deviation		,98895	
	Minimum		-1,96	
	Maximum		3,52	
	Range		5,49	
	Interquartile Range		1,61	
	Skewness		,565	,251
	Kurtosis		,474	,498

The skewness for "share of unemployed youth to total unemployed (per cent for males)" [*ythunemm*] was $0.608 > 0$.

We now attempt to correct violation of assumptions by re-expressing "share of unemployed youth to total unemployed (per cent for males)" on a logarithmic scale and not the +1/-1 criteria to determine normality:

- ***Transform -> Compute Variable***

- Type the name for the transformed variable (*LG_ythuunemm*) in the ***Target Variable*** text box.

  Type the formula for the transformation in the ***Numeric Expression*** text box: LG10(*ythunemm)*

  Click on the ***OK*** button to close the dialog box.

- To repeat the one-way analysis of variance, click on ***Dialog Recall*** tool button.

- From the drop down menu, select the ***Univariate*** command.

  Replace the raw dependent variable *ythunemm* with the log transformed variable *LG_ythunemm*.

  OK.

- Select the ***Explore*** command from the drop down menu for the ***Dialog Recall*** tool button.

  Remove the variables from the ***Dependent List*** and add variable for the standardized residual from the second run of the ***Univariate*** command, *ZRE_2*.

  OK.

Output and Interpretation:

Levene's Test of Equality of Error Variances[a]

Dependent Variable: LG_ythuunemm

F	df1	df2	Sig.
1,835	2	89	,166

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + urbaniz

Because of

$$p-value = 0.166 > 0.05 = \alpha$$

the hypothesis of equal variance is not rejected.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual for LG_ythuunemm	,054	92	,200*	,977	92	,104

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Because of

$$p-value = 0.977 > 0.05 = \alpha$$

the hypothesis of normality is not rejected.

3.

Tests of Between-Subjects Effects

Dependent Variable: LG_ythuunemm

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	,643[a]	2	,322	13,659	,000
Intercept	192,544	1	192,544	8179,743	,000
urbaniz	,643	2	,322	13,659	,000
Error	2,095	89	,024		
Total	225,603	92			
Corrected Total	2,738	91			

a. R Squared = ,235 (Adjusted R Squared = ,218)

Because of

$$p-value = 0.000 < 0.050 = \alpha$$

the hypothesis that the means of the populations represented by the groups in the sample were all equal will be rejected.
Therefore, at least one of the means of the populations represented by the groups in the sample was different from the other means.

4. – 5.

Multiple Comparisons

Dependent Variable: LG_ythuunemm

Bonferroni

(I) Degree of urbanization	(J) Degree of urbanization	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
bottom third	middle third	,1886*	,04715	,000	,0735	,3036
	top third	,2414*	,04630	,000004	,1284	,3544
middle third	bottom third	-,1886*	,04715	,000	-,3036	-,0735
	top third	,0528	,03504	,405480	-,0327	,1383
top third	bottom third	-,2414*	,04630	,000	-,3544	-,1284
	middle third	-,0528	,03504	,405	-,1383	,0327

Based on observed means.

The error term is Mean Square(Error) = ,024.

*. The mean difference is significant at the ,05 level.

Because of

$$p-value = 0.000004 < 0.050 = \alpha$$

"Countries where the degree of urbanization was in the bottom third of all nations had a different of unemployed youth to total unemployed (per cent for males) with those where the degree of urbanization was in the top third of all nations."

We have

$$p-value = 0.405480 > 0.050 = \alpha.$$

Therefore, the mean "share of unemployed youth to total unemployed (per cent for males)" for countries where the degree of urbanization was in the middle third of all nations was not statistically different for the mean for those where the degree of urbanization as in the top third of all nations.