How to Easily Compare Data Sets in Excel Using the Kolmogorov-Smirnov Test

Authored by

stats writer

December 2, 2025

The Kolmogorov-Smirnov Test, often abbreviated as the K-S Test, is a powerful and widely utilized statistical method. It falls under the category of non-parametric tests, meaning it does not rely on rigid assumptions about the specific distribution of the population from which the sample is drawn. The primary purpose of the K-S Test is to assess whether a single sample comes from a specified theoretical probability distribution, or whether two independent samples originate from the same underlying distribution. When applied to a single sample, it helps determine if the data adheres to a standard distribution, such as the normal distribution.

The **Kolmogorov-Smirnov test** is fundamentally employed to ascertain whether a sample of data is consistent with a specified theoretical probability distribution. This comparison is vital for validating statistical assumptions, especially the assumption of normality.

This non-parametric test is widely used precisely because many powerful statistical tests and procedures--known as parametric methods--make the crucial assumption that the underlying data is normally distributed. If the assumption of normality is violated, the results derived from these tests may be invalid or misleading.

The following comprehensive, step-by-step example illustrates how to perform a one-sample Kolmogorov-Smirnov test for normality on a sample dataset directly within Microsoft Excel, focusing on the calculation of the D statistic and its proper interpretation against critical values.

Understanding the Theoretical Basis of the K-S Test

The core mathematical concept behind the Kolmogorov-Smirnov Test is the comparison between two cumulative distribution functions (CDFs). When testing a single sample against a hypothesized distribution, we first generate the empirical CDF. The empirical CDF is a step function representing the observed proportion of data points below a certain value (x). We then contrast this observed function with the theoretical CDF, which is a smooth curve based on the hypothesized distribution defined by the sample's estimated parameters (mean and standard deviation).

The test statistic D is mathematically defined as the supremum, or largest absolute vertical distance, between the empirical CDF and the theoretical CDF across all possible values in the dataset. A smaller D value indicates a better fit between the sample data and the hypothesized distribution. Since perfect matches are practically non-existent in real-world data, the test determines if the observed difference is large enough to be statistically significant, leading to the rejection of the null hypothesis.

While the standard K-S test is designed for situations where the distribution parameters are known beforehand, in the context of testing a sample for normal distribution, we must estimate the mean and standard deviation from the sample itself. This modification technically makes it the Lilliefors

test, which uses slightly different critical values. However, the calculation methodology in Excel remains the same, relying on these two derived parameters to define the theoretical normal curve.

Step 1: Data Preparation and Initial Entry

The initial crucial step involves organizing and entering the raw observations. These values should be placed into a single column, which we designate as Column A (cells A2 onwards). For this demonstration, we use a dataset containing **n = 20** observations. Although sorting the data can simplify manual checks, the Excel formulas provided below handle the sequential ranking required for the CDF calculation.

Next, we must establish a clear sequential ranking for the data points in Column B. This ranking (R) is essential for calculating the empirical cumulative probabilities. In cell B2, we enter the formula **B2:**

=ROW() - 1

. This formula is then copied down through B21, ensuring the first data row (row 2) receives the rank 1, the second rank 2, and so forth, reflecting the position of the observation.

This structured data layout allows for parallel calculation of the empirical and theoretical probabilities, setting the stage for the calculation of the D statistic.

	A	B	C	D	E	F
1	**Data**					
2	5.57					
3	8.32					
4	8.35					
5	8.74					
6	8.75					
7	9.38					
8	9.91					
9	9.96					
10	10.36					
11	10.65					
12	10.77					
13	10.97					
14	11.15					
15	11.18					
16	11.47					
17	11.64					
18	11.88					
19	12.24					
20	13.02					
21	13.19					
22						
23						
24						
25						

Step 2: Calculating Statistical Parameters

Before we can generate the theoretical cumulative distribution function, we must estimate the parameters of the hypothetical normal distribution from our sample data. These parameters--the mean (mu) and the standard deviation ($sigma$)--are the backbone of the theoretical curve and should be calculated in designated cells, such as J1 and J2, for stability.

The sample mean is calculated in cell J1 using the Excel **AVERAGE** function across the entire data range (A2:A21): **J1:**

=AVERAGE(A2:A21)

. This value represents the best central estimate of the population distribution.

The sample standard deviation is computed in cell J2. It is critical to use the sample standard deviation function, **STDEV.S**, as we are working with a sample: **J2:**

=STDEV.S(A2:A21)

. These two resulting values are then used as absolute references in the subsequent steps to define the precise normal curve against which our observations will be tested for goodness-of-fit.

Step 3: Generating Cumulative Distribution Functions (CDFs)

The K-S test requires calculating two forms of cumulative probability for each data point: the empirical (observed) CDF and the theoretical (expected) CDF.

The empirical CDF is calculated using the rank (B2) divided by the total count of observations (20). We calculate two bounds for the step function: the upper bound in C2 and the lower bound in D2. The upper bound is: **C2:**

=B2/COUNT($A$2:$A$21)

. The lower bound is: **D2:**

=(B2-1)/COUNT($A$2:$A$21)

. These two calculations define the height of the jump in the empirical CDF at each data point, allowing for a thorough check of the maximum deviation.

The expected cumulative probability, assuming normal distribution, is determined in Column F using the **NORM.DIST** function. This function takes the data point (A2), the calculated mean ($J$1), the calculated standard deviation ($J$2), and a final argument of **TRUE** (for cumulative probability): **F2:**

=NORM.DIST(A2, $J$1, $J$2, TRUE)

. This result in F2 represents the expected proportion of data that should fall below A2 if the sample were perfectly normally distributed.

Step 4: Determining the Maximum Absolute Difference (D Statistic)

The final calculation step involves determining the discrepancy between the observed and expected distributions. The Kolmogorov-Smirnov Test statistic, D, is the largest absolute difference between the theoretical CDF (F) and the empirical CDF (C and D).

In column G, we calculate the absolute difference between the theoretical probability (F2) and the lower bound empirical probability (D2). This calculation determines how far the data point's observed probability step function deviates from the theoretical curve at that point: **G2:**

=ABS(F2-D2)

. This formula must be copied down for all 20 rows.

The resulting D statistic, the maximum divergence, is found by calculating the maximum value in Column G. This is performed in cell J4 using the formula: **J4:**

=MAX(G2:G21)

. For this dataset, this calculation yields the **Maximum** D value of **0.10983**. This value is the metric we must compare against the critical threshold to determine statistical significance.

	A	B	C	D	E	F	G	H	I	J
1	**Data**	**Cumulative**	**Expected**	**(RANK-1)/N**	**NORM.S.INV**	**Actual**	**Difference**		**Mean**	10.375
2	5.57	1	0.05	0	-1.64485363	0.00426	0.004264		**Std. Dev.**	1.82672
3	8.32	2	0.1	0.05	-1.28155157	0.1303	0.080302			
4	8.35	3	0.15	0.1	-1.03643339	0.13381	0.033814		**Maximum**	0.10983
5	8.74	4	0.2	0.15	-0.84162123	0.18538	0.035381			
6	8.75	5	0.25	0.2	-0.67448975	0.18685	0.013152			
7	9.38	6	0.3	0.25	-0.52440051	0.29298	0.042983			
8	9.91	7	0.35	0.3	-0.38532047	0.39953	0.099534			
9	9.96	8	0.4	0.35	-0.2533471	0.41014	0.060141			
10	10.36	9	0.45	0.4	-0.12566135	0.49672	0.096724			
11	10.65	10	0.5	0.45	0	0.55983	0.109832			
12	10.77	11	0.55	0.5	0.125661347	0.5856	0.085597			
13	10.97	12	0.6	0.55	0.253347103	0.62768	0.077682			
14	11.15	13	0.65	0.6	0.385320466	0.66431	0.064311			
15	11.18	14	0.7	0.65	0.524400513	0.67028	0.020278			
16	11.47	15	0.75	0.7	0.67448975	0.72556	0.025558			
17	11.64	16	0.8	0.75	0.841621234	0.75569	0.005688			
18	11.88	17	0.85	0.8	1.036433389	0.795	0.005004			
19	12.24	18	0.9	0.85	1.281551566	0.84636	0.003638			
20	13.02	19	0.95	0.9	1.644853627	0.92618	0.026184			
21	13.19	20	1	0.95		0.93834	0.011657			
22										
23										
24										
25										

Here is the formula summary for the calculation array (starting from Row 2):

B2:

=ROW() - 1

(Assigns Rank)
C2:

=B2/COUNT($A$2:$A$21)

(Empirical CDF Upper Bound)
D2:

=(B2-1)/COUNT($A$2:$A$21)

(Empirical CDF Lower Bound)
E2:

=IF(C2<1, NORM.S.INV(C2),"")

(Z-Score calculation for visual checks)
F2:

=NORM.DIST(A2, $J$1, $J$2, TRUE)

(Theoretical Normal CDF)
G2:

=ABS(F2-D2)

(Absolute Difference)

Summary of Parameter Cells:

J1:

=AVERAGE(A2:A21)

(Sample Mean)
J2:

=STDEV.S(A2:A21)

(Sample Standard Deviation)

J4:

=MAX(G2:G21)

(Kolmogorov-Smirnov D Statistic)

Step 5: Interpreting the Results Using Critical Values

Hypothesis testing for the K-S test follows the standard format:

H0: The data is normally distributed (i.e., the difference D is due to random chance).
HA: The data is not normally distributed (i.e., the difference D is statistically significant).

To determine if we should reject the null hypothesis, we must refer to the critical value that corresponds to our sample size ($n = 20$) and chosen significance level ($alpha = .05$). This critical value defines the threshold of acceptable deviation from the theoretical distribution. The critical value must be obtained from a specialized Kolmogorov-Smirnov Table.

The decision rule dictates that if the calculated maximum difference (D = 0.10983) is greater than the critical value, we reject H_0. If D is less than or equal to the critical value, we fail to reject H_0. For $n = 20$ and $alpha = .05$, the critical value (which is typically around 0.190 for the Lilliefors correction often used when parameters are estimated) is significantly higher than our calculated D statistic.

Since our maximum value of **0.10983** is not greater than the critical value, we fail to reject the null hypothesis. This positive outcome allows us to assume that our sample data is sufficiently normally distributed for the application of parametric statistical methods.

Conclusion: Validating Statistical Assumptions

The successful application of the Kolmogorov-Smirnov Test in Excel provides analysts with a powerful, quantifiable tool for validating the fundamental assumptions underlying many statistical models. By systematically comparing the observed empirical CDF against the theoretical normal distribution, we derive a robust measure of fit.

The result of this analysis confirms that our sample data, with a maximum deviation D of 0.10983, does not significantly deviate from a normal distribution. This validation is a crucial step in ensuring that subsequent statistical findings are reliable and that the choice of statistical tests (e.g., t-tests or ANOVA) is appropriate for the nature of the data collected.

The following tutorials explain how to perform other common statistical tests in Excel: