Kolmogorov-Smirnov Test for Normality

Hypothesis Testing

Definition 1: Let $x_1,\dots,x_n$ be an ordered sample with $x_1 \le \dots \le x_n$ and define $S_n(x)$ as follows:

$$S_n(x) = \begin{cases} 0, & x < x_1 \\ k/n, & x_k \le x < x_{k+1} \\ 1, & x \ge x_n \end{cases}$$

Now suppose that the sample comes from a population with cumulative distribution function $F(x)$ and define D_n as follows:

$$D_n = \max_x |F(x) - S_n(x)|$$

Observation: It can be shown that D_n doesn't depend on F. Since S_n(x) depends on the sample chosen, D_n is a random variable. Our objective is to use D_n as a way of estimating $F(x)$.

The distribution of D_n can be calculated (see Kolmogorov Distribution), but for our purposes now the important aspect of this distribution are the critical values. These can be found in the Kolmogorov-Smirnov Table.

If $D_{n,\alpha}$ is the critical value from the table, then $P(D_n \le D_{n,\alpha}) = 1 - \alpha$. D_n can be used to test the hypothesis that a random sample came from a population with a specific distribution function $F(x)$. If

$$\max_x |F(x) - S_n(x)| \le D_{n,\alpha}$$

then the sample data is a good fit with $F(x)$.

Also from the definition of D_n given above, it follows that

$$1 - \alpha = P(D_n \le D_{n,\alpha}) = P\left(\max_x |F(x) - S_n(x)| \le D_{n,\alpha}\right)$$
$$= P\left(S_n(x) - D_{n,\alpha} \le F(x) \le S_n(x) + D_{n,\alpha} \text{ for all x}\right)$$
$$= P\left(|F(x) - S_n(x)| \le D_{n,\alpha} \text{ for all x}\right)$$

Thus $S_n(x) \pm D_{n,\alpha}$ provides a confidence interval for F(x)

Example 1: Determine whether the data represented in the following frequency table is normally distributed.

x	100	200	300	400	500	600	700	800	900	1000
Freq	8	25	88	172	243	252	144	49	13	6

Figure 1 – Frequency table for Example 1

This means that 8 elements have an x value less than 100, 25 elements have an x value between 101 and 200, etc. We need to find the mean and standard deviation of this data. Since this is a frequency table, we can't simply use Excel's AVERAGE and STDEV functions. Instead we first use the midpoints of each interval and then use an approach similar to that described in Frequency Tables as follows:

	I	J	K	L	M	N	O	P
1	Calculate mean and std dev							
2								
3	midpt	freq	midpt-sq					
4	50	8	2500		n = count	1000		=J14
5	150	25	22500		mean	481.4		=SUMPRODUCT(I4:I13,J4:J13)/B14
6	250	88	62500		mean-sq	231746		=N5^2
7	350	172	122500		sq-sum/n	255800		=SUMPRODUCT(J4:J13,K4:K13)/B14
8	450	243	202500		varp	24054.04		=N7-N6
9	550	252	302500		var	24078.12		=N8*J14/(J14-1)
10	650	144	422500		stdev	155.1713		=SQRT(N9)
11	750	49	562500					
12	850	13	722500					
13	950	6	902500					
14		1000						

Figure 2 – Calculating mean and standard deviation for data in frequency table

Thus, the mean is 481.4 and the standard deviation is 155.2. We can now build the table that allows us to carry out the KS test, namely:

	A	B	C	D	E	F	G
1	Kolmogrov-Smirnov						
2							
3	x	Freq	Cumul	$S_n(x)$	Z-Score	F(x)	Difference
4	100	8	8	0.008	-2.4579295	0.00698703	0.001013
5	200	25	33	0.033	-1.8134802	0.03487889	0.0018789
6	300	88	121	0.121	-1.169031	0.12119558	0.0001956
7	400	172	293	0.293	-0.5245817	0.29993701	0.006937
8	500	243	536	0.536	0.11986756	0.54770597	0.011706
9	600	252	788	0.788	0.76431682	0.77766077	0.0103392
10	700	144	932	0.932	1.40876608	0.92054783	0.0114522
11	800	49	981	0.981	2.05321534	0.97997415	0.0010258
12	900	13	994	0.994	2.6976646	0.99650861	0.0025086
13	1000	6	1000	1		1	0
14		1000					
15						D_n = max	0.011706
16						$D_{n,\alpha}$	0.043007

Figure 3 – Kolmogorov-Smirnov test for Example 1

Columns A and B contain the data from the original frequency table. Column C contains the corresponding cumulative frequency values and column D simply divides these values by the sample size (n = 1000) to yield the cumulative distribution function $S_n(x)$

Column E uses the mean and standard deviation calculated previously to standardize the values of x from column A. E.g. the formula in cell E4 is =STANDARDIZE(A4,N$5,N$10), where cell N5 contains the mean and cell N10 contains the standard deviation. Column F uses these standardized values to calculate the cumulative distribution function values assuming that the original data is normally distributed. E.g. cell F4 contains the formula =NORMSDIST(E4). Finally column G contains the differences between the values in columns D and F. E.g. cell G4 contains the formula =ABS(F4—D4). If the original data is normally distributed these differences will be zero.

Now D_n = the largest value in column G, which in our case is 0.0117. If the data is normally distributed then the critical value $D_{n,\alpha}$ will be larger than D_n. From the Kolmogorov-Smirnov Table we see that

$D_{n,\alpha} = D_{1000,.05}$ = 1.36 / SQRT(1000) = 0.043007

Since $D_n = 0.0117 < 0.043007 = D_{n,\alpha}$, we conclude that the data is a good fit with the normal distribution.

Example 2: Using the KS test, determine whether the data in Example 1 of Graphical Tests for Normality and Symmetry is normally distributed.

We follow the same procedure as in the previous example to obtain the following results. Since the frequencies are all 1, this example should be a bit easier to understand.

	A	B	C	D	E	F	G	H
1	**Kolmogrov-Smirnov**							
2								
3		x	Freq	Cumul	$S_n(x)$	Z-Score	F(x)	Difference
4		1.2	1	1	0.066667	-1.09281	0.137238	0.0705711
5		1.6	1	2	0.133333	-0.88981	0.186783	0.0534494
6		1.8	1	3	0.2	-0.78831	0.215256	0.0152563
7		1.9	1	4	0.266667	-0.73756	0.230389	0.0362772
8		1.9	1	5	0.333333	-0.73756	0.230389	0.1029439
9		2.0	1	6	0.4	-0.68681	0.2461	0.1539003
10		2.2	1	7	0.466667	-0.58532	0.279168	0.1874988
11		2.6	1	8	0.533333	-0.38232	0.351114	0.1822197
12		3.0	1	9	0.6	-0.17932	0.428845	0.1711553
13		3.5	1	10	0.666667	0.074433	0.529667	0.1369995
14		4.0	1	11	0.733333	0.328183	0.628613	0.1047202
15		4.8	1	12	0.8	0.734181	0.768581	0.0314191
16		5.6	1	13	0.866667	1.14018	0.872894	0.0062278
17		6.6	1	14	0.933333	1.647679	0.950291	0.0169574
18		7.6	1	15	1	2.155178	0.984426	0.015574
19	mean	3.4						
20	stdev	1.97					D_n =	0.1874988
21	count	15					$D_{n,\alpha}$ =	0.338

Figure 4 – KS test for data from Example 2

The Kolmogorov-Smirnov Table shows that the critical value $D_{n,\alpha} = D_{15,.05} = .338$

Since $D_n = 0.1874988 < 0.338 = D_{n,\alpha}$, we conclude that the data is a reasonably good fit with the normal distribution (more precisely that there is no significant difference between the data and data which is normally distributed). Note that is not the same conclusion we reached from looking at the histogram and QQ plot.

Real Statistics Excel Function: The following functions are provided in the Real Statistics Resource Pack:

> **KSCRIT**(*n, α, tails, h*) = the critical value of the Kolmogorov-Smirnov test for a sample of size *n*, for the given value of alpha (default = .05) and *tails* = 1 (one tail) or 2 (two tails, default), based on the KS Table. If *h* = TRUE (default) harmonic interpolation is used; otherwise linear interpolation is used.
>
> **KSPROB**(*x, n, tails, iter, h*) = an approximate p-value for the KS test for the D_n value equal to *x* for a sample of size *n* and *tails* = 1 (one tail) or 2 (two tails, default) based on a linear interpolation (if *h* = FALSE) or harmonic interpolation (if *h* = TRUE, default) of the values in the Kolmogorov-Smirnov Table, using *iter* number of iterations (default = 40).

Note that the values for *α* in the Kolmogorov-Smirnov Table range from .001 to .2 (for tails = 2) and .0005 to .1 for tails = 1. If the p-value is less than .001 (tails = 2) or .0005 (tails = 1) then the p-value is given as 0 and if the p-value is greater than .2 (tails = 2) or .1 (tails = 1) then the p-value is given as 1.

For Example 2, KSCRIT(15, .05, 2) = .338 (the same as shown in cell H21 of Figure 4). Also note that the p-value = KSPROB(H20, B21) = KSPROB(0.184177, 15) = 1 (meaning that p-value > .2), and so once again we can't reject the null hypothesis that the data is normally distributed.

If the value of D_n had been .35 in Example 2, then $D_n = .35 > .338 = D_{crit}$, and so we would have rejected the null hypothesis that the data is normally distributed. In this case we would have seen that p-value = KSPROB(.35,15) = .0427, which once again leads us to reject the null hypothesis.

Kolmogorov Distribution

As referenced above, the Kolmogorov distribution can be useful in conducting the Kolmogorov-Smirnov test. Click here for more information about this distribution, including some useful functions provided by the Real Statistics Resource Pack.

Lilliefors Test

When the population mean and standard deviation for the Kolmogorov-Smirnov Test is estimated from the sample mean and standard deviation, as was done in Example 1 and 2, then the **Kolmogorov-Smirnov Table** yields results that are too conservative. More accurate results can be derived from the Liiliefors Table as described in the Lilliefors Test for Normality.

91 Responses to *Kolmogorov-Smirnov Test for Normality*

Daniel *says:*
August 15, 2018 at 2:28 am

Hi,

I see that Real Statistics offers KSCRIT and KSPROB functions for the 1 sample KS test; is there a function that computes the KS test statistic for the 1 sample test for normality? I also see that there is a KS2TEST function for the two sample test. Is there a KS1TEST function?

Reply

Charles *says:*
August 15, 2018 at 9:09 am

Daniel,
There is no KS1TEST function. Instead the Real Statistics Resource Pack provides the ADTEST function which provides similar capabilities based on the Anderson-Darling test. See the following webpage:
http://www.real-statistics.com/non-parametric-tests/goodness-of-fit-tests/anderson-darling-test/
Charles

Reply

Giovanni *says:*
May 14, 2018 at 7:15 pm

MR. Charles it is necesary know degree fredom for normality using Shapiro Wilk?

Reply

Charles *says:*
May 14, 2018 at 8:12 pm

Giovanni,
There are no degrees of freedom for Shapiro Wilk-s Test
Charles

Reply

Yoan *says:*
May 2, 2018 at 4:48 pm

Hello Charles,

After reading the different approach to test for normality, I still don't understand what are the criterions (if any) for choosing a KS test over Shapiro-Wilk or a Chi squared ?

Thank you,

Yoan

Reply

Charles *says:*
May 2, 2018 at 7:35 pm

Yoan,
In general, I recommend Shapiro-Wilk over KS or Chi-squared. SW does a better job of determining that data from a normal distribution is normal and data that is not from a normal distribution is not normal.
Charles

Reply

Yoan *says:*
May 4, 2018 at 5:15 pm

Thank you for your answer,

I have a few questions on the example 2 thought, I performed the double tailed KS test in R on the same values and it returns me :
D = 0,88493
p-value = 1.254e-10

How does one compute the p-value ? Is it the probability of
F(Max Dn + Sn(x)) ?
And do you know where does this D comes from ?

Thanks again, your website is really great for self learner

Yoan

Reply

Wilson Fandino *says:*
March 31, 2018 at 11:26 pm

Thank you for your quick reply, Charles, it is more complex than I thought!

Wilson

Reply

Wilson Fandino *says:*
March 31, 2018 at 8:50 pm

Dear Charles,

I have read with great interest the use of Kolmogorov Smirnov for testing normality. In the Figure 3 you have nicely explained the test, based on the example 1. However, I cannot figure out where the data for the F(x) column came from. You have pointed out that these data came from the NORMSDIST function from Excel, but I am interested in the actual equation for calculation of the Standard Normal Density Function. I have found that this distribution follows the following equation:

F(x) = [e exp -(z exp 2)/2]/[Square root of (Pi*2)]

This is the distribution function of a theoretical set of data with mean = 0 and Standard Deviation (and therefore variance) =1

However, when I replace the z value in the equation, the resulting values do not correspond with the provided in Figure 3 for F(x). Could you provide us with the right equation, please? Thanks a lot for your help.

Wilson

Reply

Charles *says:*
March 31, 2018 at 10:53 pm

Wilson,
This is the formula for f(x), the pdf, and not the cdf, cumulative distribution function. The formula that you are looking for is the integral of f(x).
Charles

Reply

M *says:*
November 3, 2017 at 5:40 pm

My Kolmogorov-Smirnov is 0.000 significant, yet my supervisor wants me to have significant result, which is above 0.05, pleaseeeee someone help me. I don't know how to to get significant result and what is affecting the komogorov test to increase or decrease

Reply

Charles *says:*
November 3, 2017 at 6:12 pm

M,
If your supervisor wants a result above .05, then he wants a non-significant result (not a significant result).
Unless you have made a mistake a p-value = 0 is a significant result. You can't make it non-significant unless you change your data.
Charles

Reply

Daniel *says:*
November 2, 2017 at 7:58 pm

Dear Charles,

When i use KSCRIT(n=60 ; alpha=0.05) i get 0.1753, and when i use KSPROB(X=0.1753 ; n=60) the value is 0.0109. I checked the KS Critic but the obtained value for KSPROB is not logical even in the Kolmogorov Smirnov Table). Does function KSPROB has any restriction or problem?.

Reply

Charles *says:*
November 4, 2017 at 10:18 am

Daniel,
There is an error in the KSCRIT and KSPROB functions. Thank you very much for catching this error. I will fix it in the next release of the Real Statistics software, which I hope to issue shortly. I really appreciate your help in making the Real Statistics software more accurate and giving users confidence in the results obtained.
Charles

Reply

Mino *says:*
August 21, 2017 at 7:51 pm

Hi,

Do you have an example for a wind data application?

Something with weibull.

Thanks,

Reply

Charles *says:*
August 22, 2017 at 6:45 am

Mino, sorry but I don-t have such an example.
Charles

Reply

Anna Cherian *says:*
August 18, 2017 at 2:23 pm

Dear Sir,
Kolmogorov- Smirnov test the data analysis showed p=0 .271. No other information what is it mean?

Reply

Charles *says:*
August 18, 2017 at 8:16 pm

Anna,
Since p = .271 >.5 = alpha, you can't reject the null hypothesis that the data is normally distributed.
Charles

Reply

Muhammad Irfan *says:*
August 12, 2017 at 11:21 am

Dear Sir !
Please tell me how to calculate the X min value from the data by using K.S test

Reply

Charles *says:*
August 14, 2017 at 10:45 am

The KS test is not used to calculate the X min value. To calculate the minimum value of a range X, you use the Excel formula =MIN(X).
Charles

Reply

Neeraj *says:*
April 29, 2017 at 2:13 am

Dear Sir,
I am writing a java program to perform Kolmogorov Smirnov Test.
So I want to know the mathematical formulae to calculate the Z-Score and F(x).
And what does it mean by estimating population mean and standard deviation from the sample mean and standard deviation. Does it mean taking the mean and standard deviation of population same as sample mean and standard deviation.

Thank You

Reply

Charles *says:*
April 29, 2017 at 6:57 am

Neeraj,
Yes
Charles

Reply

César *says:*
March 13, 2017 at 11:48 pm

Dear Sir,

Is possible to determinate Dn parameter froma a data, using your Real Statistics addin. I was checking your examples and addin. However I cant find that formule.

Reply

Charles *says:*
March 14, 2017 at 7:23 am

César,
Sorry, but I have not yet implemented this is the Real Statistics addin. This value depends on the specific distribution that you are trying to fit and since there are so many distribution, I haven't tried to create a single formula for this. I will look into doing this shortly.
Charles

Reply

BASANTES S. Ruben *says:*
February 23, 2017 at 4:45 pm

Dear Charles,

Thanks for your application Real Statistics is ver useful
However, I have some questions about the Kolmogorov-Smironov test.
I have a huge sample of 3160 observations. My sample desribe a normal distribution, I can say that is a ideal sample.
When I conduct K-S test on this data the null hypothesis is rejected.
Why does it happen? I could expect that Ho was accepted!
thanks for your response but I am new in this test,

Best regards,

Reply

Charles *says:*
February 23, 2017 at 4:54 pm

Ruben,
If you send mean Excel file with your data and the KS test that you have conducted I will try to figure out what is going on. You can send the file to the email listed on the Contact Us webpage.
Note: You should use the Lilliefors version of the KS test for normality when you are estimating the the population mean and variance from the sample mean and variance.
Charles

Reply

Shaks *says:*

December 14, 2016 at 7:29 pm

Hi Charles:

If i want to use the Ks test to test for the Benford Distribution, would that be a two sample test? So I would use what you have at this url?
http://www.real-statistics.com/non-parametric-tests/two-sample-kolmogorov-smirnov-test/

Thanks,

Shaks

Reply

Charles *says:*
December 14, 2016 at 11:13 pm

Shaks,
No, it sounds like a one sample test.
Charles

Reply

Shaks *says:*
December 14, 2016 at 11:23 pm

Thank you. So in this case, my critical value would be 1.22/sqrt(n), assuming alpha = 10%?

Reply

Charles *says:*
December 15, 2016 at 8:07 am

Shaks,
Yes, if n > 50.
Charles

Reply

Shaks *says:*
December 15, 2016 at 9:06 pm

thanks!

Daniel Vincent *says:*
December 14, 2016 at 1:05 pm

Charles, can you help me with what kind of normality test & hypothesis test that I should use with my survey?

I created a survey about student's perception on business ethics, based on gender (male / female), ethics education (formal / informal), & age of maturity (junior / senior). n = 160 with unequal distribution: 44 male – 116 female, 77 informal – 83 formal, 28 junior – 132 senior.

Thank you in advance.
DV

Reply

Charles *says:*
December 14, 2016 at 2:43 pm

Daniel,
I suggest that you use the Shapiro-Wilk test for normality.

Regarding the hypothesis test, first you need to determine what hypothesis/hypotheses you want to test.
Charles

Reply

Richard *says:*
November 26, 2016 at 1:20 am

Hi Charles,

How should one choose n if the sample size is really small, say ~10?

Cheers,
Richard

Reply

Charles *says:*
November 26, 2016 at 7:09 pm

Richard,
If the sample size is 10, then n = 10.
Charles

Reply

Santosh *says:*
November 14, 2016 at 11:05 am

Hi Charles,
Do you have a example of applying K-S test to check for conformance with Benford's distribution. Do you think K-S method will be an appropriate method to check for deviation from Benford's law.
Thanks
Santosh

Reply

Charles *says:*
November 14, 2016 at 2:28 pm

Santosh,
Yes, I can see how you could use KS test to check or conformance with Benford's law.I have not provided an example of this.
Charles

Reply

aditya garg *says:*
April 20, 2016 at 2:15 am

Sir,
would u like to prescribe any book in which all these tests can be found

Reply

Charles *says:*
April 22, 2016 at 10:45 am

Aditya,
I will be publishing a series of books shortly which cover these tests.
Charles

Reply

aditya garg *says:*
May 18, 2016 at 9:54 pm

Waiting for your books Plz make it quick
Till then plz suggest some good literature for above tests

Reply

Charles *says:*
May 19, 2016 at 3:46 pm

I think that my website has a pretty good explanation of the Kolmogorov-Smirnov Test. You should also look at the following webpages:
One Sample Kolmogorov-Smirnov Test
Lilliefors Test
Charles

Reply

Ijeoma *says:*
April 11, 2016 at 1:15 pm

Dear Charles
Is it possible to get a conflicting result when the explore command and the one sample K-S are used to check the normality of the same data?

Reply

Charles *says:*
April 12, 2016 at 8:30 am

I don't know what "explore command" is.
It is not surprising that two different tests for normality will give different p-values.
Charles

Reply

Rob *says:*
November 14, 2015 at 10:02 am

Hi,
great site I learn from it a lot.

Can I please ask, how did you calculate a column K? midpt-sq?

Thank you.

Reply

Charles *says:*
November 14, 2015 at 1:22 pm

Rob,
Cell K4 contains the formula =I4^2, and similarly for the other cells in column K.
Charles

Reply

Rob *says:*
November 17, 2015 at 3:08 pm

Thank you very much Charles.
I dont know if I get it right, about numbers in column A (x values).
Let me explain on this example.

I have scale of loneliness and results can be
10-20 – low loneliness
20-30 – average loneliness
30-40 – high loneliness
so I calculate:

data midp freq midp^2
10-20 15 24 225
20-30 25 32 625
30- 40 35 33 1225

So:
n: 89
M: 26,011
Msq: 676,584
Sq-sum/n: 739,606
Varp: 63,022
Var: 63,738
Stdev: 7,983

x freq Cum Snx Z F(x) D
15 24 24 0,270 -1,379 0,084 0,186
25 32 56 0,629 -0,127 0,450 0,180
35 33 89 1 1,126 0,870 0,130

Dmax 0,185
Dkrit 0,144

0,185>0,144 so data is not normally distributed.

Is this rigt calculated or not?, I am not sure about choosing data for column A.

Reply

Charles *says:*
November 28, 2015 at 4:48 pm

Rob,

If you assume that the data in each interval is concentrated at the midpoint then the calculation is correct. I have typically assumed this for the calculation of the mean, but have used the right end-point of the intervals for the KS calculation. I can see advantages with both approaches.

I suggest that instead of using the KS table to calculate the critical value you use the Lilliefors Table instead. It is more accurate for determining whether data is normal when you use the sample mean and standard deviation. See the following webpage:

Lilliefors Test

Charles

Reply

Issam *says:*
November 1, 2015 at 1:33 pm

Dear Sir,

Thank you very much, I'm learning a lot from your website.

Unfortunately, my data set dose not fit with normal distribution.

I have very large data and I read in this paper (Open Access): Langlois, T. J., Fitzpatrick, B. R., Fairclough, D. V., Wakefield, C. B., Hesp, S. A., McLean, D. L., … Meeuwig, J. J. (2012). Similarities between Line Fishing and Baited Stereo-Video Estimations of Length-Frequency: Novel Application of Kernel Density Estimates. PLoS ONE, 7(11), 1–9. doi:10.1371/journal.pone.0045973

"We used Monte Carlo simulations to overcome uncertainty regarding the asymptotic distributions of KS test statistics under the null hypothesis".

How can I do the simulation in excel so my data can fit with normal distribution, so I can run the KS test for my data.

Thank you very much

Reply

Issam *says:*
November 1, 2015 at 5:36 pm

I want to do KS tow sample test

Reply

Charles *says:*
November 3, 2015 at 3:46 pm

See the following webpage
Kolmogorov-Smirnov Two Sample Test.

harles

Reply

Charles *says:*
November 3, 2015 at 3:50 pm

Sorry, but I don't understand your question. You don't do simulations to make data fit a distribution. If you knew the data was normally distributed then you wouldn't need to perform the KS test. Please clarify what you are trying to do.
Charles

Reply

Edgar *says:*
August 27, 2015 at 10:11 pm

Dear Charles I appreciate your contributions.

Please consider the following, in your second example you state the following:

"Since Dn = 0.1874988 < 0.338 = Dn,α, we conclude that the data is a reasonably good fit with the normal distribution (more precisely that there is no significant difference between the data and data which is normally distributed). Note that is not the same conclusion we reached from looking at the histogram and QQ plot"

So the same remains for Dn = 0.1874988 .338 = Dcrit, and so we would have rejected the null hypothesis that the data is normally distributed. In this case we would have seen that p-value = KSPROB(.35,15) = .0427, which once again leads us to reject the null hypothesis"

But if the α=0.01 then the critical value is 0.404 and Dn = 0.35 < 0.404 = Dn,α, Then, should we conclude that data is normally distributed ???

I´ll appreciate your comments,

Kind regards

Edgar

Reply

Charles *says:*

September 3, 2015 at 5:04 pm

Edgar,
Changing the value of alpha from .05 to .01, changes the value for Dcrit, but doesn't change the value of Dn. I don't see where the Dn = 0.35 comes from?
The null hypothesis that the data comes from a normal population cannot be rejected if Dn < Dcrit. Charles

Reply

Brody *says:*

August 25, 2015 at 5:21 pm

Charles, thanks, but I too have a question.

May I perform the KS test on two samples with different counts or n values?

For example, there are 7 possible categories, and there are 3 individual samples that will distribute within those 7 categories (dealing with sediment, sieves, and weights). I need to compare this set of samples to another set of samples, however, the number of samples here is 7. So, 2 sets of samples. The first, has 3 samples, and the second has 7. There are 7 sieve sizes or categories into which the samples are distributed. Can KS test be run on them?

If not, then would it be permissible to take the means of each sample, thus giving congruency to the data (same n values, but with means), and use the n from the sample size (n=10 (3 from first, and 7 from second)), rather than the mean (n=2) to establish the critical value, or would I need to use the n from the mean sample size to establish the CV?

Thanks,
Brody

Reply

Charles *says:*

August 26, 2015 at 9:41 am

Brody,
Although I don't completely understand your description, you can perform a two sample KS test with samples of different sizes to determine whether these samples come from populations with the same distributions. See the following webpage for more details
Two Sample KS Test
Charles

Reply

sathyapriya *says:*

August 19, 2015 at 5:46 am

how to i find cumulative distribution function F(x)?

Reply

Charles *says:*

August 22, 2015 at 10:17 pm

Examples 1 and 2 on the referenced webpage explain how to compute the cumulative distribution function F(x).
Charles

Reply

Gianma *says:*
July 2, 2015 at 4:03 pm

Dear Sir,
Thank you for sharing this.
I have a question: why in the first example we calculate the Z-score with x=100, 200, etc., but with mean and standard deviation calculated from the mid points (150, 250, etc.)?
Shouldnt'e be correct to have the mid points of the intervals in column A for Z-score calculation?
Best regards,
Gianma

Reply

Charles *says:*
July 7, 2015 at 8:01 am

Gianma,
Probably either approach is acceptable, but here I have used the endpoints of the various intervals with the mean and stdev based on the midpoint of the intervals.
Charles

Reply

Gianma *says:*
July 8, 2015 at 2:07 pm

Dear Charles,
Sorry for insisting, but it's not a negligible difference: using the midpoints of the intervals for calculating the Z-score, the resulting Dn is equal to 0.117>Dn,a, so the overall result is the opposite (the data is NOT a good fit with the normal distribution)…
Considering that the definition of Z is (Xi-u)/S, where u is the mean of the X values and S is their stdev, I think that only the midpoints of the intervals should be used, if u is calculated as their mean.
Otherwise, we can use the endpoints of the intervals as Xi, but in this case also mean and stdev should be calculated on these values, and not on the midpoints.
Do you agree?

Reply

Charles *says:*
July 15, 2015 at 4:43 pm

Gianma,

I realize that depending on the choice you make you might come to a different conclusion. This is why it is important to view significance values such as alpha = .05 not as absolute things. In fact if you set alpha = .05 as your significance value, any p-value near .05 can be viewed with some caution.

Unfortunately, this is the nature of statistics. If you get a p-value of .0003 you are fairly confident of your result (at least as far as type I error is concerned), but often depending on which test you choose to use (or which version of a test you use), you might get different outcomes.

Charles

Reply

Gianma *says:*
July 24, 2015 at 9:52 am

Dear Sir,
Thank you for your answer. Sincerely, I'm not 100% convinced, but at least this discussion forced me to look deeper into this topic, and review some forgotten page of statistics!
Best regards,
Gianmarco

Avi *says:*
June 24, 2015 at 3:39 am

Great article. I know understand how you calculate the P value for the KS test. Thanks so much. However when I try to replicate in excel, the NORMDIST function does not return the same values. Is there something different you are doing, as excel is asking me for the mean and stDev (which i input) but does not return the same values you have in your sheet

many thanks

Reply

Charles *says:*
June 24, 2015 at 7:44 am

Avi,
I don-t see any reference to the NORMDIST function on the referenced webpage. There is a reference to =NORMSDIST(E4(, which is the standard normal distribution function (mean = 0 and standard deviation = 1).
Charles

Reply

abhijit bora *says:*
May 12, 2015 at 10:43 am

Hello Sir,

I am searching for Kolmogorov-Smirnov Test two sample data in excel. Can you help me?

Reply

Charles *says:*
May 13, 2015 at 6:51 am

See the webpage Two Sample Kolmogorov-Smirnov Test
Charles

Reply

Sven *says:*
May 9, 2015 at 5:56 pm

Charles,

with your tool it is possible to use the Shapiro-Wilk-Test on a time series and get a besides the p-value a "yes" or "no" for the normal assumption. Therefore I can do this test for multiple series parallel with only one formula which is very nice.

Is there also a possibility to test for other distributions (Poisson, Stuttering Poisson, Gamma, Negative Binomial, etc.) for multiple series (KS-Test or Chi-Square-Test), so I can see which distribution would fit best?

Sven

Reply

Charles *says:*
May 11, 2015 at 11:20 am

Sven,
I haven't yet implemented software versions of chi-square or KS to test for a fit with a specific distribution. The One Sample Kolmogorov-Smirnov Test and Goodness of Fit webpages explain how this can be done, however.
Charles

Reply

Cathy *says:*
December 29, 2014 at 8:55 am

Can we use the Kolmogorov Smirnov test if we want to know whether the data follow a binomial distribution?

Reply

Charles *says:*
December 29, 2014 at 2:01 pm

Cathy,
Yes, you can use the KS test for this purpose. In addition to the referenced webpage, which shows how to use the KS test to determine whether data fits with the normal distribution, I give an example of how to do this for the exponential distribution on the webpage http://www.real-statistics.com/non-parametric-tests/one-sample-kolmogorov-smirnov-test/. The approach for the binomial distribution is similar. Also note that if the sample size is sufficiently large the binomial distribution can be approximated by a normal distribution, as described on the webpage http://www.real-statistics.com/binomial-and-related-distributions/relationship-binomial-and-normal-distributions/.
Charles

Reply

Jerome Gomes *says:*
August 10, 2015 at 9:12 pm

Dear Sir,
Can you give an example, where we can use KS table to determine whether the distribution follows poisson dist. or not, an excel worksheet will be helpful.

Regards,
Jerome
jeromegomes89@gmail.com

Reply

Charles *says:*
August 11, 2015 at 7:41 am

Jerome,

You can use the one sample KS test as described on the webpage
One Sample KS Test
The only problem is that the test is more accurate if you know the mean of the distribution instead of estimating it from the sample.

You can also use the chi-square goodness of fit test as described on the webpage
Goodness of Fit

Charles

Reply

Charles *says:*
June 20, 2014 at 8:22 am

Renato,
I have now provided another example of how to apply the KS test to determine whether a sample follows a specified distribution. See the webpage http://www.real-statistics.com/non-parametric-tests/one-sample-kolmogorov-smirnov-test/.
Charles

Reply

zohreh *says:*
June 13, 2014 at 5:34 am

Hi all,
I am trying to fit an appropriate probability distribution with my data. I have known that I can use K-S test, but my problem is that, as I am going to use MATLAB or EXCEL softwares for this purpose, I do not know how I can use these softwares for this test. My problem is that I have not ever seen any example of this test for exponential or other distributions rather than normal and lognormal distributions. How can I decide whether for example lognormal distribution is appropriate or exponential distribution?
Thank you very much for your help inn advance.

Reply

Charles *says:*
June 13, 2014 at 5:49 am

Hi Zohreh,
The approach for using the KS Test to test whether the data is exponentially distributed is very similar to that shown on the referenced webpage. I will add an example using the exponential distribution to the website in the next couple of days. This should help you.
Charles

Reply

Charles *says:*
June 20, 2014 at 8:25 am

Zohreh,
I have now added a description of how to determine whether data fits the exponential distribution using the KS test. See the webpage http://www.real-statistics.com/non-parametric-tests/one-sample-kolmogorov-smirnov-test/.
Charles

Reply

Kevin *says:*
June 10, 2014 at 6:31 am

Sir,

I am trying to determine if Rokeach value survey (RVS) responses for two different groups are statistically significant. The RVS has subjects rank 18 values in order of importance to them. I have calculated the mean response for each value within each group and ordered them from most important (lowest mean) to least important (highest mean). I was told I could use the Kolmogorov-Smirnov Test to determine if differences in mean value rankings between groups are statistically significant. I would appreciate an explanation of this process in Excel.

Thank you in advance,
Kevin, Excel expert, stats neophyte

P.S. I have learned more practical statistics from your site than my undergrad and masters professors have been able to drill into me… Well done, Sir!

Reply

Charles *says:*
June 11, 2014 at 10:23 am

Kevin,

It is good to hear that the site has been helpful. My goal was exactly as you stated, to help people make practical use of (and understand) statistics in the environment is probably the most available for most people, namely Excel.

If your goal is to determine whether there is a significant difference between the means of the two groups, you probably want to use the t test (if the data in the two groups are normally distributed) or the Mann-Whitney test if they are not. You could also use the two-sample Kolmogorov-Smirnov Test to determine whether the two groups of data come from the same population. I have already described the one sample Kolmogorov-Smirnov Test on the website, but not the two sample test.

Fortunately, I have just implemented the two sample test in the Real Statistics Resource Pack (Release 2.15) and have written the description for the website (including two examples). I plan to release these in the next couple of days. Stay tuned.

Charles

Reply

Charles *says:*
June 20, 2014 at 8:14 am

Kevin,
The two-sample KS test is now included in the Real Statistics Resource Pack. The procedure is described on the webpage http://www.real-statistics.com/non-parametric-tests/two-sample-kolmogorov-smirnov-test/.
Charles

Reply

Sally *says:*
May 28, 2014 at 3:36 am

hai, may I know what the p-value mean by and how to find the p-value of kolmogorov-smirnov ?

Reply

Sally *says:*
May 28, 2014 at 3:38 am

besides that is it possible to use the statistical value of other distribution as a critical value to find the p-value of KS test?
for example, use the z value of normal distribution to find the p-value by KS test.

Reply

Charles *says:*
May 31, 2014 at 2:17 pm

Sally,
Sorry, but I don't understand your question. In any case I will be adding the KS p-value shortly.
Charles

Reply

Charles *says:*
May 31, 2014 at 2:16 pm

Sally,
I am revising the KS part of the website/software and will add the p-value. Stay tuned.
Charles

Reply

Charles *says:*

June 20, 2014 at 8:19 am

Sally,
I have now provided a way of calculating the p-value for the KS test, using the functions KSPROB and KSDIST. These are available in the latest release of the Real Statistics Resource Pack (Rel 2.15).
Charles

Reply

Sandeep R *says:*
May 11, 2014 at 9:09 am

hello sir, i found this article very helpful. i need to fit log normal distribution either from chi square or K-S test. you have explained only normal distribution. please explain log normal distribution also.
here is my test data
mean= 5.1439
δ= 0.2506
median= 4.99
σlnz = 0.247

interval observed frequency
1.81 2.759 9
2.759 3.708 61
3.708 4.657 116
4.657 5.606 155
5.606 6.555 120
6.555 7.504 42
7.504 8.453 7
8.453 9.402 2
9.402 10.351 2
10.351 11.3 3
sum= 517

Reply

Charles *says:*
May 11, 2014 at 9:05 pm

The procedure for using the K-S test with the log normal distribution is pretty much the same as for the normal distribution. E.g. in Figure 3, you won't need the E column. Simply enter the formula for the log-normal distribution in column F. E.g. cell F4 would contain a formula like =LOGNORMDIST(A4,N5,N10). The rest is the same as in the examples provided on the webpage.
Charles

Reply

Sandeep R *says:*
May 12, 2014 at 8:04 am

thank you very much sir for your reply.
sir i have one more doubt, should we use "mean and standard deviation" or "Median and σlnz in lognormal distribution?

Reply

Charles *says:*
May 15, 2014 at 7:23 am

Sandeep,
If I understand your original question correctly, then you should use the mean and std dev, esp. since Excel has the LOGNORM.DIST function available which use these two parameters. Why do you think the median and σlnz might be good choices? Perhaps this is correct and I am not answering the right

question.
Charles

Reply

Sandeep R *says:*
May 25, 2014 at 3:22 pm

thank you sir

Bruno *says:*
January 4, 2016 at 4:18 pm

Hi,

Critical values are the same for all distributions tested?

Thank you

Reply

Charles *says:*
January 4, 2016 at 9:16 pm

Hi Bruno,

If you know the distribution parameters (e.g. mean and variance for the normal distribution) then the table of critical values is the same. Usually these parameter values are estimated from the sample, in which case different critical values should be used. I have provided the table for the normal distribution. This is call the Lilliefors Test. See the following webpages

Lilliefors Test
Lilliefors Table

Charles

Reply