The Kolmogorov-Smirnov Test

Karl Juhl, Juan Gil de Gomez, Oriol Vall

Abstract

The purpose of this paper is to explore the usability of the Kolmogorov-Smirnov test, including how it works, why one would use it and in what context, and how to use it. We will use the test in different environments, namely by hand as well as R programming. This paper should be a clear introduction and guide to using this statistical test, whether the user wants to apply it in their Python environment or understand how the test works manually. Additionally the reader should clearly understand the advantages and disadvantages of the Kolmogorov-Smirnov Test, allowing them to know when to use it and when not to use it, as well as similar options to the test.

Table of Contents

Introduction

Imagine that you are trying to run a simple linear regression, you are trying to predict how much money people make based on their GPAs from university. A simple linear regression has a lot of assumptions that need to be met so the prediction is accurate and with low bias. If one of the assumptions are not met and we run a model, our outcome will be biased and therefore not useful, and we may not even realize it because we didn’t test for it. Normality is one of the assumptions and the most important one. If there is no normality you cannot run a linear regression .

In statistics a vast majority of tests rest upon the assumption of normality, therefore if the data collected is not normally distributed the reliability and accuracy of the test decreases. Those tests are the parametric tests, which unlike the non-parametric tests, are based on the assumption that the data are distributed on some well-known distributions like a normal distribution for example. How do we know if the data follows those distributions? The Kolmogorov-Smirnov test lets us check it. This unique test allows us to check if two datasets differ significantly, like whether a sample comes from a specific distribution or not, e.g. if it has a normal distribution. It is also important to note that the KS - test can be either for one sample or for two samples. You can run this test to see if the two samples that you have are normally distributed, but in this paper we will go more in depth in the one sample test.

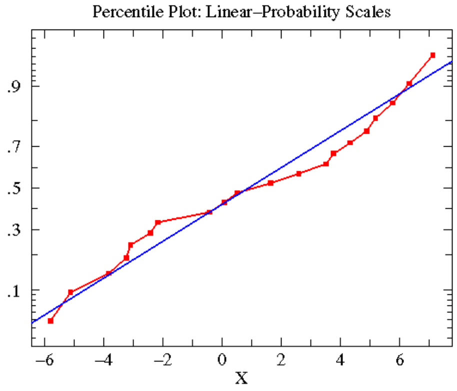
When we refer to normality in data is when the sample follows a normal distribution i.e. the sample has the majority of its points happening in a small range of values but also having a small number of outliers in both the upper and lower end. Not only that but normally distributed samples also happen to have the same number for the mean, mode, and median.

Theory

The one sample KS-test is used for testing whether a sample of data follows a normal distribution or not. This characteristic of the data tested is crucial when analysing, testing, and modelling using the data, and knowing whether the data is normal helps one pick the most appropriate test, or make transformations to make the data normal.

In order to do this, the KS-test examines the distribution of the sample data, and compares it to the cumulative distribution of a theoretical normal distribution. A visual representation of the test is shown below, with the red line representing the sample data, and the blue line being the theoretical normal distribution. In order to decide whether the sample data follows a normal distribution, the test uses the following formula to calculate the test statistic D:

$$D \ = max \ _{1\leq Y\leq N} \left[F(Yi) - \frac{i-1}{N}, \frac{i}{N} - F(Yi)\right]$$



Where F is frequency, Yi is the respective observed data values of the sample (holding i as the index), N is the sample size. In simpler terms, the formula takes the maximum of the first half of the square brackets, then the maximum of the second half of the square brackets, and takes the maximum between these two. The resulting value is the test statistic D, which will then be compared to the critical value D_{α}. This can be further simplified in a visual manner to say that the test statistic D is the maximum distance vertically between the sample data distribution and the cumulative form of the empirical normal distribution.

The Kolmogorov-Smirnov normality test includes a few assumptions which must be met in order for the results to be reliable - not too conservative, and not too strict. The first is the sample must be a random sample. The second is that the theoretical normal distribution used for comparison in the test must be specified fully, meaning the parameters must be defined before-hand rather than estimated from the data. Furthermore, the theoretical distribution, as well as the sample distribution should be of continuous nature. With these assumptions met, the KS-test will produce reliable results which will allow one to move on with their analysis with confidence of the normality or lack thereof in their data. One limitation is the test is usually more sensitive near the center of the distribution, and more conservative near the tails of the distribution. The other main limitation is the requirement of the parameters of the theoretical distribution being fully defined. In some cases this limits the use of the test depending on the data, and leads to many erroneous tests being conducted.

Implementation Example

We are trying to make a model with multilinear regression about the prediction of students' GPA based on their attendance, age, and gender. But for this model, we first need to check the assumptions required to create said model and we find that one of those is the normality of data. Therefore what we do is apply the Kolmogorov-Smirnov test. The sample collected regarding the GPA in the following:

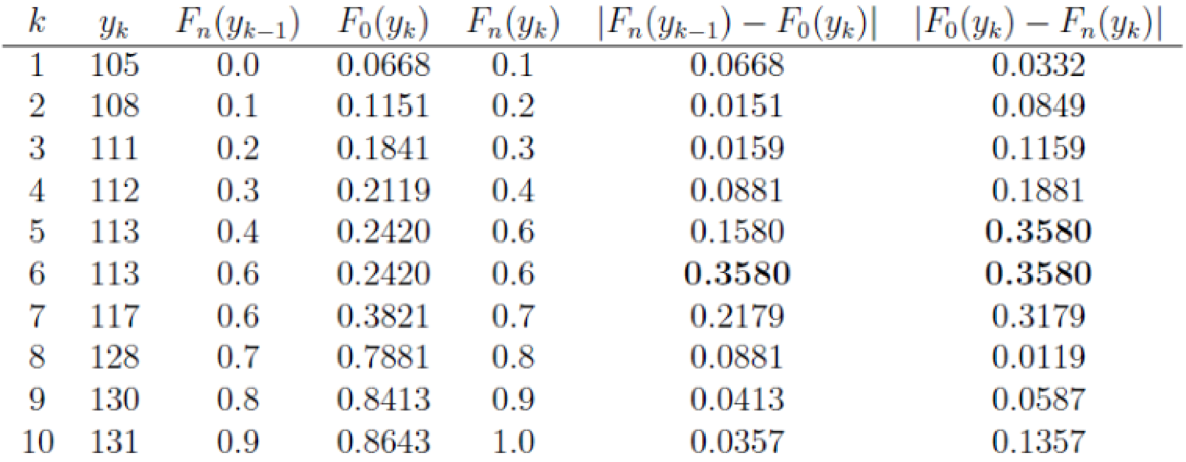
i	1	2	3	4	5	6	7	8	9	10
Sample Data (Yi)	108	112	117	130	111	131	113	113	105	128

Once we know what the data is we establish the parameters of what would be a perfect normal distribution of this data and we have that the mean is 120 and the variance 10.

Then we establish that the null hypothesis is that the data is normally distributed and the alternative hypothesis is that the data is not normally distributed, we reject Ho when D is higher than Da.

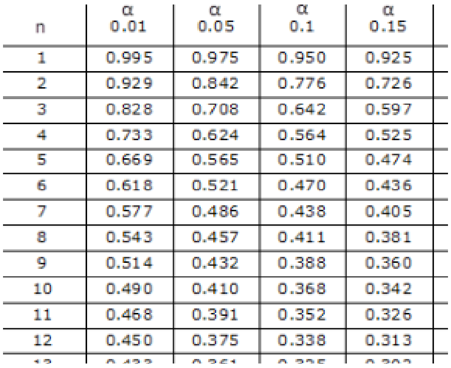
We could solve this example by either doing all the calculations by hand or by using tools such as R studio. First of all, if we want to do it by hand, we have to keep in mind the formula and the theory explained before.

The first step is to order the observations from smaller to higher and then we calculate the cumulative probabilities of the observed values $F_o(y)$ and the cumulative probabilities of $F_o(y_{k-1})$and we compare those to the cumulative probabilities of the theoretical normal distributed sample $Fn(y_k)$.



k	y_k	$F_n(y_{k-1})$	$F_0(y_k)$	$F_n(y_k)$	$\|F_n(y_{k-1}) - F_0(y_k)\|$	$\|F_0(y_k) - F_n(y_k)\|$
1	105	0.0	0.0668	0.1	0.0668	0.0332
2	108	0.1	0.1151	0.2	0.0151	0.0849
3	111	0.2	0.1841	0.3	0.0159	0.1159
4	112	0.3	0.2119	0.4	0.0881	0.1881
5	113	0.4	0.2420	0.6	0.1580	**0.3580**
6	113	0.6	0.2420	0.6	**0.3580**	**0.3580**
7	117	0.6	0.3821	0.7	0.2179	0.3179
8	128	0.7	0.7881	0.8	0.0881	0.0119
9	130	0.8	0.8413	0.9	0.0413	0.0587
10	131	0.9	0.8643	1.0	0.0357	0.1357

When calculating the difference we use the highest of them, i.e. the highest D, and then we use the Kolmogorov-Smirnov table to see whether the D obtained during the operations is higher or lower than the one on the table(Da). To obtain the Da from the table we need to use the number of observations as the degrees of freedom and the significance level. In this case, we see on the table below, that the D we found was 0.358 and compared to the Da is lower and therefore we reject the null hypothesis.

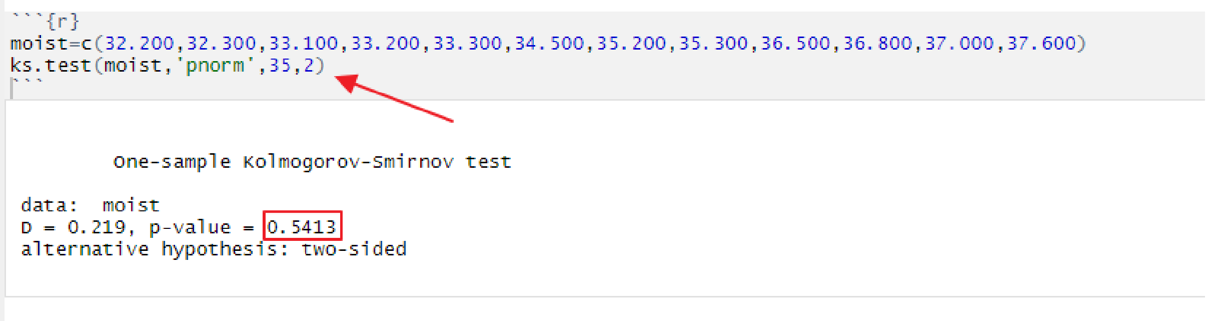


n	α 0.01	α 0.05	α 0.1	α 0.15
1	0.995	0.975	0.950	0.925
2	0.929	0.842	0.776	0.726
3	0.828	0.708	0.642	0.597
4	0.733	0.624	0.564	0.525
5	0.669	0.565	0.510	0.474
6	0.618	0.521	0.470	0.436
7	0.577	0.486	0.438	0.405
8	0.543	0.457	0.411	0.381
9	0.514	0.432	0.388	0.360
10	0.490	0.410	0.368	0.342
11	0.468	0.391	0.352	0.326
12	0.450	0.375	0.338	0.313

Example with R studio.

Testing for normality is way easier in RStudio. We obtained the data of horn moisture in cows, that the population's mean is 35 and variance 2. When writing the code is really important to know what parameters are we comparing our data too. We need to know the population's mean and variance to specify them in the code.

i	1	2	3	4	5	6	7	8	9	10	11	12
Sample data (Yi)	32.2	32.3	33.1	33.2	33.3	34.5	35.2	35.3	36.5	36.8	37	37.6



~~~
```{r}
moist=c(32.200,32.300,33.100,33.200,33.300,34.500,35.200,35.300,36.500,36.800,37.000,37.600)
ks.test(moist,'pnorm',35,2)
```
~~~

```

        One-sample Kolmogorov-Smirnov test

data:  moist
D = 0.219, p-value = 0.5413
alternative hypothesis: two-sided
```

First of all you need to create a vector with all the data points that were collected. Then it is just as simple as writing the main code that is ks.test with its parameters. First is the vector with all the data, after is to what type of distribution you want to compare it to, in this case pnorm since we want to test if it is normal distribution. And finally you need to introduce the population's mean and variance. The output that we get is composed of two main things. Da which you can use to compare to the table. Or even easier the P-value, if it is higher than 0.05 then we fail to reject, which means that the sample seems to be normally distributed.

Conclusion

The Kolmogorov-Smirnov test has become a standard for checking the assumption of normality in data, which is vital for much analysis and research. However the limitations of the test have led to more refined tests being created based on the KS-test, namely the Anderson-Darling or Cramer Von-Mises tests.

References

College of Saint Benedict & Saint John's University. *Kolmogorov-Smirnov Test*: http://www.physics.csbsju.edu/stats/KS-test.html

NIST. *1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test:* https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

InfluentialPoints. *Kolmogorov-Smirnov test one- & two-sample, and related tests:* https://influentialpoints.com/Training/kolmogorov-smirnov_test-principles-properties-assumptions.htm

MIT. (2006). *Section 13 Kolmogorov-Smirnov test*: https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf