### GOODNESS OF FIT USING SPSS

#### WHAT IS GOODNESS OF FIT ?

A goodness-of-fit is a statistical technique. It is Applied to measure "how well the actual(observed) data points Fit into a Machine Learning model ". It summarizes the divergence between actual observed data points and expected data points in context to a statistical or machine learning model.

Assessment of divergence between the observed data points and model predicted data points is critical to understand, a decision made o poorly fitting models might be badly misleading. A seasoned practitioner must examine the fitment of actual and model predicted data points.

#### WHY DO WE TEST GOODNESS OF FIT ?

Goodness-of-fit tests are statistical tests to determine values match those predicted by the model . Goodness-of-fit tests are frequently applied in business decision making .

### WHAT ARE THE MOST COMMMON GOODNESS OF FIT TEST ?

There are multiple methods for determining goodness-of-fit. Some of the most popular methods used in statistics include the <u>chi-square</u>, the <u>Kolmogorov - Smirnov</u> test, the <u>Anderson-Darling</u> test and the <u>Shipiro - Wilk test.</u>

#### FOR EXAMPLE :

The below image depict the linear regression function. The Goodness-of-fit tests here will compare the actual observed values denoted by bule dots to the predicted vaules denoted by the red regression line.



# EXPLAIN TWO TEST ONLY

1. CHI – SQUARE TEST GOODNESS OF FIT USING SPSS

2. NORMAL DISTRIBUTION FIT USING SPSS

### 1. CHI – SQUARE TEST FOR GOODNESS OF FIT USING SPSS<sup>A</sup>

Chi-Square goodness of fit test is a non-parametric test that is used to find out how the observed value of a given phenomena is significantly different from the expected value. In Chi-Square goodness of fit test, the term goodness of fit is used to compare the observed sample distribution with the expected probability distribution. Chi-Square goodness of fit test determines how well theoretical distribution (such as normal, binomial, or Poisson) fits the empirical distribution. In Chi-Square goodness of fit test, sample data is divided into intervals. Then the numbers of points that fall into the interval are compared, with the expected numbers of points in each interval.

# The chi – square test for a goodness of fit test is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

 $x_{c}^{2}$  = chi – square goodness of fit test  $O_{i}$  = an obsrved count for bin *i*   $E_{i}$  = an expected count for bin *i*, asserted by the null hypothesis.

#### The Expected frequency is calculated by

$$E_i = \left(F(Y_u) - F(Y_l)\right)N$$

F =the cumulative distribution function for the probability distribution being tested Y<sub>u</sub> =the upper limit for class I Y<sub>i</sub> =the lower limit for class i , and N =the sample size

# Application of Chi-square as the goodness of fit

The Chi-square is applied to establish or refute that a relationship exists between actual observed values and predicted values. The chi-squared test is a very useful tool for predictive analytics professionals. It is used very commonly in Clinical research, Social sciences, and Business research.

It is also right tail test.

### Procedure for Chi-Square Goodness of Fit Test:

HYPOTHESIS :

Null Hypothesis : Ho There is no diference between observed and expected hypothesis

Alternative Hypothesis : H<sub>1</sub> There is a significant difference between observed and expected hypothesis

# Example of chi – square goodness of fit test in SPSS

A shop owner claims that an epual numbers of customers come into his shop each weekday. To test his Hypothesis , a researcher records the number of customers that come into the shop on a given week and Find the following. Monday-50 customers Tuesday-60 customers Wednesday-40 customers Thursday-47 customers Friday-53 customers Use the following steps to perform a Chi-Square goodness of fit test in SPSS to Determine if the data is consistent with the shop owner's claims.

### Procedure for chi-square test

Step 1: Open SPSS softwear and select variable view and enter the variables.

Step 2 : Select the data view and enter the data.
Step 3 : Select the analysis and click Non-parametric test.
Step 4 : Click legacy dialogs and select chi-square and transform the test variables list box into the no.of.customers.
Step 5 : Select option and tick the discriptive and select continue.

Step 6 : Finally click ok to get output.

### Step 1 : Enter the variable view and enter variables in SPSS

🔒 *Untitled1	[DataSet0] - IBM	SPSS Statistics D	ata Editor										- 0 ×	
<u>File</u> dit	<u>V</u> iew <u>D</u> ata	Transform	<u>A</u> nalyze	<u>Graphs</u> <u>U</u>	tilities Extension	s <u>W</u> indow	<u>H</u> elp							
			× 🔛		<b>H</b>									
	Name	Туре	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role			
1	Days	Numeric	8	0	No of the days	{1, Monday}	None	8	■ Center	🙈 Nominal	🔪 Input			
2	Count	Numeric	8	0	No of the custo	None	None	8	를 Center	Scale 🔗	S Input			
3														
4											1			
5			-						-					
6														
/														
õ														
10			2											
11														
12				2			2							
13														
14				2		2	20							
15											1			
16			2								1			
17														
18														
19														
20		2	1						1		(			
21		2	2								1			
22					0									
23										1	1			
24														
25														
26														
	1												4	
Data View	ariable View													
										IBM SPS	S Statistics Proces	sor is ready Unicode:ON	Weight On	

## Step 2 : Select data view and enter data in SPSS

🔒 *Untitled1	[DataSet0] - IBM	SPSS Statistics	Data Editor														ð X	
<u>F</u> ile <u>E</u> dit	<u>V</u> iew <u>D</u> ata	<u>T</u> ransform	<u>A</u> nalyze	<u>G</u> raphs <u>U</u> t	tilities E <u>x</u> te	nsions <u>W</u> in	dow <u>H</u> elp											
<b>a</b> H			<b>1</b>		<b>H</b>													
2 : Count	60															Visible: 2 d	f 2 Variables	
	💰 Days	Sount 🔗	var	var	var	var	var	var	var	var	var	var	var	var	var	var	var	
1	Monday	50															4	
2	Tuesday	60																
3	Wednesday	40																
4	Thursday	47																
5	Friday	53																
6							1	-										
7																		
8																		
9																		
10																		//N
11																		MV.
12									-									1/1//
13																		
14																		
16							N											
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24					1													
25																	*	
	1																	
Data View	/ariable View																	
											IBM S	SPSS Statistic	s Processor i	s ready	Unicode:0	DN Wei	ght On	

### Step 3 : Select the analysis and click Non-parametric test and Click legacy dialogs and select chi-square.

: Count	60		Bayesian Statistics											Visible: 2 of	f 2 Variables	
	💑 Days	Sount 🔗	Ta <u>b</u> les	ar	var	var	var	var	var	var	var	var	var	var	var	
1	Monday	50	Compare Means													
2	Tuesday	60	General Linear Model													
3	Wednesday	40	Generalized Linear Models													
4	Thursday	47	Mixed Models													
5	Friday	53	<u>C</u> orrelate											1		
6			Regression •	_												
0	<u></u>		Loglinear	-												
9			Neural Networks	-												
10			Classify	-												
11			Dimension Reduction	-				-								
12			Sc <u>a</u> le •				_									
13			Nonparametric Tests		One Sample											
14			Forecasting •	Δ	Independent Sar	nples										
15			<u>S</u> urvival		Related Sample	S										
16			Multiple Response		Legacy Dialogs	E.	Chi-se	Juare								
17			Missing Value Analysis				Dia Binom	nial								
18			Multiple Imputation				Dunc.							1		
19			Complex Samples				Turis.									
20			Bimulation				<u>1</u> -San	ipie K-S								
21			Quality Control	_			2 inde	pendent Sam	pies							
22			ROC Curve				K Inde	ependent Sam	ples							
23			Spatial and Temporal Modeling •	_			2 Rela	ated Samples.								
24			Direct Marketing	-			K Rela	ated <u>S</u> amples								
20	1															

### Step 4 : transform the test variables list box into the no . Of . customers

📲 Chi-square Test	×	
No of the days [Days]	Test Variable List:	
Expected Range © Get from data © Use specified range Lower: Upper:	Expected Values All categories equal Values: Add Change	
OK Paste	Remove Reset Cancel Help	

# Step 5 : Select option and tick the discriptive and select continue and click ok.

🝓 Chi-square Test: Options	×
Statistics	
Missing Values © Exclude cases <u>t</u> est-by-test © Exc <u>l</u> ude cases listwise	
Cancel Help	

### **OUTPUT**:

🚡 *Output1 [Document1] - IBM SPSS	S Statistics Viewer	– 🗆 X
<u>File Edit View Data Tran</u>	nsform Insert F <u>o</u> rmat <u>A</u> nalyze <u>G</u> raphs <u>U</u> tilities Extensions <u>W</u> indow <u>H</u> elp	
😑 🔳 🖨 🙇 🤚	A 💷 🗠 🛥 🧱 💒 🚍 🖻 🔊 💷 💽	
🖲 Output In Log	Descriptive Statistics	
NPar Tests	N Mean Std. Deviation Minimum Maximum	
Title	No of the customers 250 50.87 6.558 40 60	
→ ▲ Active Dataset     → ▲ Descriptive Statistics     → ▲ Chi-Square Test     → ▲ Title     → ▲ Frequencies     → ▲ Title	Chi-Square Test Frequencies	
Test Statistics	No of the customers	
	Observed N Expected N Residual	
	40 40 50.0 -10.0	
	47 47 50.0 -3.0	
	50 50 50.0 .0	
•	53 53 50.0 3.0	
	60 60 50.0 10.0 Tatal 250	
	Test Statistics	
	No of the	
	Chi-Square 4.360 <sup>a</sup>	
	df 4	
	Asymp. Sig	
(	a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.	-
	IBM SPSS Statistics Processor i	s ready Unicode:ON

#### **Conclusion :**

Chi-Square: The Chi-Square test statistic, found to be 4.36.

**df** : The degrees of freedom, calculated as #categories-1 = 5-1 = 4. **Asymp. Sig:** The p-value that corresponds to a Chi-Square value of 4.36 with 4 degrees of freedom, found to be .359. This value can also be found by using the <u>Chi-Square Score to P</u> <u>Value Calculator</u>.

Since the p-value (.359) is not less than 0.05, we fail to reject the null hypothesis. This means we do not have sufficient evidence to say that the true distribution of customers is different from the distribution that the shop owner claimed

#### 2. NORMAL DISTRIBUTION FIT USING IN SPSS

Normality test using SPSS :

An normality test is used to determine whether sample data has been drawn from a normal distribute population. What is Normal distribution ?

The normal distribution is always symmetrical about the mean which look like a "bell curve".

When testing for normality:

Probabilities > 0.05 indicate that the data are normal.

Probabilities < 0.05 indicate that the data are NOT normal.</li>

#### Normal Distribution Formula:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{3x - \mu}{\sigma}\right)^2}$$

where

σ is a population standard deviation;
µ is a population mean;
x is a value or test statistic;
<u>e</u> is a mathematical constant of roughly 2.72;
<u>π</u> a mathematical constant of roughly 3.14.

#### The following numerical and visual output must be investigated :

Skewness and kurtosis z-values
(should be somewhere in the span of -1.96 to +1.96)
The shapiro-wilk test p-value
(should be above 0.05)
Histograms , normal Q-Q plots and Box plots
(should be indicated that our data are approximately normal
Distributed).

### In a many statistical analysis, there are dependent variables and independent variable :

Dependent variable = a that variable depend on other factors.

For example : exam scores, as a variable, may be change depending on the student gender.

Independent variable= a variable that does not depend on the other factor.For example : gender does not change depending

on exam scores.

In this example, the exam scores should be approximately normally distributed for **both** males and females.





#### **Example:**

The students Gender = male and female Exam scores = 47,53,60,90,70,45,35,62,84,

#### Step 1 : Enter the variable view and enter variables in SPSS File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help r 🤉 🖾 🛓 🗐 H A 0 • 14 Name Width Decimals Label Values Missing Columns Align Measure Role Type 1 Gender Numeric 8 0 {1, Male}. None 8 Right 💑 Nominal > Input Exam Scores Numeric None I Right Scale 8 > Input 2 8 0 None 8 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 4 Data View Variable View IBM SPSS Statistics Processor is ready Unicode:ON

### Step 2 : Select data view and enter data in

							1											
) : Exam_Sco	res 84								1				1			Visible: 2 o	f 2 Variables	
	🗞 Gender	Exam_Sc ores	var	var	var	var	var	var	var	var								
1	Male	47																
2	Male	53																
3	Female	60																
4	Male	90																
5	Male	70														1		$\Box A$
6	Female	45																
7	Male	35																M
8	Male	62																
9	Female	84																
10		-																
11		-																
12	:-																	
13																		
14										· · · · · · · · ·								
15																-		
10						0												
10	<u> </u>																	
10																		
20																-		
21																		
27																		
23																		
24																		
24	4																	

#### Step 3 : Select the analysis and click discriptive statistics and select explore :

*Untitled1 [l	DataSet0] - IBM	SPSS Statistics I	Data Editor													J X	
<u>File E</u> dit	<u>V</u> iew <u>D</u> ata	Transform	<u>Analyze</u> <u>Graphs</u> <u>Utilities</u> Exter	nsions	Window	Help											
			Reports	*		A	0										
		,	Descriptive Statistics	•	123 Erequend	ies											
9 : Exam_Scor	res 84	1	Bayesian Statistics	۲	🔚 Descripti	ves		1	1	1	1	10	10	1	Visible: 2 or	f 2 Variables	
	Sender .	Exam_Sc	Tables	•			var	var	var	var	var	var	var	var	var	var	
1	Male	ores	Compare Means	•	Crosstab	s											
2	Male	53	<u>G</u> eneral Linear Model	•	Ratio									2	-		
3	Female	60	Generalized Linear Models	•	P-P Plots										-		
4	Male	90	Mixed Models	1	Q-Q Plots												(******/)
5	Male	70	Correlate														
6	Female	45	Regression														
7	Male	35	Loglinear														
8	Male	62	Classify														
9	Female	84	Dimension Reduction			_				-				-			
10			Scale											2			
11			Nonnarametric Tests														///////////////////////////////////////
12		-	Forecasting			-								-			
13			Survival							_							
14			Multiple Response			_									_		
16			Missing Value Analysis			_											
17			Multiple Imputation														
18	~		Complex Samples												-		
19	·		Simulation		· · · · · · · · · · · · · · · · · · ·									2			
20			Quality Control											1			
21																	
22			Spatial and Temporal Modeling	•										1	1		
23			Direct Marketing														
24			Diroct Margoung											1			
	1			_	_	_	***							_			
Data View V	ariable View																
Explore											IBM SPSS	Statistics Proc	essor is read	lv Ur	nicode:ON		
1///////	11/1///////////////////////////////////	///////////////////////////////////////		1////		1/////	///////	///////////////////////////////////////	///////////////////////////////////////	///////////////////////////////////////	N./M.N	///////////////////////////////////////	7//////				

# Step 4: transform the Dependent list box into the Exam\_scores and transform the Factor list box into the Gender

the Explore	×.
	Dependent List:       Statistics         ✓ Exam_Scores       Plots         Factor List:       Options         ✓ Gender       Bootstrap         Label Cases by:       Label Cases by:
Display © Both © Statistics © Plot OK Past	s e <u>R</u> eset Cancel Help

# Step 5 : Open polts and ticks histogram and normality plots with test and continue then click ok .

🛊 Explore: Plots 🛛 🗙	
Boxplots       Descriptive	
Normality plots with tests	
Spread vs Level with Levene Test	
© Power estimation	
© Transformed Power: Natural log	
© <u>U</u> ntransformed	
Cancel Help	

### **OUTPUT**:

👘 *Output1 [Document1] - IBM SP	PSS Statistics Viewer	– 🗆 🗙
<u>File Edit View Data Tra</u>	ransform Insert Format Analyze Graphs Utilities Extensions Window Help	
😑 H 🖨 🗟 성	2 II. ~ ~ III. II. II. II. II. II. II. II	
9 plore Title Notes Active Dataset Gender Title Case Processing Summary Descriptives Title Exam_Scores Title Histograms Title Gender= Male Gender= Male Gender= Female Detrended Normal 0-Q Title Gender= Female Detrended Normal 0-Q Title Gender= Female Detrended Normal 0-Q Title Gender= Female Detrended Normal 0-Q Fille Gender= Female Detrended Normal 0-Q Fille Gender= Female Detrended Normal 0-Q Fille Gender= Female Fille Gender= Female Fille Fille Gender= Female Fille Gender= Female Fille Gender= Female Fille Fille Gender= Female Fille Fille Gender= Female Fille Fille Fille Gender= Female Fill	Your temporary usage period for IBM SPSS Statistics will expire in 5517 days.         EXAMINE VARIABLES=Exam_Scores BY Gender         /PLOT BOXEDOT HISTOGRAM NEPLOT         /OWRARE GROUPS         /STATISTICS DESCRIPTIVES         /CINTERVAL 95         /MISSING LISTWISE         /NOTOTAL.         Explore         [DataSet0]         Gender         Valid       Nessing         Valid       Nessing         Valid       Nessing         Valid       Nessing         Exam_Scores       Male         Oender       Valid         Descriptives         Order       Statistic         Statistic       Std. Error         Exam_Scores       Male         Oender       Statistic         Statistic       Std. Error         Exam_Scores       Male         Off Maan       Lower Bound         Jose Found       Std. Error	
	IBM SPSS Statistics Processor is ready	Unicode:ON

output [bocument] - bin bra	5 Statistics	, memer						
ile <u>E</u> dit <u>V</u> iew <u>D</u> ata <u>T</u> rai	nsform	Insert For	mat <u>A</u> na	alyze <u>G</u> raphs <u>U</u> tilities	Extensions V	/indow <u>H</u> e	elp	
🚔 H 🖨 🗟 🕹			า 🗍	🗑 🔛 📩 🗐	2			
				Descriptive	5			
ore			Gender			Statistic	Std. Error	
Notes	E	am Scores	Male	Mean		59.50	7.843	
Active Dataset		-		95% Confidence Interval	Lower Bound	39.34		
Gender				for Mean	Upper Bound	79.66		
🖹 Inte 🛱 Case Processing Summary				5% Trimmed Mean		59.17		
Descriptives				Median		57.50		
Tests of Normality				Variance		369.100		
Exam_scores				Std. Deviation		19.212		
🖬 🖶 Histograms				Minimum		35		
Title				Maximum		90		
Gender= Male				Range		55		
Normal Q-Q Plots				Interquartile Range		31		
Title	-			Skewness		.534	.845	
Gender= Male	1			Kurtosis		.245	1.741	
Detrended Normal Q-Q			Female	Mean		63.00	11.358	
🖻 Title				95% Confidence Interval	Lower Bound	14.13		
Gender= Male				Ior Mean	Upper Bound	111.87		
Boxplot				5% Trimmed Mean				
				Median		60.00		
				Variance		387.000		
				Std. Deviation		19.672		
				Minimum		45		
				Maximum		84		
				Range		39		
				Interquartile Range				
				Skewness		.670	1.225	
				Kurtosis			с.	











#### **Conclusion :**

The skewness and kurtosis measure should be as to zero as possible in spss.

As a consequence you must divide the measure by its standard error.and you need to do this hand using a calculator.

This will give the z-vaule, which should be somewhere between -1.96 to +1.96.

Male : to calculate the skewness z-vaule , divide the skewness by its standard error.

0.534/0.845 = 0.63

To value is 0.63 neither below -1.96 nor above +1.96 0.245/1.741 = 0.14 **Female:** 0.670/1.225 = 0.54

All the z-values are within +/- 1.96.

Conclusion : Regarding skewness and kurtosis .our example data are little skewed and kurtotic, for both males and females, but it does not differ significantly from normality.

The null hypothesis for the test of normality is that the data are normally distributed.

In spss p-value is labeled by "sig".

Both p-values are above 0.05.we keep null hypothesis.