KOLMOGOROV–SMIRNOV ONE-SAMPLE TEST

OBJECTIVES

In this lecture, you will learn the following items:

- How to perform a Kolmogorov–Smirnov one-sample test to determine if a data sample meets acceptable levels of normality.

- How to use SPSS to perform a Kolmogorov–Smirnov one-sample test to determine if a data sample meets acceptable levels of normality.

COMPUTING THE KOLMOGOROV–SMIRNOV ONE-SAMPLE TEST

The Kolmogorov–Smirnov one-sample test is a procedure to examine the agreement between two sets of values. For our purposes, the two sets of values compared are an observed frequency distribution based on a randomly collected sample and an empirical frequency distribution based on the sample's population. Furthermore, the observed sample is examined for normality when the empirical frequency distribution is based on a normal distribution.

The Kolmogorov–Smirnov one-sample test compares two cumulative frequency distributions. A cumulative frequency distribution is useful for finding the number of observations above or below a particular value in a data sample. It is calculated by taking a given frequency and adding all the preceding frequencies in the list.

In other words, it is like making a running total of the frequencies in a distribution. Creating cumulative frequency distributions of the observed and empirical frequency distributions allow us to find the point at which these two distributions show the largest divergence.

Then, the test uses the largest divergence to identify a two-tailed probability estimate p to determine if the samples are statistically similar or different.

To perform the Kolmogorov–Smirnov one-sample test, we begin by determining the relative empirical frequency distribution $\hat{f}_{x_i}$ based on the observed sample. This relative empirical frequency distribution will approximate a normal distribution since we are examining our observed values for sample normality. First, calculate the observed frequency distribution's midpoint M and standard deviation s. The midpoint and standard deviation are found using Formula

$$M = (x_{\max} + x_{\min}) \div 2$$

where $x_{\max}$ is the largest value in the sample and $x_{\min}$ is the smallest value in the sample, and

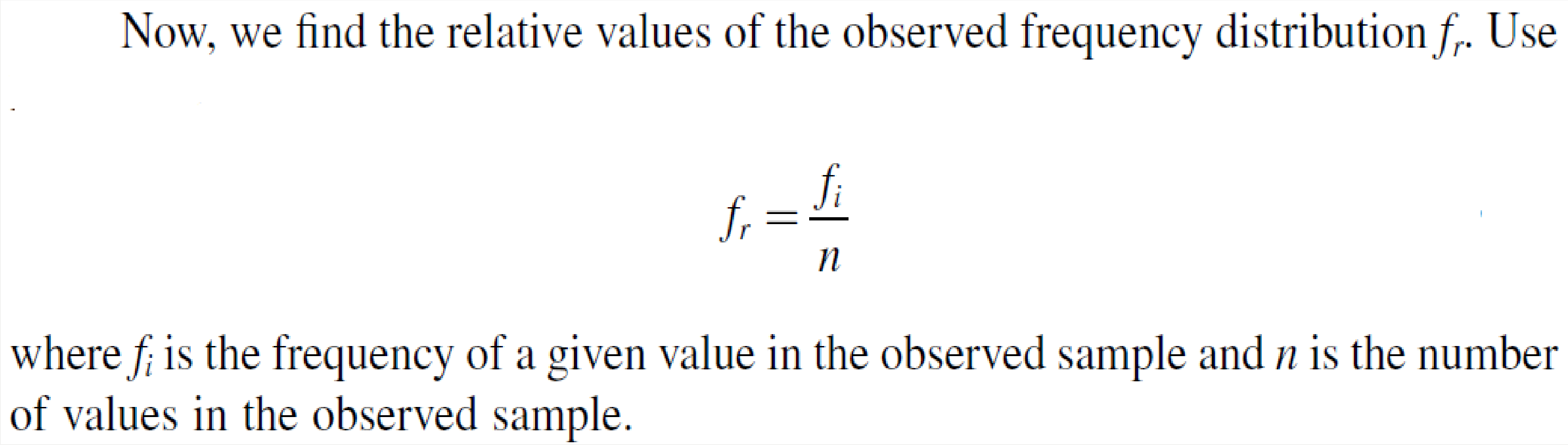
$$s = \sqrt{\frac{\sum (f_i x_i^2) - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}}$$

where x_i is a given value in the observed sample, f_i is the frequency of a given value in the observed sample, and n is the number of values in the observed sample.

Next, use the midpoint and standard deviation to calculate the z-scores for the sample values x_i, using the following formula:

$$z = \left| \frac{x_i - M}{s} \right|$$

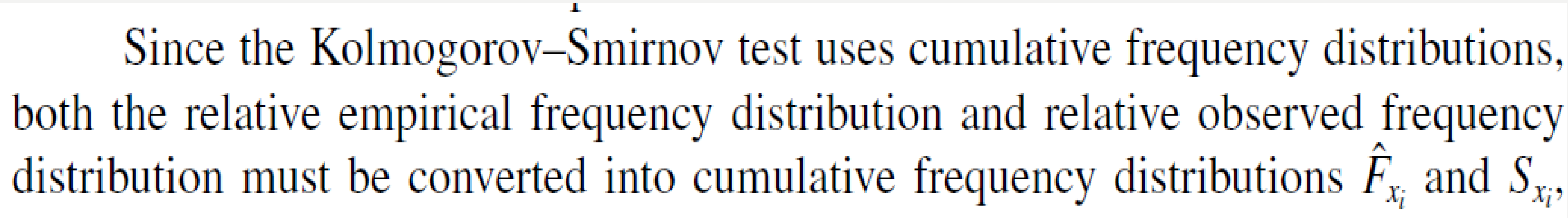
Use those z-scores and standard normal distribution table to determine the probability associated with each sample value, $\hat{p}_{x_i}$. These p-values are the relative frequencies of the empirical frequency distribution $\hat{f}_r$.



Now, we find the relative values of the observed frequency distribution f_r. Use

$$f_r = \frac{f_i}{n}$$

where f_i is the frequency of a given value in the observed sample and n is the number of values in the observed sample.



Since the Kolmogorov–Smirnov test uses cumulative frequency distributions, both the relative empirical frequency distribution and relative observed frequency distribution must be converted into cumulative frequency distributions $\hat{F}_{x_i}$ and S_{x_i},

Use the following formulas to find the absolute value divergence $\tilde{D}$ and D between the cumulative frequency distributions:

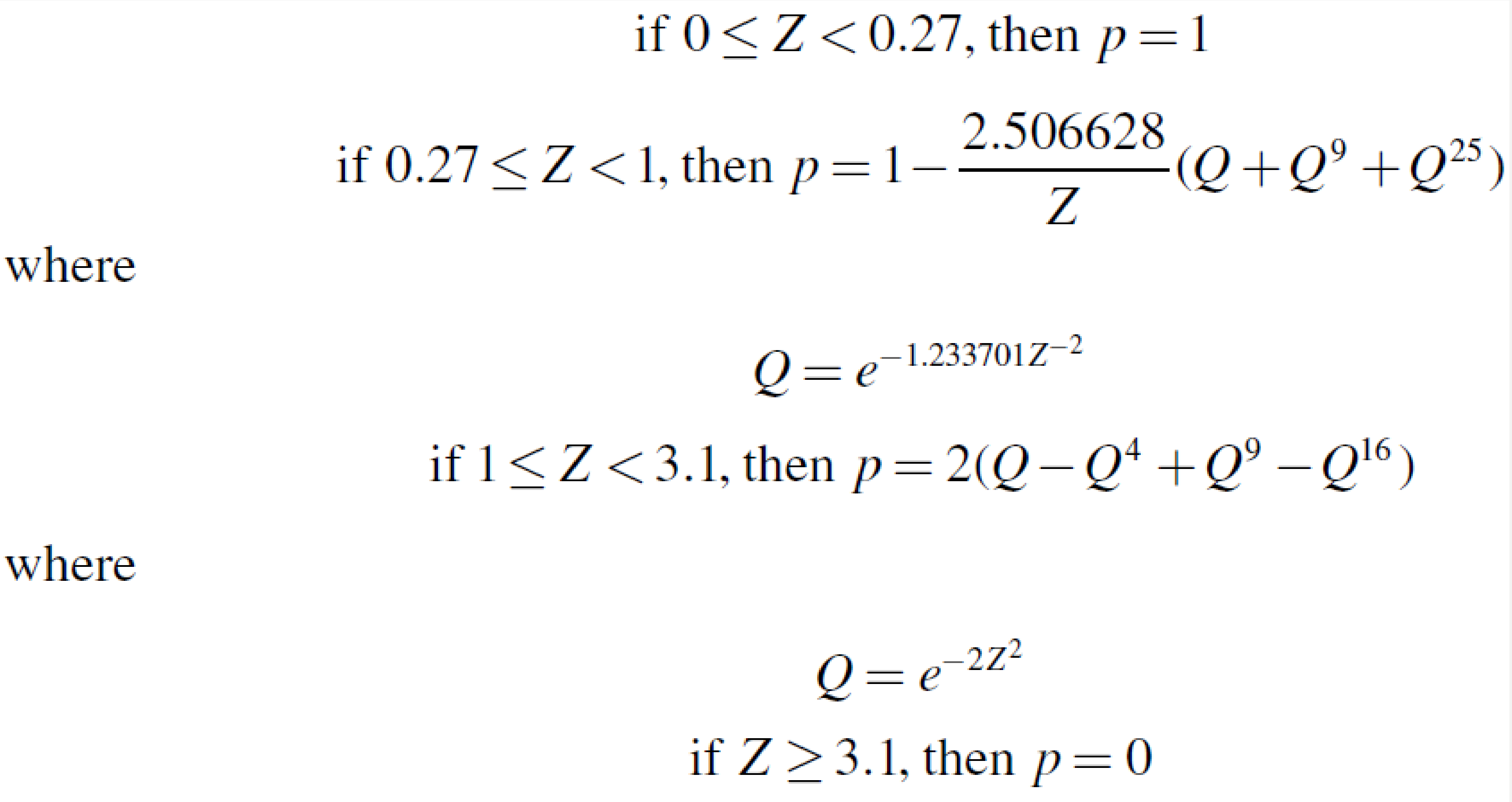
$$\tilde{D} = \left| \hat{F}_{x_i} - S_{x_i} \right|$$

$$D = \left| \hat{F}_{x_i} - S_{x_{i-1}} \right|$$

Use the largest divergence with the next formula to calculate the Kolmogorov–Smirnov test statistic Z:

$$Z = \sqrt{n} \max\left(|D|, |\tilde{D}| \right)$$

Then, use the Kolmogorov–Smirnov test statistic Z and the next Smirnov formulas to find the two-tailed probability estimate p:



$$\text{if } 0 \le Z < 0.27, \text{ then } p = 1$$

$$\text{if } 0.27 \le Z < 1, \text{ then } p = 1 - \frac{2.506628}{Z}(Q + Q^9 + Q^{25})$$

where

$$Q = e^{-1.233701Z^{-2}}$$

$$\text{if } 1 \le Z < 3.1, \text{ then } p = 2(Q - Q^4 + Q^9 - Q^{16})$$

where

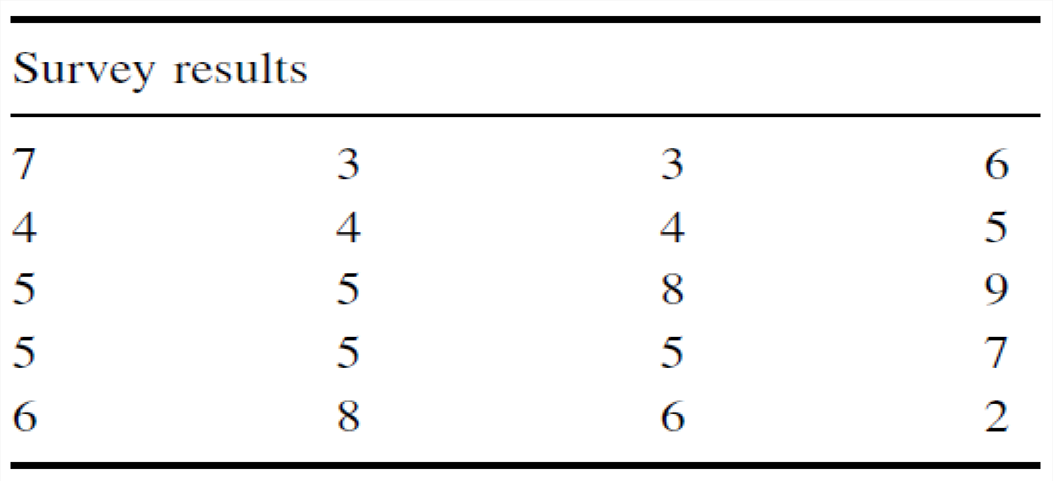
$$Q = e^{-2Z^2}$$

$$\text{if } Z \ge 3.1, \text{ then } p = 0$$

A p-value that exceeds the level of risk associated with the null hypothesis indicates that the observed sample approximates the empirical sample. Since our empirical distributions approximated a normal distribution, we can state that our observed sample is sufficiently normal for parametric statistics. Conversely, a p-value that is smaller than the level of risk indicates an observed sample that is not sufficiently normal for parametric statistics. The nonparametric statistical tests in this book are useful if a sample lacks normality.

Sample Kolmogorov–Smirnov One-Sample Test

A department store has decided to evaluate customer satisfaction. As part of a pilot study, the store provides customers with a survey to rate employee friendliness. The survey uses a scale of 1–10 and its developer indicates that the scores should conform to a normal distribution. Use the Kolmogorov–Smirnov one-sample test to decide if the sample of customers surveyed responded with scores approximately matching a normal distribution. The survey results are shown in the following table.



Survey results			
7	3	3	6
4	4	4	5
5	5	8	9
5	5	5	7
6	8	6	2

1 State the Null and Research Hypotheses

The null hypothesis states that the observed sample has an approximately normal distribution. The research hypothesis states that the observed sample does not approximately resemble a normal distribution.

The null hypothesis is

***H_0*:** There is no difference between the observed distribution of survey scores and a normally distributed empirical sample.

The research hypothesis is

H_A: There is a difference between the observed distribution of survey scores
and a normally distributed empirical sample.

2. Set the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

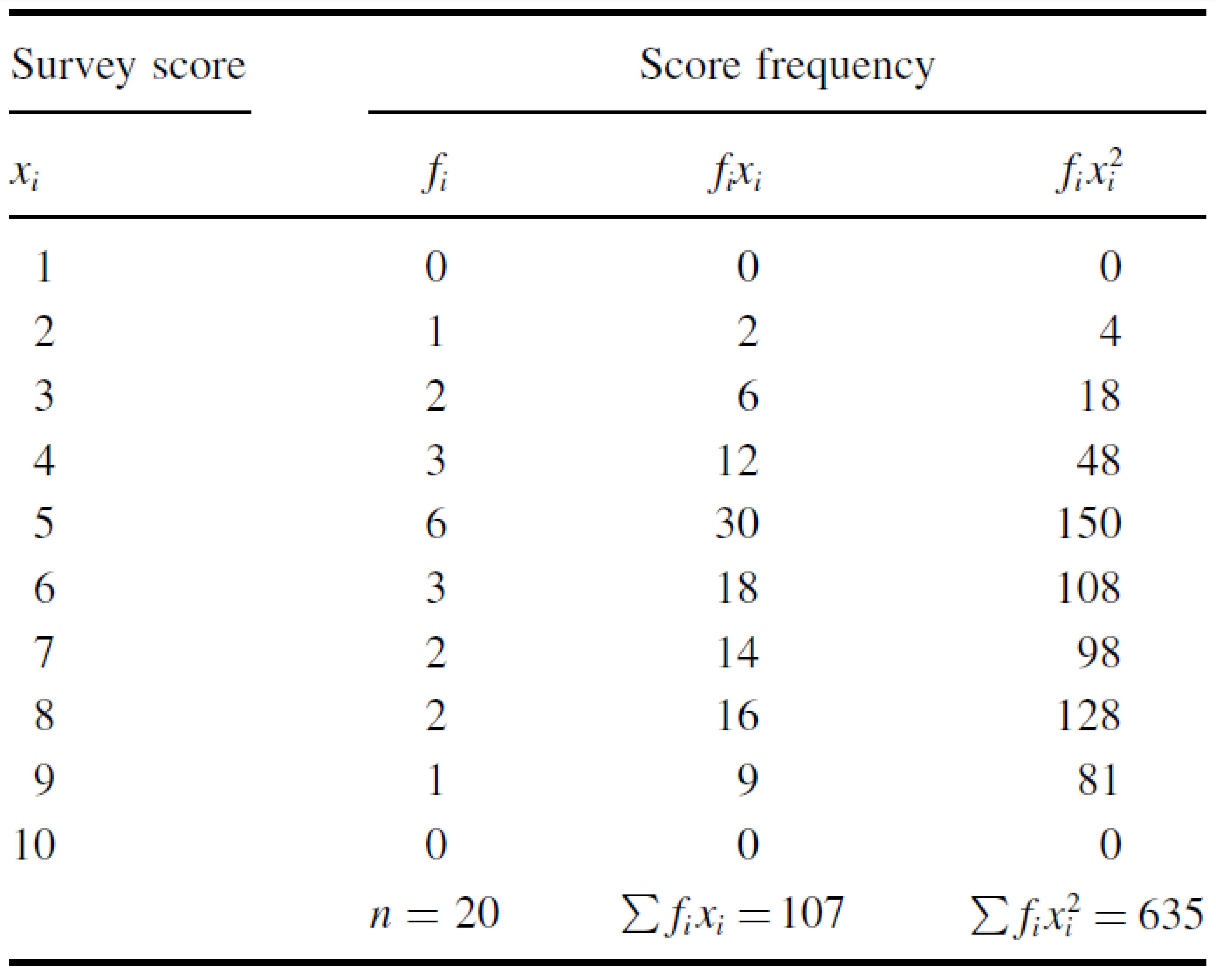
The level of risk, also called an alpha (α), is frequently set at 0.05. We will use an $\alpha = 0.05$ in our example. In other words, there is a 95% chance that any observed statistical difference will be real and not due to chance.

3 Choose the Appropriate Test Statistic

We are seeking to compare our observed sample against a normally distributed empirical sample. The Kolmogorov–Smirnov one-sample test will provide this comparison.

4 Compute the Test Statistic

First, determine the midpoint and standard deviation for the observed sample. The following table helps to manage the summations for this process.



Survey score	Score frequency		
x_i	f_i	$f_i x_i$	$f_i x_i^2$
1	0	0	0
2	1	2	4
3	2	6	18
4	3	12	48
5	6	30	150
6	3	18	108
7	2	14	98
8	2	16	128
9	1	9	81
10	0	0	0
	$n = 20$	$\sum f_i x_i = 107$	$\sum f_i x_i^2 = 635$

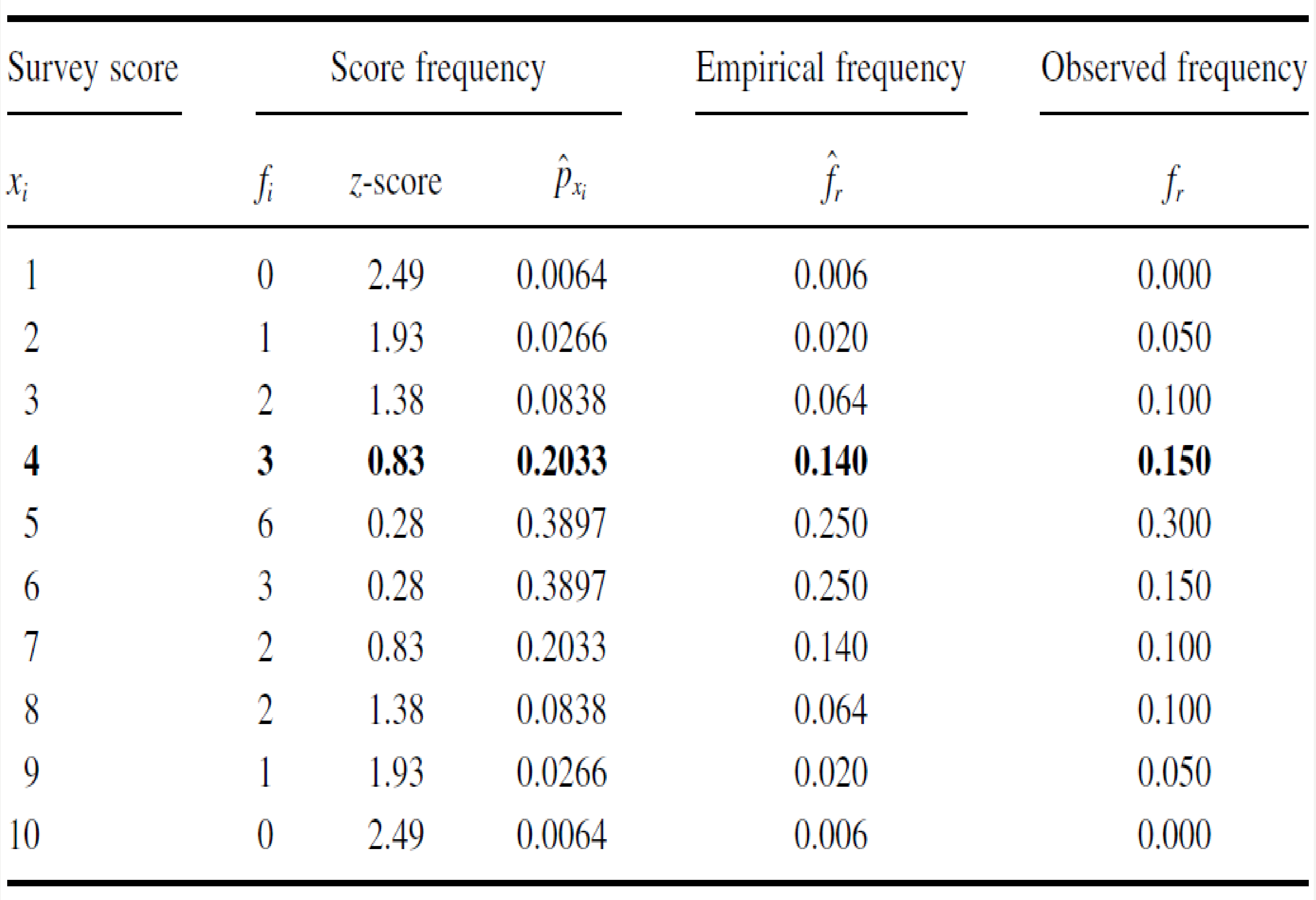
Use the following formula to find the midpoint:

$$\begin{aligned} M &= (x_{\max} + x_{\min}) \div 2 \\ &= (9+2) \div 2 \\ M &= 5.5 \end{aligned}$$

Then, use following formula to find the standard deviation:

$$\begin{aligned} s &= \sqrt{\frac{\sum (f_i x_i^2) - \frac{\left(\sum f_i x_i\right)^2}{n}}{n-1}} \\ &= \sqrt{\frac{635 - \frac{107^2}{20}}{20-1}} \\ s &= 1.81 \end{aligned}$$

Now, determine the z-scores, empirical relative frequencies, and observed relative frequencies for each score value.



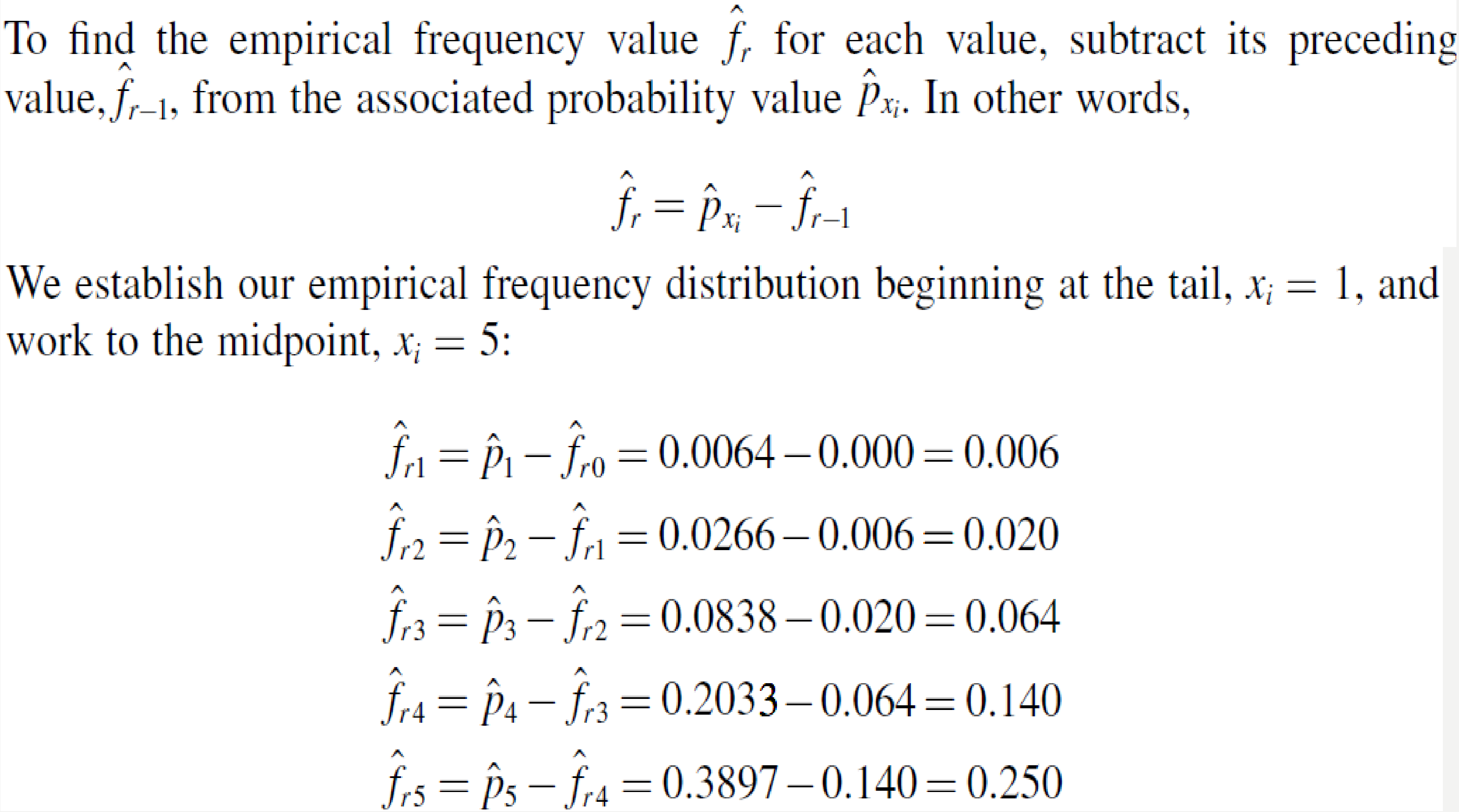
Survey score	Score frequency			Empirical frequency	Observed frequency
x_i	f_i	z-score	$\hat{p}_{x_i}$	$\hat{f}_r$	f_r
1	0	2.49	0.0064	0.006	0.000
2	1	1.93	0.0266	0.020	0.050
3	2	1.38	0.0838	0.064	0.100
4	**3**	**0.83**	**0.2033**	**0.140**	**0.150**
5	6	0.28	0.3897	0.250	0.300
6	3	0.28	0.3897	0.250	0.150
7	2	0.83	0.2033	0.140	0.100
8	2	1.38	0.0838	0.064	0.100
9	1	1.93	0.0266	0.020	0.050
10	0	2.49	0.0064	0.006	0.000

We will provide a sample calculation for survey score = 4 as seen in above table. Use the following formula to calculate the z-scores:

$$\begin{aligned} z &= \left|\frac{x_i - M}{s}\right| \\ &= \left|\frac{4 - 5.5}{1.81}\right| \\ z &= 0.83 \end{aligned}$$

Use each z-score and standard normal distribution table to determine the probability associated with the each value

$$\hat{p}_4 = 0.2033$$



To find the empirical frequency value $\hat{f}_r$ for each value, subtract its preceding value, $\hat{f}_{r-1}$, from the associated probability value $\hat{p}_{x_i}$. In other words,

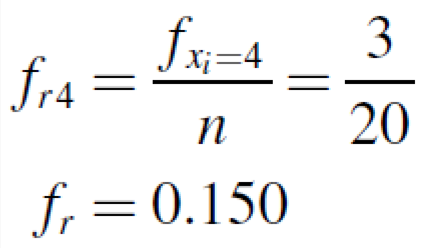
$$\hat{f}_r = \hat{p}_{x_i} - \hat{f}_{r-1}$$

We establish our empirical frequency distribution beginning at the tail, $x_i = 1$, and work to the midpoint, $x_i = 5$:

$$\begin{aligned}
\hat{f}_{r1} &= \hat{p}_1 - \hat{f}_{r0} = 0.0064 - 0.000 = 0.006 \\
\hat{f}_{r2} &= \hat{p}_2 - \hat{f}_{r1} = 0.0266 - 0.006 = 0.020 \\
\hat{f}_{r3} &= \hat{p}_3 - \hat{f}_{r2} = 0.0838 - 0.020 = 0.064 \\
\hat{f}_{r4} &= \hat{p}_4 - \hat{f}_{r3} = 0.2033 - 0.064 = 0.140 \\
\hat{f}_{r5} &= \hat{p}_5 - \hat{f}_{r4} = 0.3897 - 0.140 = 0.250
\end{aligned}$$

Our empirical frequency distribution is based on a normal distribution, which is symmetrical. Therefore, we can complete our empirical frequency distribution by basing the remaining values on a symmetrical distribution. Those values are in above Table

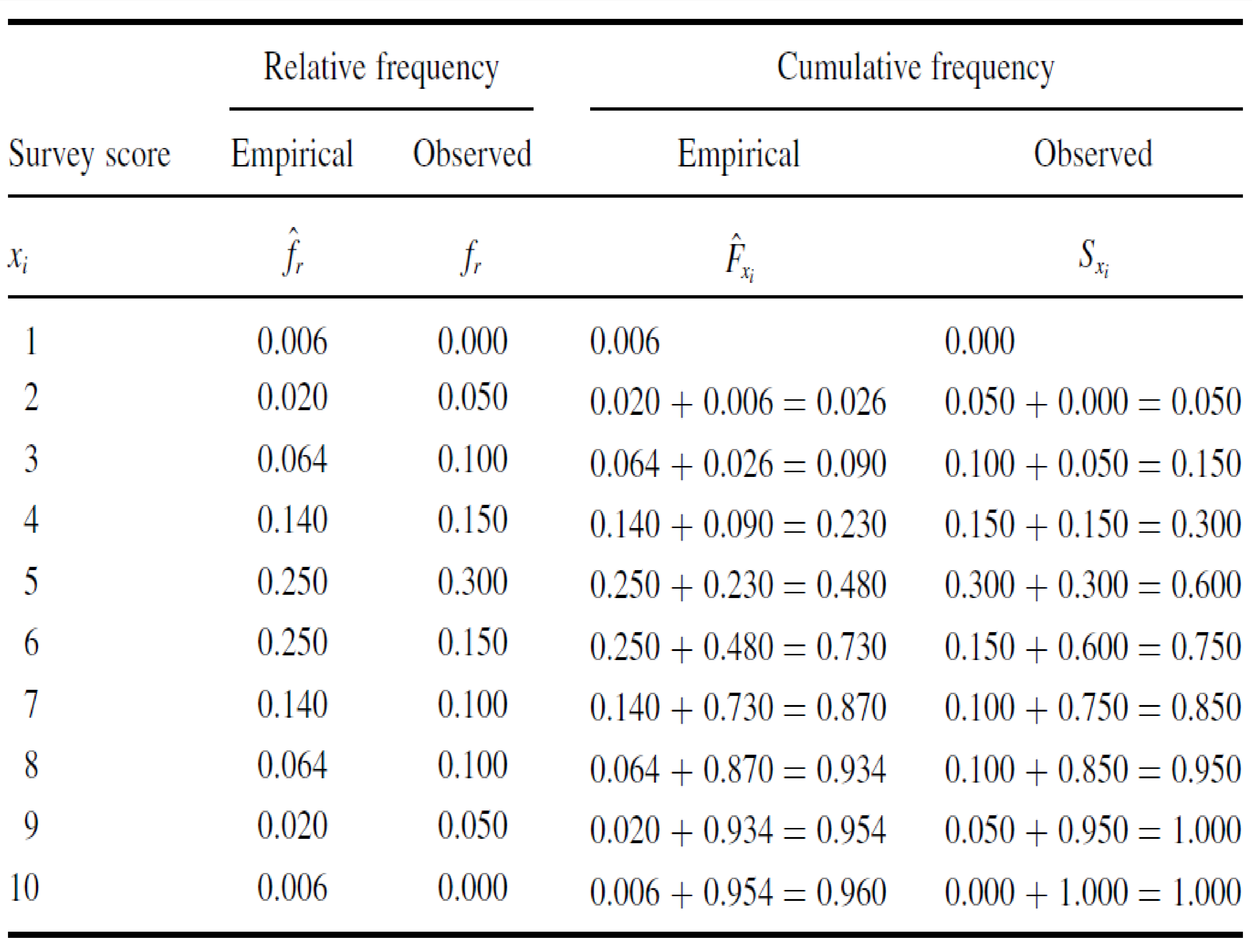
Now, we find the values of the *observed frequency* distribution f_r . We provide a sample calculation with survey result = 4. That survey value occurs three times:



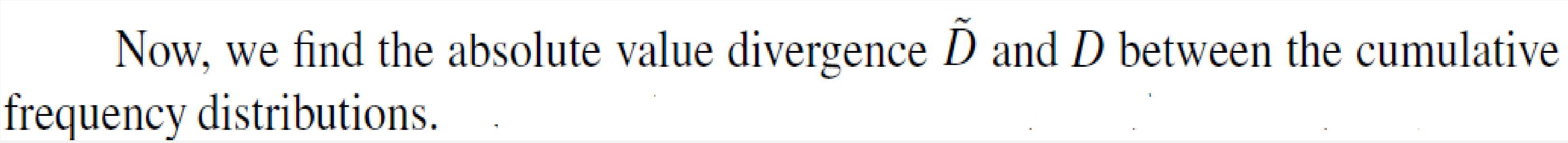
$$f_{r4} = \frac{f_{x_i=4}}{n} = \frac{3}{20}$$

$$f_r = 0.150$$

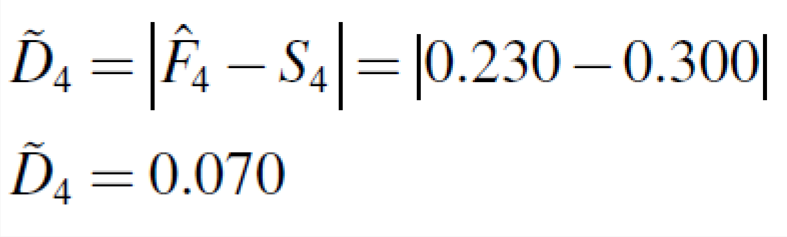
Next, we create cumulative frequency distributions using the empirical and observed frequency distributions. A cumulative frequency distribution is created by taking a frequency and adding all the preceding values. We demonstrate this in next table.



	Relative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Empirical	Observed
x_i	$\hat{f}_r$	f_r	$\hat{F}_{x_i}$	S_{x_i}
1	0.006	0.000	0.006	0.000
2	0.020	0.050	$0.020 + 0.006 = 0.026$	$0.050 + 0.000 = 0.050$
3	0.064	0.100	$0.064 + 0.026 = 0.090$	$0.100 + 0.050 = 0.150$
4	0.140	0.150	$0.140 + 0.090 = 0.230$	$0.150 + 0.150 = 0.300$
5	0.250	0.300	$0.250 + 0.230 = 0.480$	$0.300 + 0.300 = 0.600$
6	0.250	0.150	$0.250 + 0.480 = 0.730$	$0.150 + 0.600 = 0.750$
7	0.140	0.100	$0.140 + 0.730 = 0.870$	$0.100 + 0.750 = 0.850$
8	0.064	0.100	$0.064 + 0.870 = 0.934$	$0.100 + 0.850 = 0.950$
9	0.020	0.050	$0.020 + 0.934 = 0.954$	$0.050 + 0.950 = 1.000$
10	0.006	0.000	$0.006 + 0.954 = 0.960$	$0.000 + 1.000 = 1.000$



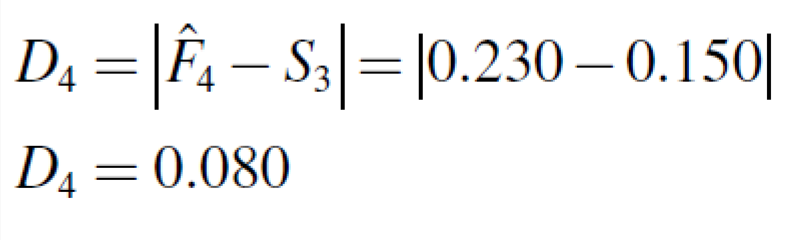
Now, we find the absolute value divergence $\tilde{D}$ and D between the cumulative frequency distributions.



$$\tilde{D}_4 = \left|\hat{F}_4 - S_4\right| = |0.230 - 0.300|$$

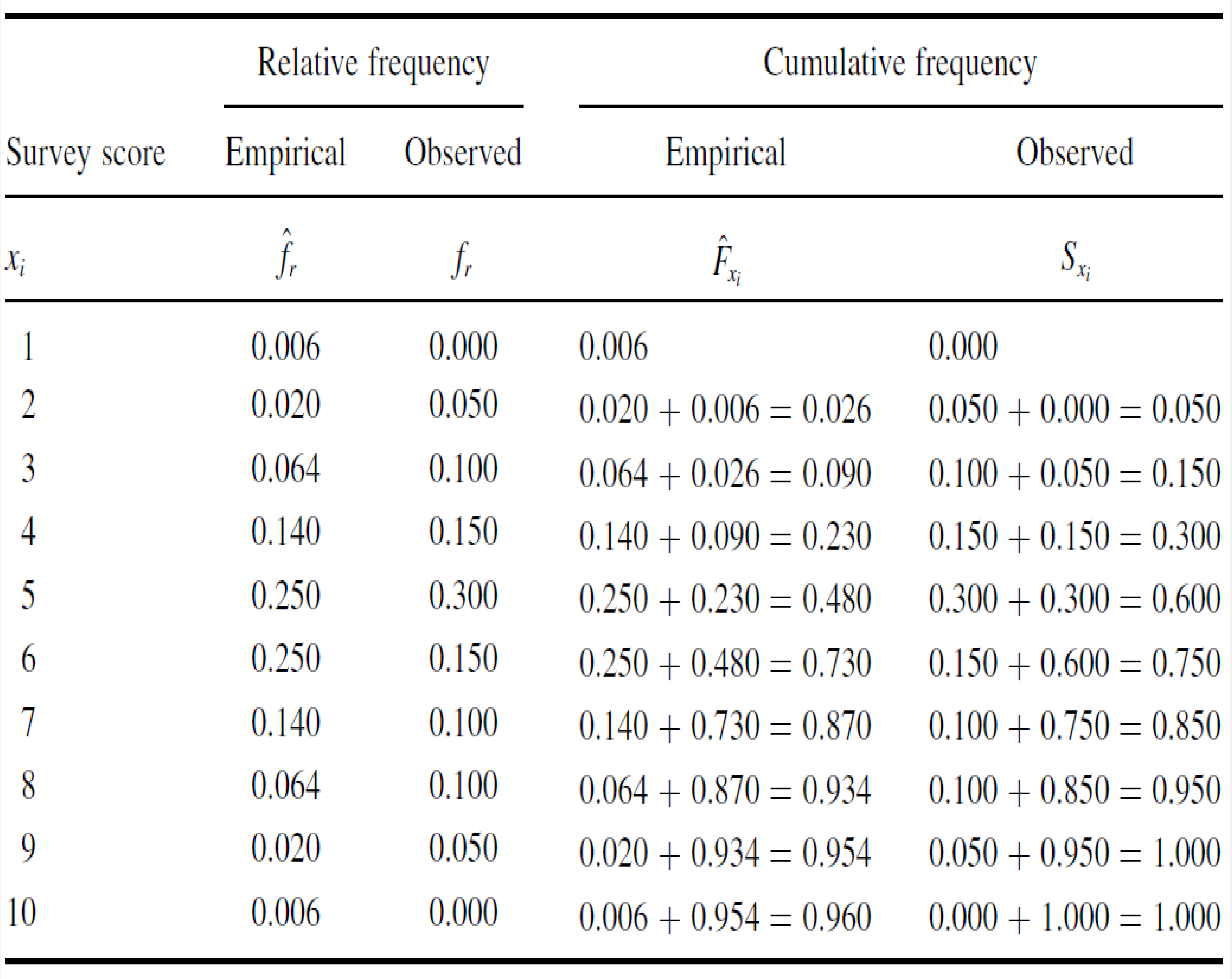
$$\tilde{D}_4 = 0.070$$

and

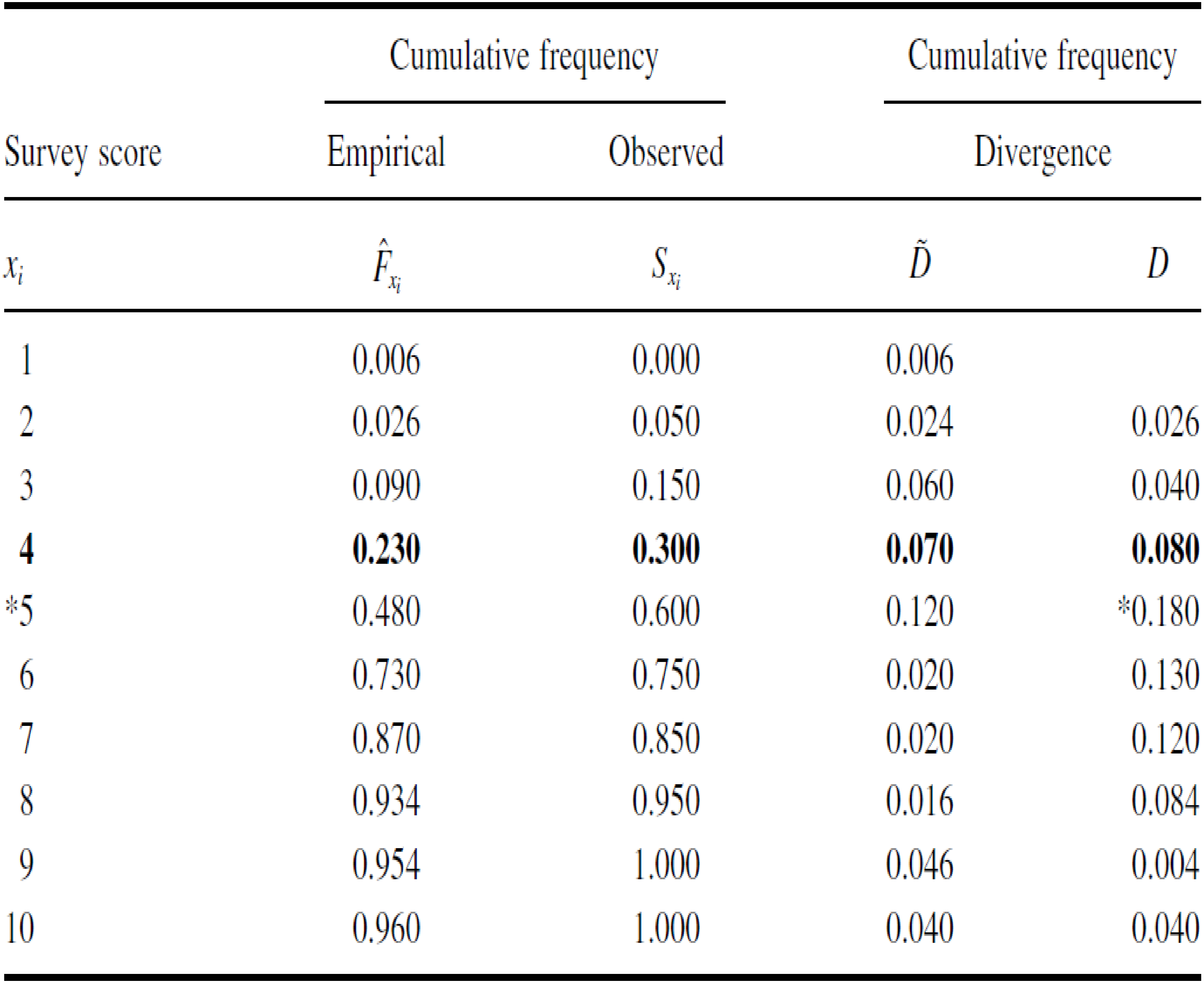


$$D_4 = \left|\hat{F}_4 - S_3\right| = |0.230 - 0.150|$$

$$D_4 = 0.080$$



	Relative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Empirical	Observed
x_i	$\hat{f}_r$	f_r	$\hat{F}_{x_i}$	S_{x_i}
1	0.006	0.000	0.006	0.000
2	0.020	0.050	$0.020 + 0.006 = 0.026$	$0.050 + 0.000 = 0.050$
3	0.064	0.100	$0.064 + 0.026 = 0.090$	$0.100 + 0.050 = 0.150$
4	0.140	0.150	$0.140 + 0.090 = 0.230$	$0.150 + 0.150 = 0.300$
5	0.250	0.300	$0.250 + 0.230 = 0.480$	$0.300 + 0.300 = 0.600$
6	0.250	0.150	$0.250 + 0.480 = 0.730$	$0.150 + 0.600 = 0.750$
7	0.140	0.100	$0.140 + 0.730 = 0.870$	$0.100 + 0.750 = 0.850$
8	0.064	0.100	$0.064 + 0.870 = 0.934$	$0.100 + 0.850 = 0.950$
9	0.020	0.050	$0.020 + 0.934 = 0.954$	$0.050 + 0.950 = 1.000$
10	0.006	0.000	$0.006 + 0.954 = 0.960$	$0.000 + 1.000 = 1.000$



	Cumulative frequency		Cumulative frequency	
Survey score	Empirical	Observed	Divergence	
x_i	$\hat{F}_{x_i}$	S_{x_i}	$\tilde{D}$	D
1	0.006	0.000	0.006	
2	0.026	0.050	0.024	0.026
3	0.090	0.150	0.060	0.040
4	**0.230**	**0.300**	**0.070**	**0.080**
*5	0.480	0.600	0.120	*0.180
6	0.730	0.750	0.020	0.130
7	0.870	0.850	0.020	0.120
8	0.934	0.950	0.016	0.084
9	0.954	1.000	0.046	0.004
10	0.960	1.000	0.040	0.040

To find the test statistic Z, use the largest value from $\tilde{D}$ and D. The last above table has an asterisk next to the largest divergence. That value is located at survey value = 5.

It is $\max\left(|D|, |\tilde{D}|\right) = 0.180$:

$$\begin{aligned} Z &= \sqrt{n}\max\left(|D|, |\tilde{D}|\right) \\ &= \sqrt{20}(0.180) \\ Z &= 0.805 \end{aligned}$$

5 Determine the p-Value Associated with the Test Statistic

The Kolmogorov–Smirnov test statistic Z and the Smirnov (1948) formulas are used to find the two-tailed probability estimate p.

Since 0.27 Z < 1, we have:

$$
\begin{aligned}
Q &= e^{-1.233701Z^{-2}} \\
&= e^{-(1.233701)(0.805)^{-2}} \\
Q &= 0.149
\end{aligned}
$$

and

$$
\begin{aligned}
p &= 1 - \frac{2.506628}{Z}(Q + Q^{9} + Q^{25}) \\
&= 1 - \frac{2.506628}{0.805}(0.149 + 0.149^{9} + 0.149^{25}) \\
p &= 0.536
\end{aligned}
$$

6. Compare the p-Value with the Level of Risk (or the Level of Significance) Associated with the Null Hypothesis

The critical value for rejecting the null hypothesis is $\alpha = 0.05$ and the obtained p-value is p = 0.536. If the critical value is greater than the obtained value, we must reject the null hypothesis. If the critical value is less than the obtained p-value, we must not reject the null hypothesis. Since the critical value is less than the obtained value ($0.05 < 0.536$), we do not reject the null hypothesis.

7 Interpret the Results

We did not reject the null hypothesis, suggesting the customers’ survey ratings of employee friendliness sufficiently resembled a normal distribution. This means that a parametric statistical procedure may be used with this sample.

8 Reporting the Results

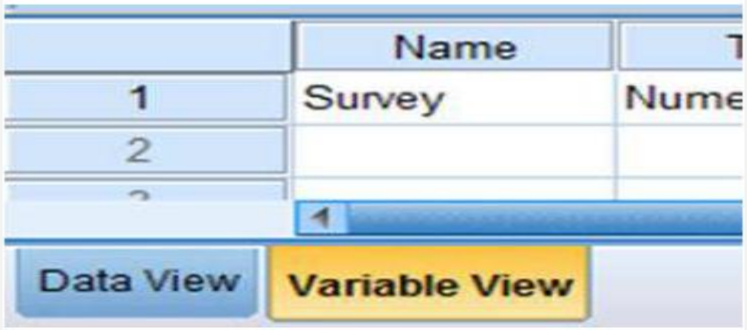
When reporting the results from the Kolmogorov–Smirnov one-sample test, we include the test statistic (D), the degrees of freedom (which equals the sample size), and the p-value in terms of the level of risk . Based on our analysis, the sample of customers is approximately normal, where D(20) = 0.180, $p > 0.05$.

Performing the Kolmogorov–Smirnov One-Sample Test Using SPSS

We will analyze the data from the example earlier using SPSS.

1 Define Your Variables

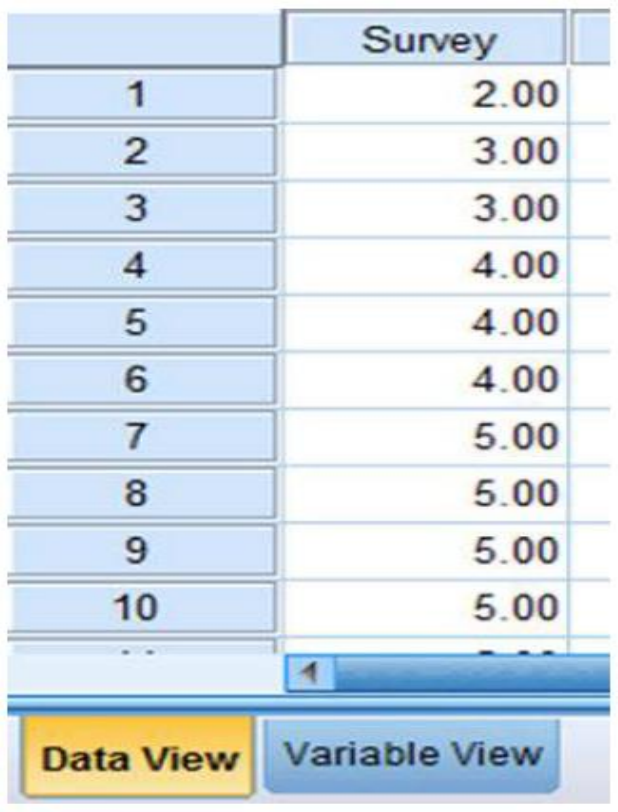
First, click the “Variable View” tab at the bottom of your screen. Then, type the names of your variables in the “Name” column. As shown in Figure 2.13, the variable is called “Survey.”



	Name	T
1	Survey	Nume
2		

2 Type in Your Values

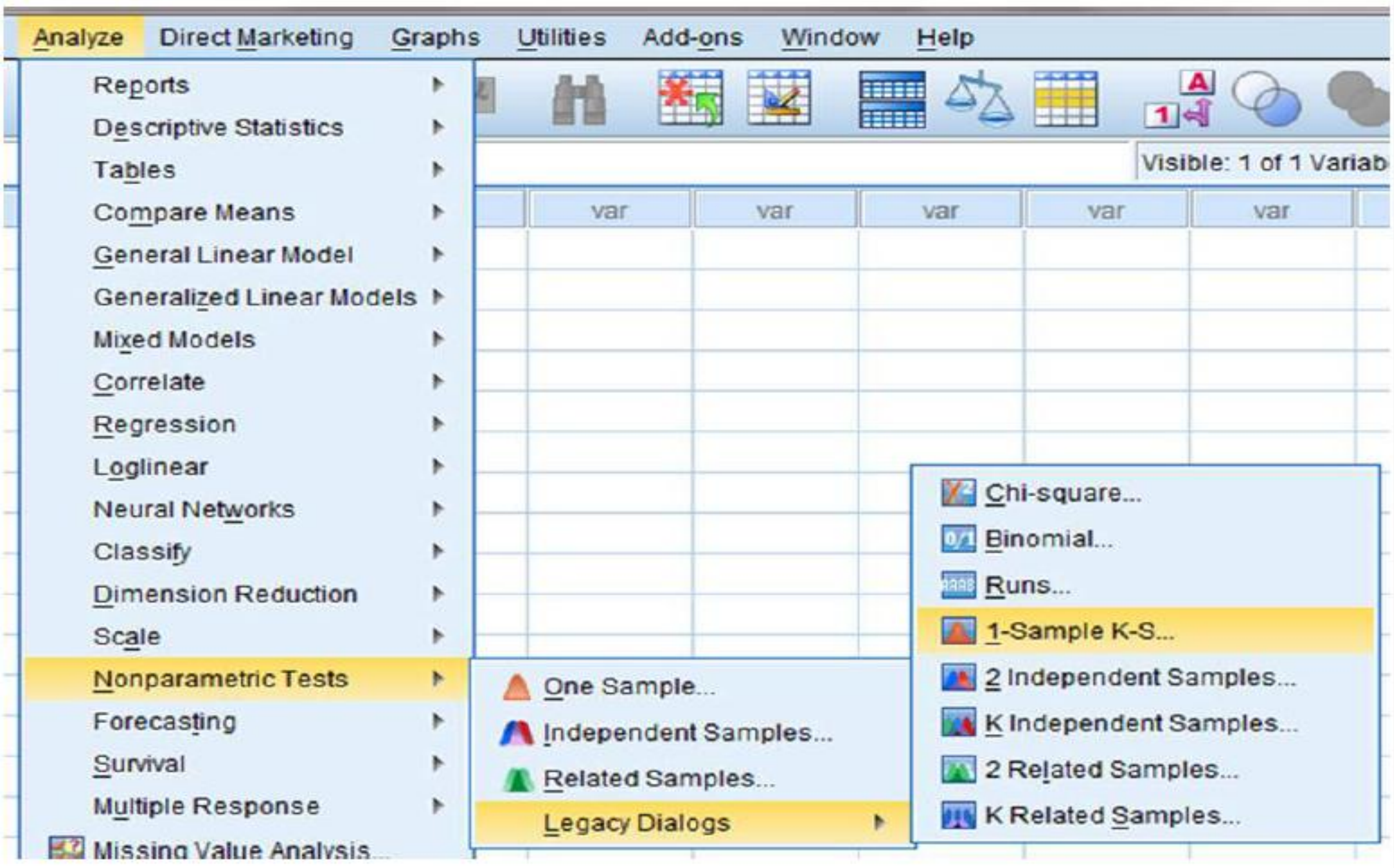
Click the “Data View” tab at the bottom of your screen. Type your sample values in the “Survey” column as shown in the following figure:



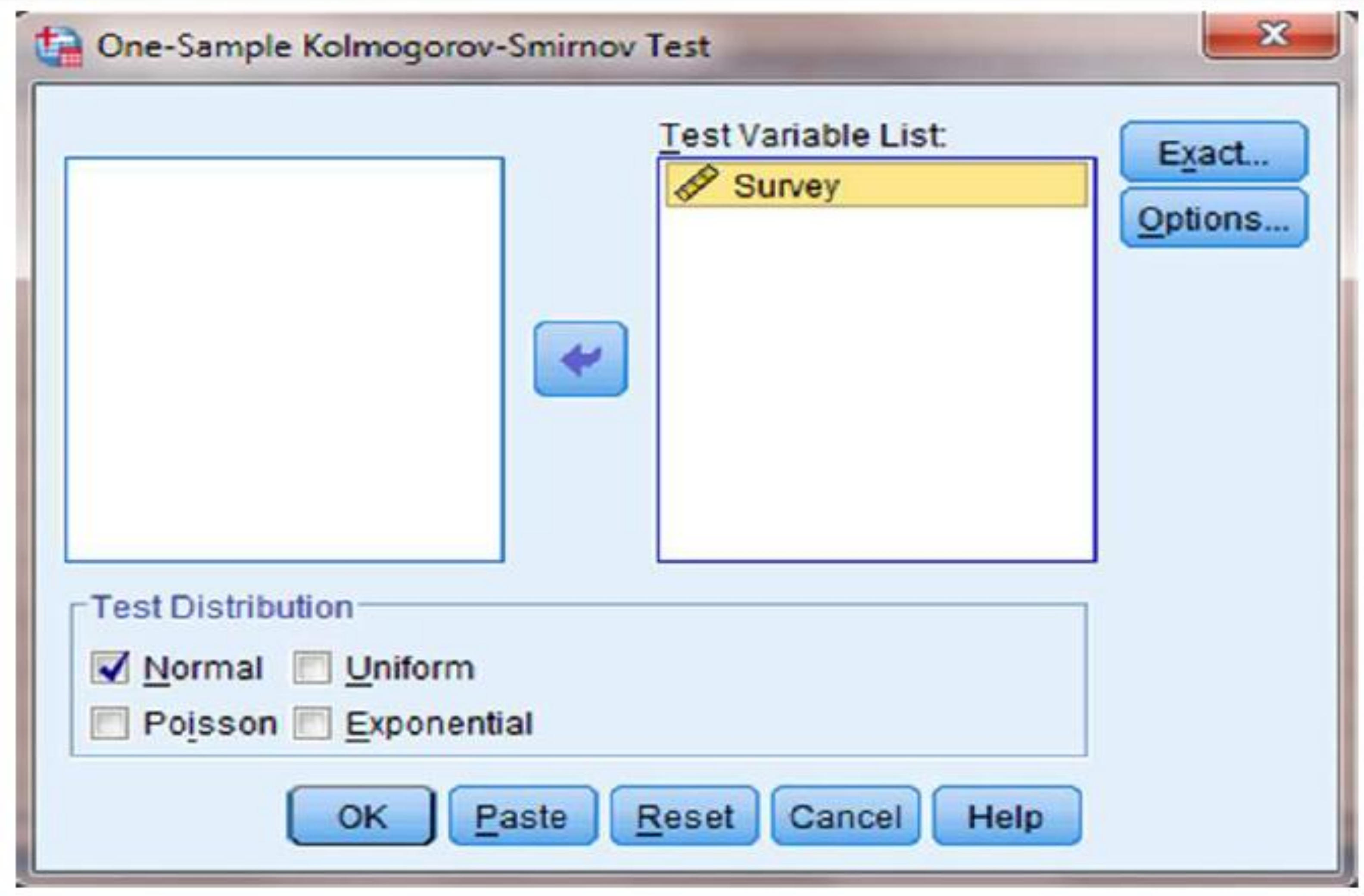
	Survey
1	2.00
2	3.00
3	3.00
4	4.00
5	4.00
6	4.00
7	5.00
8	5.00
9	5.00
10	5.00

3 Analyze Your Data

As shown in following figure, use the pull-down menus to choose “Analyze,” “Nonparametric Tests,” “Legacy Dialogs,” and “1-Sample KS . . .”



Use the arrow button to place your variable with your data values in the box labeled “Test Variable List:” as shown in following figure. Finally, click “OK” to perform the analysis.

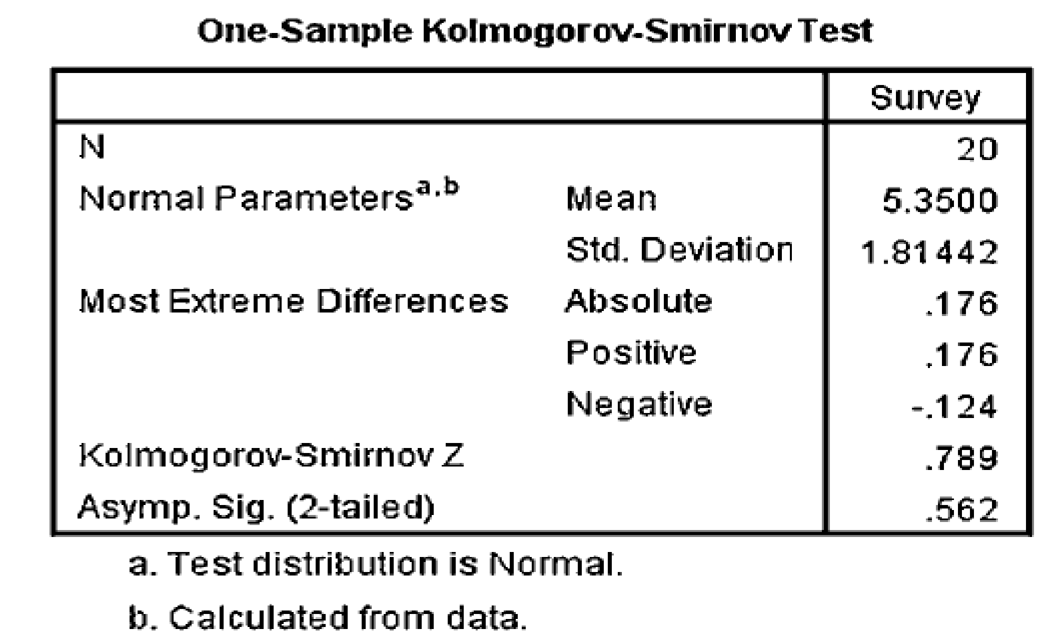


One-Sample Kolmogorov-Smirnov Test

Test Variable List:
Survey

Test Distribution
☑ Normal ☐ Uniform
☐ Poisson ☐ Exponential

7 Interpret the Results from the SPSS Output Window SPSS



One-Sample Kolmogorov-Smirnov Test

		Survey
N		20
Normal Parameters[a,b]	Mean	5.3500
	Std. Deviation	1.81442
Most Extreme Differences	Absolute	.176
	Positive	.176
	Negative	-.124
Kolmogorov-Smirnov Z		.789
Asymp. Sig. (2-tailed)		.562

a. Test distribution is Normal.

b. Calculated from data.

SPSS OUTPUT 2.2

Output 2.2 provides the most extreme difference (D = 0.176), Kolmogorov–Smirnov Z-test statistic (Z = 0.789), and the significance (p = 0.562). Based on the results from SPSS, the p-value exceeds the level of risk associated with the null hypothesis (α= 0.05). Therefore, we do not reject the null hypothesis. In other words, the sample distribution is sufficiently normal.

Remark: Differences between the values from the sample problem earlier and the SPSS output are likely due to value precision and computational round off errors.

SUMMARY

Parametric statistical tests, such as the t-test and one-way analysis of variance, are based on particular assumptions or parameters. Therefore, it is important that you examine collected data for its approximation to a normal distribution. Upon doing that, you can consider whether you will use a parametric or nonparametric test for analyzing your data.

In Lectures 3 & 4, we presented three quantitative measures of sample normality. First, we described how to examine a sample's kurtosis and skewness. Then, we described how to perform and interpret a Kolmogorov–Smirnov one-sample test.

In the following lectures, we will describe several non-parametric procedures for analyzing data samples that do not meet the assumptions needed for parametric statistical tests. In the chapter that follows, we will begin by describing a test for comparing two unrelated samples.