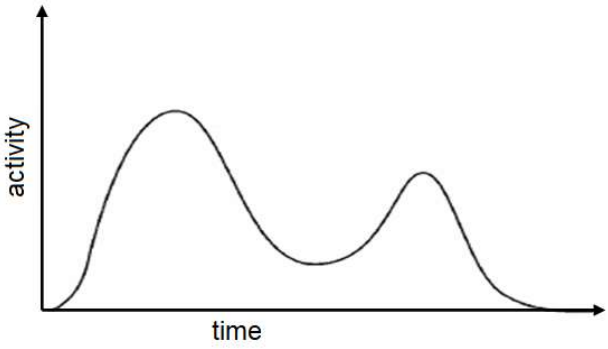
Simple guide to Kolmogorov-Smirnov test

Kolmogorov-Smirnov is one of the most commonly used **non-parametric** tests uses to assess if two data sets are distinct enough in a 'statistically significant'* sense. It can also be used to compare a distribution against normal distribution to figure out if a given data set is normal. In this case we are talking about one sample Kolmogorov-Smirnov test of Kolmogorov-Smirnov test for normality.

If you are reading this document, you are either curious or you have been told that you need to use a non-parametric test because your data is not normal. Let's expand this a bit. A test is **parametric** if it uses a **parameter(s)** derived from the sample(s) for calculations. For example, t-test and ANOVA parametric tests because they use both mean and variance of the data in order to make statistical inference. However in order for the parameters to be representative the data must be normal.



For example we can determine 'mean' for the data on the left, but the mean would sit between the two peaks and would thus represent the data poorly.

**Statistical significance as a term has several issues. One being that 'statistical significance' does not imply actually significant difference.*

A non-parametric test, such as Kolmogov-Smirkon, Kurskall-Wallis and Mann-Whitney-U do not rely on parameters (as the non-parametric part implies). Instead they make the inference using every measured data point. The methods differ but the gist remains the same. The two data sets are compared in their entirety and statistical inference is made.

As this document aims to describe Kolmogorov-Smirnov (KS) test we are going to walk through the idea behind it. To understand KS test, we must understand something called **empirical cumulative distribution function (CDF)**. So we are going to discuss mathematics. Now, of course we could just wave hands and ask the reader to trust the mathematics. But doctors and bioscientists trusting mathematics blindly is what got us into trouble in the first place, so we are not going to do that.

Despite its mathematical sounding name CDF is not particularly convoluted concept. For EDF, the value of the function at given step *t* is percentage of data points at *t* and below it

Let's have a distribution (data set) with four measurements such that: A = [1 2 2 3]. Now Empirical distribution function F(t) is formed as follows:

$$F_n(t) = \frac{number\ of\ elements\ in\ the\ sample \leq t}{total\ number\ of\ elments\ in\ the\ sample},$$

And because this is going to be Hebrew to most people, we are going to walk through this. Idea is that we run *t* from 0 to the largest measured value (or we run the *t* from the smaller measured value to the largest measured value).

We start from *t* = 0.

$$F_n(0) = \frac{number\ of\ elements\ in\ the\ sample\ \leq 0}{4}$$

How many values in A are equal or less than zero? Non. Thus F(0) = 0.

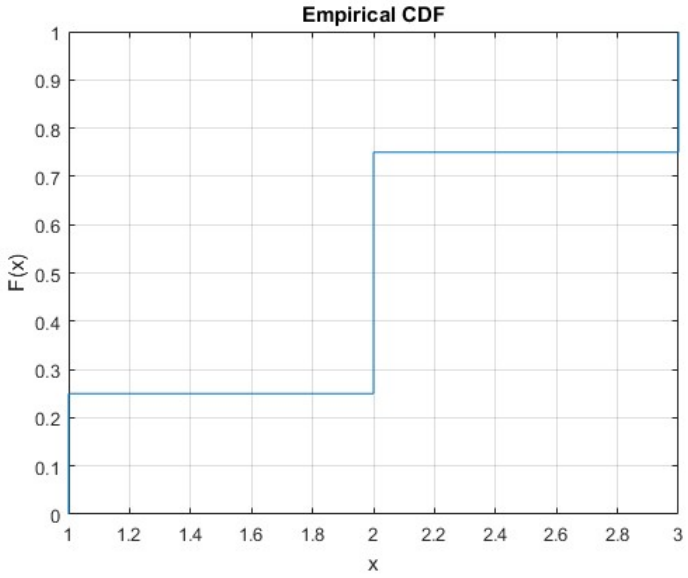
$$F_n(1) = \frac{number\ of\ elements\ in\ the\ sample\ \leq 1}{4} = \frac{1}{4} = 0.25$$

$$F_n(2) = \frac{number\ of\ elements\ in\ the\ sample\ \leq 2}{4} = \frac{3}{4} = 0.75$$

$$F_n(3) = \frac{number\ of\ elements\ in\ the\ sample\ \leq 3}{4} = \frac{4}{4} = 1$$

So our CDF F(t) has values [0 0.25 0.75 1] at locations [0 1 2 3]. If we plot it, it looks like this.

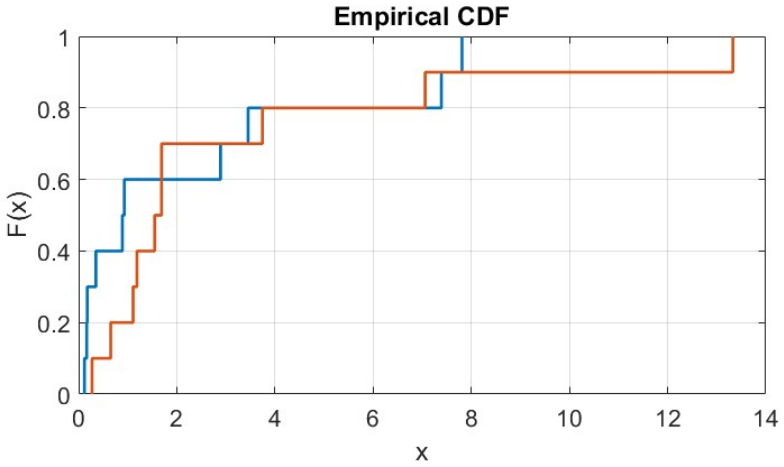
F(t)	Value
0	0
1	0.25
2	0.75
3	1



The cumulative distribution function is called 'empirical' as it is based on empirical evidence, i.e. data. It is possible to determine CDF for theoretical distributions such as normal distribution.

Of course can do this kind of calculations to more complex set of data. Say we have two samples S1 and S2.

S1	S2
7,39	1,69
0,12	13,33
0,18	0,65
2,89	0,27
0,89	1,55
7,81	1,11
3,45	1,19
0,35	3,74
0,93	1,69
0,17	7,06



Here the blue line corresponds to S1 and red line to S2.

Some text books write the expression

$$F_n(t) = \frac{number\ of\ elements\ in\ the\ sample \leq t}{n}$$ as

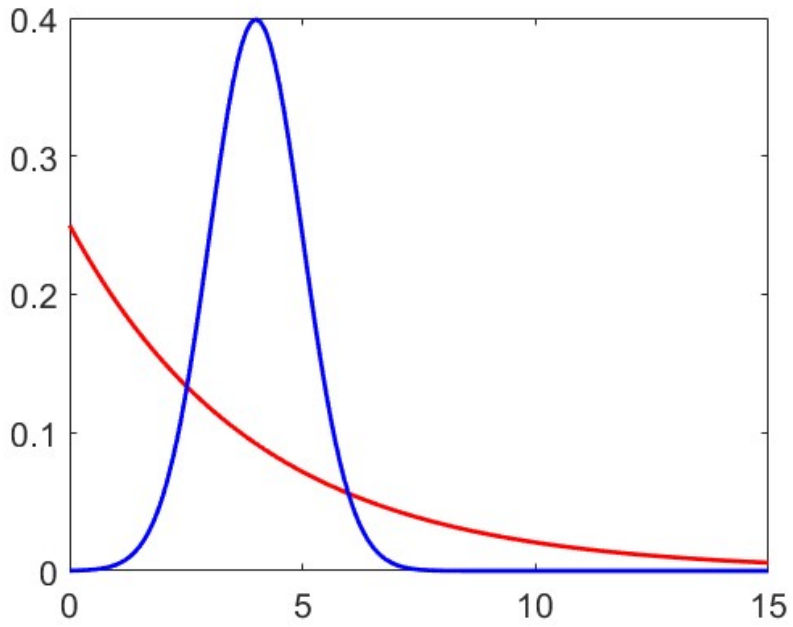
$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{x_i \leq t}$, but this is really just the same thing expressed in a more

compact form. I am noting this in case you bump into it somewhere else. Also sometimes the n is replaced by $n + 1$. The principle remains the same.

Practical example 1.

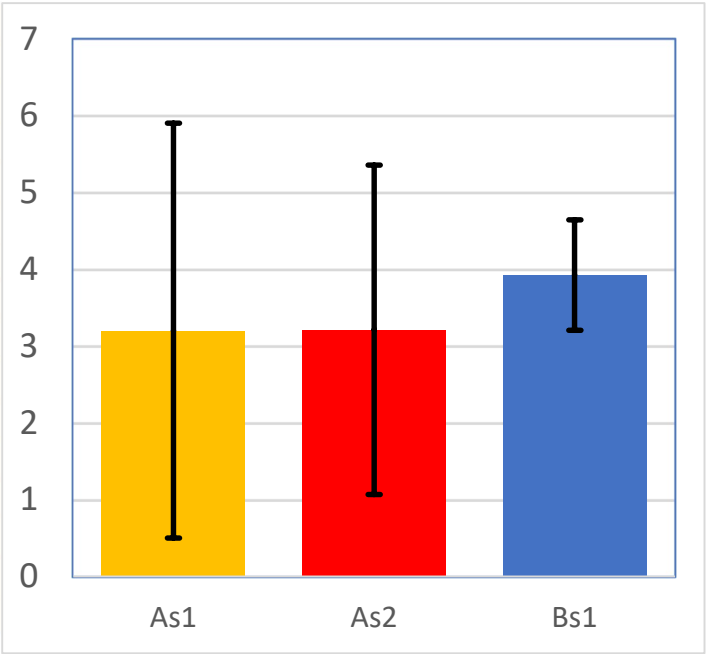
Let us have two (probability) distributions as follows.

One exponential distribution A(μ=4) and one normal distribution B(μ=4, σ = 1). As graphs these distributions look like this. It's rather obvious that they are not the same distribution.



Now let us draw two samples from the A and call them As1 (A sample 1) and As2 (A sample 2) and let us draw one sample from the B and call it Bs1 (B sample 1). These samples are shown below.

As1	As2	Bs1
0,164	5,477	3,898
4,311	0,822	3,759
2,143	5,650	4,319
5,988	0,294	4,313
1,144	4,200	3,135
5,465	6,506	3,970
2,725	5,528	3,835
1,432	1,938	4,628
0,462	2,992	5,093
0,166	4,180	5,109
2,412	0,741	3,136
7,904	2,143	4,077
7,607	2,393	2,786
5,427	0,346	2,887
0,694	5,009	3,993



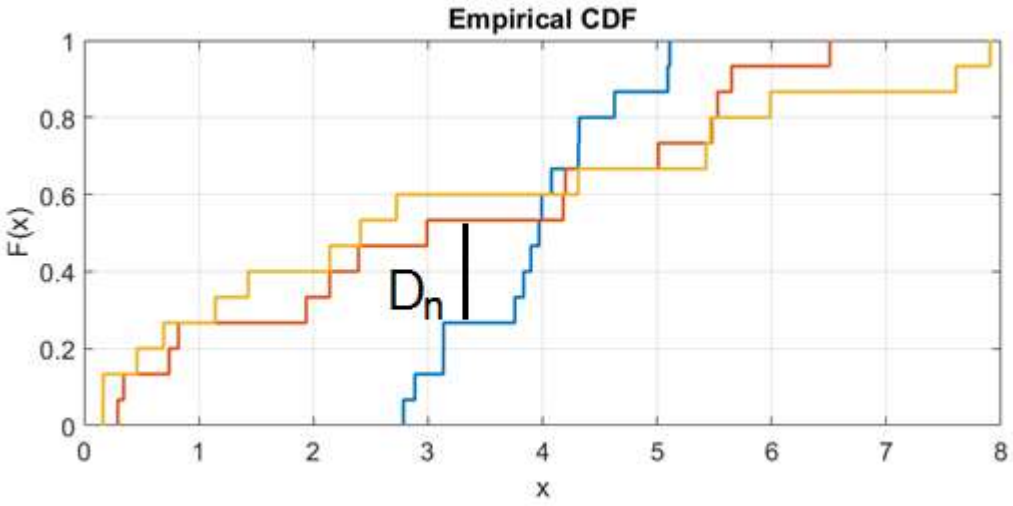
On the previous page we have the three samples plotted as bars with error bars shwing the standard deviations. While we can see that the Bs1 looks different from As1 and As2

Kolmogorov-Smirnov test statistic for two probability distributions is given by

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

and this might sound scary. But it's not. Here $F_{1,n}(x)$ and $F_{2,m}(x)$ are the corresponding empirical cumulative distribution functions of the first and the second sample respectively, *n* and *m* are the sample sizes of the first and the second sample respectively and *sup* is the supremum function.

This sounds scary but basically means 'find the biggest difference between the two CDF's. This is your test statistic.



Basically we measure the distance between the CDF's for every value of x and the largest we can find is our test statistic.

Our null is that the samples have been drawn from the same distribution. If you are reading this, you really ought to know what null hypothesis is. If you don't, find out and come back.

Null is rejected at level α if:

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}},$$

where *n* and *m* are the sample sizes of the first and the second sample respectively and c(α) is given by:

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln\alpha}.$$

How to do KS-test in practise

In practice no-one is going to do the test by hand. However, Excel (without expansions) won't do Kolmogorov-Smirnov test for you.

On matlab the syntax is:

[h,p,ks2stat] = kstest2(A,B,'Alpha',0.05),

Where A1,B1 are your data, the 0.05 is the value of your alpha. The output [h,p,ks2stat] contains the null (h), the p-value (p) and the test statistic (ks2stat).

In matlab, you feed in your data in the form A1 = [values], e.g.

A = [3.8980 3.7590 4.3190 4.3130 3.1350 3.9700 3.8350 4.6280 5.0930 5.1090 3.1360 4.0770 2.7860 2.8870 3.9930];

Everything between the [ ] is your sample and the measurements are separated by whitespace. Basically you need to feed in your samples in this manner.

Then you call the function (or copy paste it). The 0.05 is the wanted p-value.

[h,p,ks2stat] = kstest2(A,B,'Alpha',0.05)

```
>> [h,p,ks2stat] = kstest2(A1,B1,'Alpha',0.05)

h =

  logical

   1


p =

    0.0047


ks2stat =

    0.6000
```

This will yield something like this.

h=1, null rejected at alpha
h=0, null not rejected at alpha

And the p shows the corresponding p-value.