**A8Wc Chi-square test for normality**

The chi-square goodness of fit test can be used to test the hypothesis that data comes from a normal hypothesis, where H0: data are sampled from a normal distribution.

**Example 1**

Consider the sample data and test whether the data is likely to be normally distributed with population mean and standard deviation of 25 and 8 respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Data Value | ID | Data Value |
| 1 | 22 | 24 | 27 |
| 2 | 33 | 25 | 16 |
| 3 | 30 | 26 | 30 |
| 4 | 26 | 27 | 26 |
| 5 | 22 | 28 | 28 |
| 6 | 32 | 29 | 14 |
| 7 | 36 | 30 | 35 |
| 8 | 6 | 31 | 23 |
| 9 | 28 | 32 | 17 |
| 10 | 24 | 33 | 35 |
| 11 | 21 | 34 | 22 |
| 12 | 24 | 35 | 28 |
| 13 | 10 | 36 | 27 |
| 14 | 26 | 37 | 28 |
| 15 | 18 | 38 | 23 |
| 16 | 36 | 39 | 32 |
| 17 | 16 | 40 | 26 |
| 18 | 26 | 41 | 19 |
| 19 | 18 | 42 | 12 |
| 20 | 24 | 43 | 14 |
| 21 | 12 | 44 | 25 |
| 22 | 27 | 45 | 23 |
| 23 | 21 | 46 |  |

Enter data into Excel – 1st 10 data values in Figure W8.2 (data values are in Cells C4:C48)



Figure 1 First 10 data values in Excel

Calculate summary statistics as illustrated in Figure W8.3: mean, five-number summary, measure of skewness and kurtosis.



Figure 2 Summary statistics

From Excel, Fisher’s skewness value = - 0.36 with a kurtosis value = - 0.17. The skewness value is not critical and therefore we can assume that the frequency distribution is approximately symmetric. The kurtosis value is not critical and tells us that the frequency distribution is like a normal distribution with respect to peak value and the distribution tails. Furthermore, we can use the five-number summary and corresponding box-and-whisker plot to check this is a reasonable interpretation.



Figure 3 Box-and-whisker plot

To test statistically if the data values can be considered to come from a normal distribution with population mean 25 and population standard deviation 8, we can conduct a chi-square goodness-of-fit test.

Create frequency distribution using the data values

Create a frequency table for the 45 data elements in the range C4:C48 based on the bin array J6:J9 (the text “over 32” in cell J10 is not part of the bin array). In this case, I have chosen X ≤ 6, 16, 24, 32, and over 32 but you can use your own BIN values in Cells J6:J10. To produce the output, highlight the range K6:K10 (i.e. a column range with one more cell than the number of bins) and enter the formula: =FREQUENCY(B4:B48,J6:J9). Since this is an array formula, you must press **Ctrl-Shift-Enter**.



Figure4 Creating the Grouped frequency distribution

Excel now inserts frequency values in the highlighted range K6:K10. Here K6 contains the number of data elements in the input range with value in the first bin (i.e. data elements whose value is ≤ 8). Similarly, K7 contains the number of data elements in the input range with value in the second bin (i.e. data elements whose value is > 8 and ≤ 16). The final output cell (K10) contains the number of data elements in the input range with value > the value of the final bin (i.e. > 32 for this example).

OR

You could use Data > Data Analysis > Histogram to create the frequency distribution and corresponding histogram.



Figure 5 Excel Histogram menu

This will create the frequency distribution and histogram



Figure 6 Excel Histogram - frequency distribution solution



Figure 7 Excel Histogram solution

As we can see from the figure, the data is relatively symmetric.

Conduct a chi-square goodness-of-fit test to test for normality



Figure 8 Frequency distribution

Undertake chi-square goodness-of-fit test



Figure 9 Chi-square goodness-of-fit test

If we look at the expected frequencies, we observe that the first cell expected frequency is < 5. This should be merged with the second cell and re-calculate observed frequencies, expected frequencies, then undertake the chi-square test as illustrated in Figure 10.



Figure 10 Chi-square goodness-of-fit result with Cell 1 and 2 merged.

From Excel, chi-square test statistic= 2.35 < critical chi-square statistic = 7.81, accept the null hypothesis. Conclude that the data values come from a normal distribution with population mean = 25 and population standard deviation = 8. What happens if you change the BIN values?

**SPSS solution**

Enter data into SPSS and save the data file (first 5 data values illustrated)



Figure 11 SPSS data file – saved as ‘normality test.sav’.

Analyze > Descriptive Statistics > Explore

Transfer DataValues into Dependent List box

Click on Plots and choose Normality plots with tests



Figure 12 SPSS Explore menu

Click OK

SPSS will create an output file as illustrated below – saved file as ‘normality test.spv’.

This will output descriptive statistics including two normality test statistics – we will use the Shapiro-Wilks test. Figures 13 and 14 illustrates this statistic from the SPSS output solution.



Figure 13 SPSS Descriptives



Figure 14 SPSS Tests of Normality

From SPSS, Shapiro-Wilk statistic = 0.975 with a corresponding p-value = 0.446. Given the p-value = 0.446 > 0.05, we accept the null hypothesis. Conclude, the data values are normally distributed. Note: (a) the SPSS solution, the sample mean and sample standard deviation will have been calculated and used to undertake the goodness-of-fit test, and (b) the Excel solution, I stated the population values were known.