正态性检验方法

1、通过 SPSS 数据统计分析工具

在 SPSS 中，可以通过使用 Kolmogorov-Smirnov 检验来验证数据是否符合正态分布。

单样本 Kolmogorov-Smirnov 检验

N		2040
正态参数[a,b]	均值	27.22201254
	标准差	4.986570293
最极端差别	绝对值	.055
	正	.026
	负	-.055
Kolmogorov-Smirnov Z		2.488
渐近显著性(双侧)		.000

a. 检验分布为正态分布。
b. 根据数据计算得到。

比如以上结果显示，在 0.001 的显著性水平下，p 值为 0.000，原假设（H0：假设数据符合正态分布）被拒绝，说明不能接收数据服从正态分布的假设。

单样本 Kolmogorov-Smirnov 检验

N		484
正态参数[a,b]	均值	23.40427795
	标准差	4.923757137
最极端差别	绝对值	.045
	正	.045
	负	-.023
Kolmogorov-Smirnov Z		.992
渐近显著性(双侧)		.279

a. 检验分布为正态分布。
b. 根据数据计算得到。

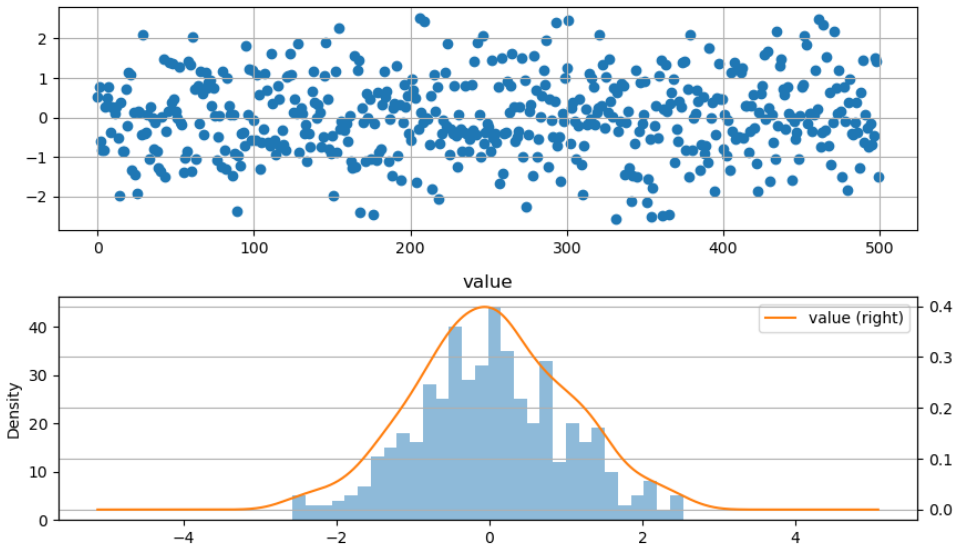
以上结果显示，在 0.001 的显著性水平下，p 值为 0.279，原假设（H0：假设数据符合正态分布）不能被拒绝，说明可以接收数据服从正态分布的假设。

2、通过 Python 进行 KS 检验

Python 的 scipy.stats 模块提供了一系列的检验函数，包括：
（1）kstest：可用于检验样本是否服从正态分布、指数分布、伽马分布等；
（2）normaltest：专门用于检验样本是否服从正态分布；
（3）shapiro：也是专门用于做正态检验。

具体方法可以参考：
https://blog.csdn.net/tszupup/article/details/108432814

（一）先通过统计直方图和密度图，查看数据是否近似服从正态分布，如下：



（二）然后，可以通过上述 kstest 等函数，来进行数据是否符合正态分布的统计检验。

如果 p-value 大于特定的显著性水平 α，则说明需要接收原假设，即数据服从正态分布；否则，需要拒绝原假设，即数据不服从正态分布。

3、通过 Excel 进行正态性检验

（一）通过 Excel，画出统计直方图，根据统计直方图来初步判断是否是正态的

（0）数据准备

a、假设数据为：（请将下面的数据放到 Excel 表格的第 A 列，并在 A1 中输入“数据”）

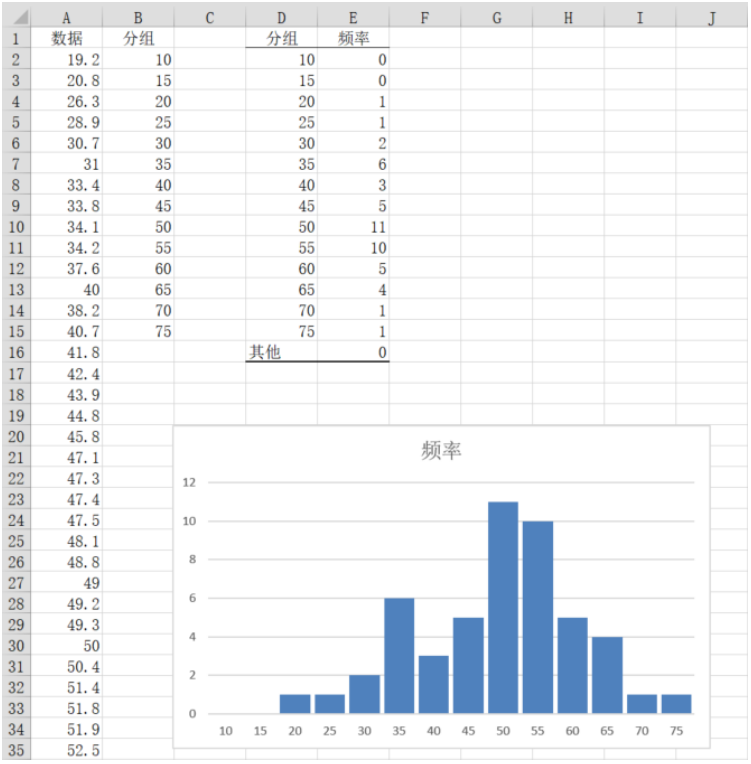
19.20	37.60	47.30	51.40	56.10
20.80	40.00	47.40	51.80	57.50
26.30	38.20	47.50	51.90	58.00
28.90	40.70	48.10	52.50	58.90
30.70	41.80	48.80	52.70	60.10
31.00	42.40	49.00	53.10	60.40
33.40	43.90	49.20	53.20	61.80
33.80	44.80	49.30	53.20	61.90
34.10	45.80	50.00	54.60	69.00
34.20	47.10	50.40	55.20	71.30

b、假设数据“分组”为：10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75，
并将该分组放到 Excel 表格的第 B 列，并在 B1 中输入“分组”。

（1）在 Excel 中，选择“数据”→“数据分析”→“直方图”

a、在弹出的对话框中，“输入区域”选择“$A$1:$A$51”；“接收区域”选择“$B$1:$B$15”；选中“标志”；“输出区域”选择“$D$1”；确定。

b、以生成出来的直方图表格，绘制“柱形图”得到直方图。

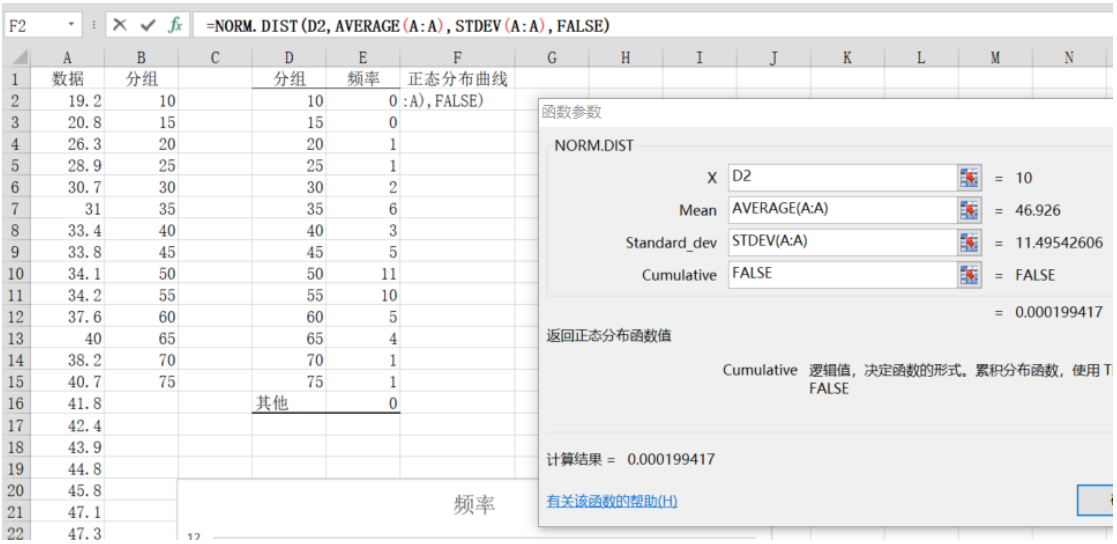


	A	B	C	D	E	F	G	H	I	J
1	数据	分组		分组	频率					
2	19.2	10		10	0					
3	20.8	15		15	0					
4	26.3	20		20	1					
5	28.9	25		25	1					
6	30.7	30		30	2					
7	31	35		35	6					
8	33.4	40		40	3					
9	33.8	45		45	5					
10	34.1	50		50	11					
11	34.2	55		55	10					
12	37.6	60		60	5					
13	40	65		65	4					
14	38.2	70		70	1					
15	40.7	75		75	1					
16	41.8			其他	0					
17	42.4									
18	43.9									
19	44.8									
20	45.8									
21	47.1									
22	47.3									
23	47.4									
24	47.5									
25	48.1									
26	48.8									
27	49									
28	49.2									
29	49.3									
30	50									
31	50.4									
32	51.4									
33	51.8									
34	51.9									
35	52.5									

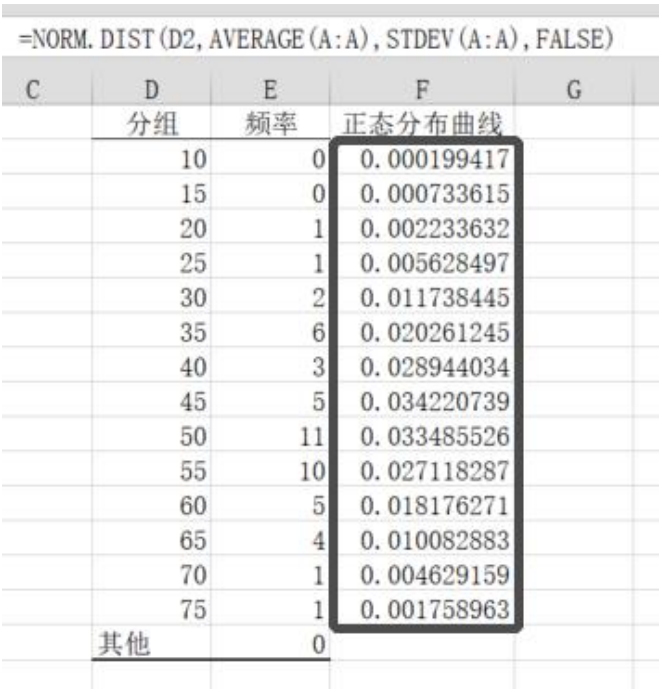
（2）在 Excel 中，通过正态分布的概率密度正态分布函数 NORMDIST 来计算，并绘制出同样分组的正态分布，并绘制到同一个直方图中，比较两者的分布是否一致

=NORM.DIST(X, Mean, StdDev, Cummulative)，即：

=NORM.DIST(D2, AVERAGE(A:A), STDEV(A:A), FALSE)



（3） 向下填充



C	D	E	F	G
	分组	频率	正态分布曲线	
	10	0	0.000199417	
	15	0	0.000733615	
	20	1	0.002233632	
	25	1	0.005628497	
	30	2	0.011738445	
	35	6	0.020261245	
	40	3	0.028944034	
	45	5	0.034220739	
	50	11	0.033485526	
	55	10	0.027118287	
	60	5	0.018176271	
	65	4	0.010082883	
	70	1	0.004629159	
	75	1	0.001758963	
	其他	0		

（4）在直方图中增加正态分布曲线图

a、在直方图内右键→选择数据→添加→

b、系列名称：选中 F1 单元格

c、系列值：选中 F2:F15

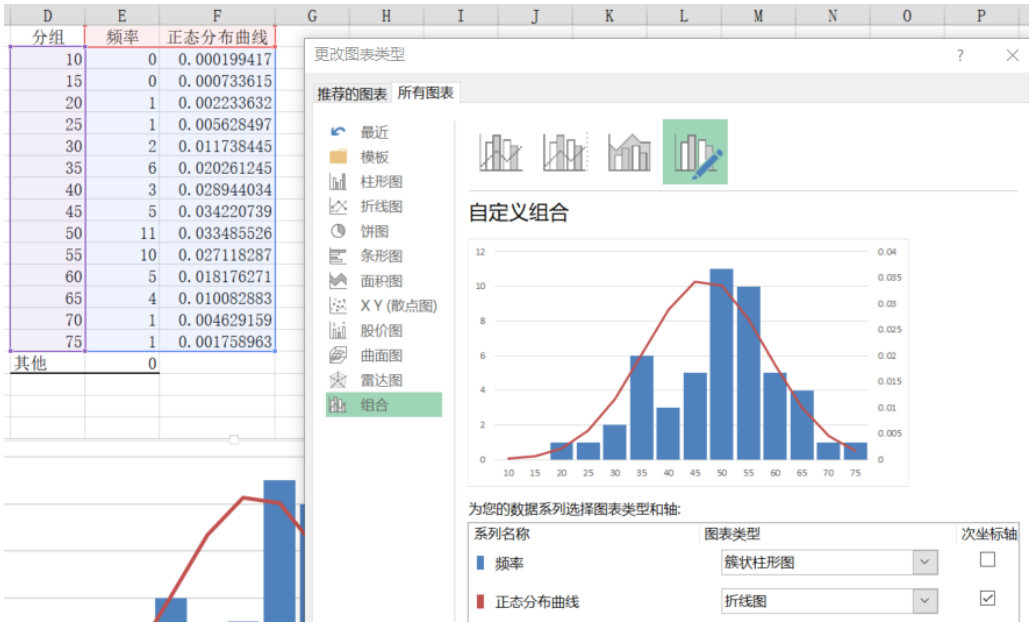
d、确定、确定

（5）修整图形

a、在直方图内右键→选择“更改图表类型”

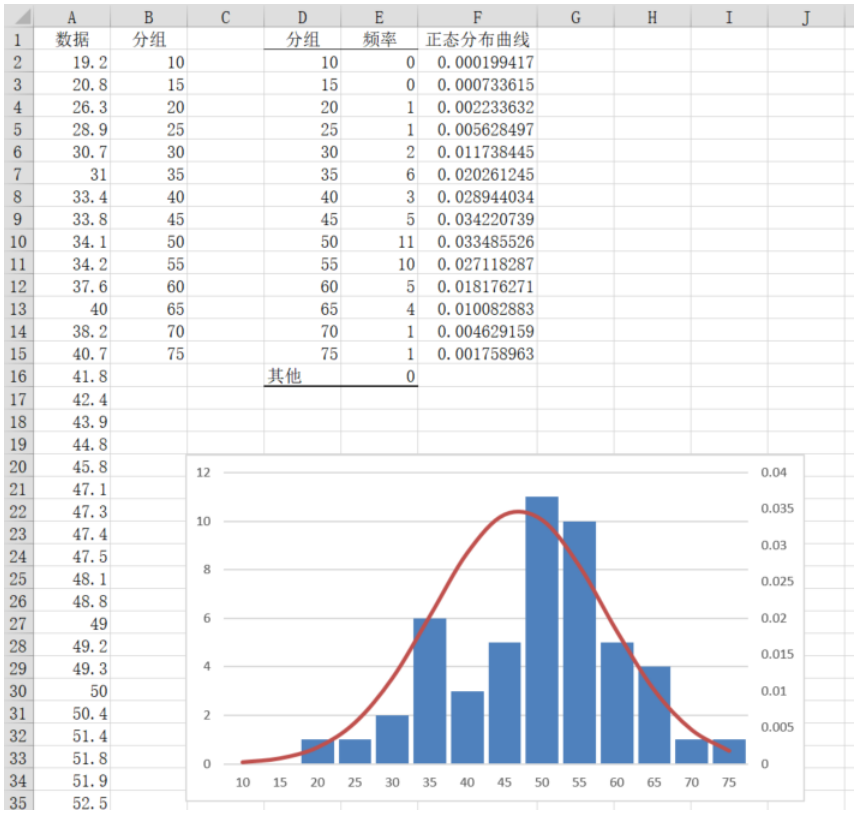
b、在“组合”中，在“正态分布曲线”部分，选择图表类型为“折线图”，选中“次坐标轴”

d、确定



（6）平滑正态分布图

选中正态分布曲线→右键→设置数据列格式→线型→勾选“平滑线”→关闭



（7）可从上述直方图和平滑正态分布图大致判断原数据（蓝色）基本符合正态分布特点。

（二）通过 Excel 进行卡方检验（Chi Squared Goodness Fit），并根据卡方检验的 p-value 来进行正态性检验

参考连接：https://www.inprolink.com/2019/02/20/normality-test-using-microsoft-excel/

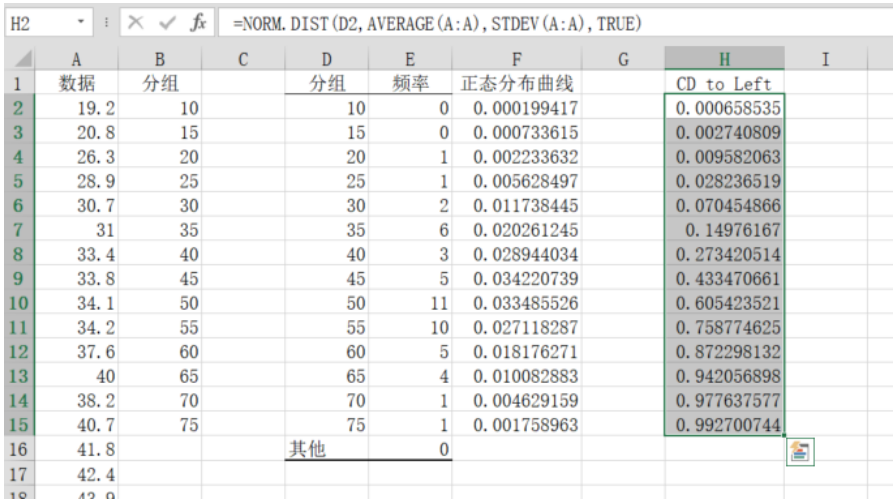
卡方检验 (chi-square test)，主要用于检验统计样本的实际观测值与理论推断值之间的偏离程度，或者是检验一批数据是否与某种理论分布相符合。
（A）卡方值是卡方检验时用到的检验统计量，卡方值越大，说明观测值与理论值之间的偏离就越大；反之，二者偏差越小。实际应用时，可以根据卡方值计算 P-value，从而选择拒绝或者接受原假设。
（B）卡方值χ^2的计算方法：

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e},$$

其中 o 代表 observation，即实际频数；e 代表 expectation，即期望频数。

（1）假设上述（一）中的操作中 D 列为“分组”列（上述参考连接中为“Bin”）

（2）计算正态分布从最左到当前分组边界值（由 D 列决定）的累计分布
将第 H 列命名为“CD to Left”，使用如下公式计算 H 列的每个值：
H2=NORM.DIST(D2, AVERAGE(A:A), STDEV(A:A), TRUE)（注意为 TRUE，表示计算累计分布）
H3=NORM.DIST(D3, AVERAGE(A:A), STDEV(A:A), TRUE)
……
（向下填充）



	A	B	C	D	E	F	G	H	I
1	数据	分组		分组	频率	正态分布曲线		CD to Left	
2	19.2	10		10	0	0.000199417		0.000658535	
3	20.8	15		15	0	0.000733615		0.002740809	
4	26.3	20		20	1	0.002233632		0.009582063	
5	28.9	25		25	1	0.005628497		0.028236519	
6	30.7	30		30	2	0.011738445		0.070454866	
7	31	35		35	6	0.020261245		0.14976167	
8	33.4	40		40	3	0.028944034		0.273420514	
9	33.8	45		45	5	0.034220739		0.433470661	
10	34.1	50		50	11	0.033485526		0.605423521	
11	34.2	55		55	10	0.027118287		0.758774625	
12	37.6	60		60	5	0.018176271		0.872298132	
13	40	65		65	4	0.010082883		0.942056898	
14	38.2	70		70	1	0.004629159		0.977637577	
15	40.7	75		75	1	0.001758963		0.992700744	
16	41.8			其他	0				
17	42.4								

（3）计算每个组（Bin）的概率值
将第 I 列命名为“Bin Only”，使用如下公式计算 I 列的每个值：
I2=H2（注意第一个为 H2）
I3=H3-H2（注意其他为两两相减，如 H3-H2、H4-H3……）
I4=H4-H3
……
I15=H15-H14

	A	B	C	D	E	F	G	H	I	J
1	数据	分组		分组	频率	正态分布曲线		CD to Left	Bin Only	
2	19.2	10		10	0	0.000199417		0.000658535	0.000659	
3	20.8	15		15	0	0.000733615		0.002740809	0.002082	
4	26.3	20		20	1	0.002233632		0.009582063	0.006841	
5	28.9	25		25	1	0.005628497		0.028236519	0.018654	
6	30.7	30		30	2	0.011738445		0.070454866	0.042218	
7	31	35		35	6	0.020261245		0.14976167	0.079307	
8	33.4	40		40	3	0.028944034		0.273420514	0.123659	
9	33.8	45		45	5	0.034220739		0.433470661	0.16005	
10	34.1	50		50	11	0.033485526		0.605423521	0.171953	
11	34.2	55		55	10	0.027118287		0.758774625	0.153351	
12	37.6	60		60	5	0.018176271		0.872298132	0.113524	
13	40	65		65	4	0.010082883		0.942056898	0.069759	
14	38.2	70		70	1	0.004629159		0.977637577	0.035581	
15	40.7	75		75	1	0.001758963		0.992700744	0.015063	
16	41.8			其他	0					
17	42.4									

（4）计算每个组的期望值（Expected Number）

将第 J 列命名为“Expected Number”，使用如下公式计算 J 列的每个值：

J2=I2*COUNT(A:A)

……

J15=I15*COUNT(A:A)

（5）有了上述期望值，即可与实际数据每组频率值（E 列）来进行比较，进行卡方检验

$$(expected-observed)^2/expected$$

将第 K 列命名为“Diff”，使用如下公式计算 K 列的每个值：

K2=(J2-E2)^2/J2

……

K15=(J15-E15)^2/J15

K2 =(J2-E2)^2/J2

	A	B	C	D	E	F	G	H	I	J	K
1	数据	分组		分组	频率	正态分布曲线		CD to Left	Bin Only	Expected Number	Diff
2	19.2	10		10	0	0.000199417		0.000658535	0.000659	0.032926747	0.032926747
3	20.8	15		15	0	0.000733615		0.002740809	0.002082	0.104113709	0.104113709
4	26.3	20		20	1	0.002233632		0.009582063	0.006841	0.34206268	1.265503497
5	28.9	25		25	1	0.005628497		0.028236519	0.018654	0.932722819	0.004852695
6	30.7	30		30	2	0.011738445		0.070454866	0.042218	2.110917336	0.005828109
7	31	35		35	6	0.020261245		0.14976167	0.079307	3.965340221	1.044006361
8	33.4	40		40	3	0.028944034		0.273420514	0.123659	6.182942178	1.638559866
9	33.8	45		45	5	0.034220739		0.433470661	0.16005	8.002507369	1.126528235
10	34.1	50		50	11	0.033485526		0.605423521	0.171953	8.597643008	0.671267592
11	34.2	55		55	10	0.027118287		0.758774625	0.153351	7.667555204	0.709521951
12	37.6	60		60	5	0.018176271		0.872298132	0.113524	5.676175326	0.080549498
13	40	65		65	4	0.010082883		0.942056898	0.069759	3.487938307	0.075175406
14	38.2	70		70	1	0.004629159		0.977637577	0.035581	1.779033948	0.341136768
15	40.7	75		75	1	0.001758963		0.992700744	0.015063	0.753158334	0.080900397
16	41.8			其他	0						

（6）通过以下公式计算卡方分布的 P 值

=chidist(x,df)

Chi-Squared Statistic　　Degrees of freedom

其中：　Chi-Squared Statistic，卡方统计量，卡方值χ^2，就是第 K 列的 Diff 的求和；

Degrees of freedom = #bins – 1 – #calculated_parameters，自由度。

在本任务中，我们有 14 个分组，因此#bins=14；为了进行卡方统计检验需要计算样本的均值和标准差，所以#calculated_parameters=2，因此最终的自由度为：14-1-2=11

K17=SUM(K2:K15)

K18=COUNT(D2:D15)-1-2

K19=CHISQ.DIST.RT(K17,K18)

K19 =CHISQ.DIST.RT(K17,K18)

	A	B	C	D	E	F	G	H	I	J	K
1	数据	分组		分组	频率	正态分布曲线		CD to Left	Bin Only	Expected Number	Diff
2	19.2	10		10	0	0.000199417		0.000658535	0.000659	0.032926747	0.032926747
3	20.8	15		15	0	0.000733615		0.002740809	0.002082	0.104113709	0.104113709
4	26.3	20		20	1	0.002233632		0.009582063	0.006841	0.34206268	1.265503497
5	28.9	25		25	1	0.005628497		0.028236519	0.018654	0.932722819	0.004852695
6	30.7	30		30	2	0.011738445		0.070454866	0.042218	2.110917336	0.005828109
7	31	35		35	6	0.020261245		0.14976167	0.079307	3.965340221	1.044006361
8	33.4	40		40	3	0.028944034		0.273420514	0.123659	6.182942178	1.638559866
9	33.8	45		45	5	0.034220739		0.433470661	0.16005	8.002507369	1.126528235
10	34.1	50		50	11	0.033485526		0.605423521	0.171953	8.597643008	0.671267592
11	34.2	55		55	10	0.027118287		0.758774625	0.153351	7.667555204	0.709521951
12	37.6	60		60	5	0.018176271		0.872298132	0.113524	5.676175326	0.080549498
13	40	65		65	4	0.010082883		0.942056898	0.069759	3.487938307	0.075175406
14	38.2	70		70	1	0.004629159		0.977637577	0.035581	1.779033948	0.341136768
15	40.7	75		75	1	0.001758963		0.992700744	0.015063	0.753158334	0.080900397
16	41.8			其他	0						
17	42.4									Chi Sq Statistics	7.18087083
18	43.9									Degree of Freedom	11
19	44.8									Chi Sq Distribution	0.784252717

（7）根据卡方检验的 p-value 来进行正态性检验

在本任务中，卡方检验的 p-value 是 0.784252717。因为 p-value 大于 0.05 的显著性水平，原假设（H0：假设数据符合正态分布）不能被拒绝，说明可以接收数据服从正态分布的假设。

（完）