The Kolmogorov-Smirnov Test

Introduction

During data analysis, you have to satisfy a number of assumptions for the underlying dataset. One of the most common assumptions that you will come across is the "Normality Assumption", i.e., the underlying data roughly follows a normal distribution.

If the data is not found to be normally distributed (i.e. data with kurtosis and skew while doing linear regression), you may first answer a question like: “Given my data … if there is a deviation from normality, will there be a material impact on my results?”

In this lesson, we'll look at a popular statistical test for satisfying the normality assumption, the Kolmogorov-Smirnov test, or simply, the K-S test.

Objectives

You will be able to:

- Explain the role of the normality assumption in statistical tests
- Calculate a one-and two-sample Kolmogorov-Smirnov test
- Interpret the results of a one- and two-sample Kolmogorov-Smirnov test

Normality assumption

Formal normality tests always reject the huge sample sizes we work with today. When n (our sample size) gets large, even the smallest deviation from perfect normality will lead to a significant result. And as every dataset has some degree of random noise, no single dataset will be a **perfectly** normally distributed sample.

> **In applied statistics, the question is not whether the data/residuals are perfectly normal, but normal enough for the assumptions to hold**.

This question is answered through visualization techniques like qqplots, boxplots, or more advanced statistical tests including:

- The Shapiro-Wilk test;
- The Anderson-Darling test, and;
- The Kolmogorov-Smirnov test

In this lesson, we'll focus on the Kolmogorov-Smirnov test (K-S test) which will give you a strong foundation to help you understand and implement other tests when needed.

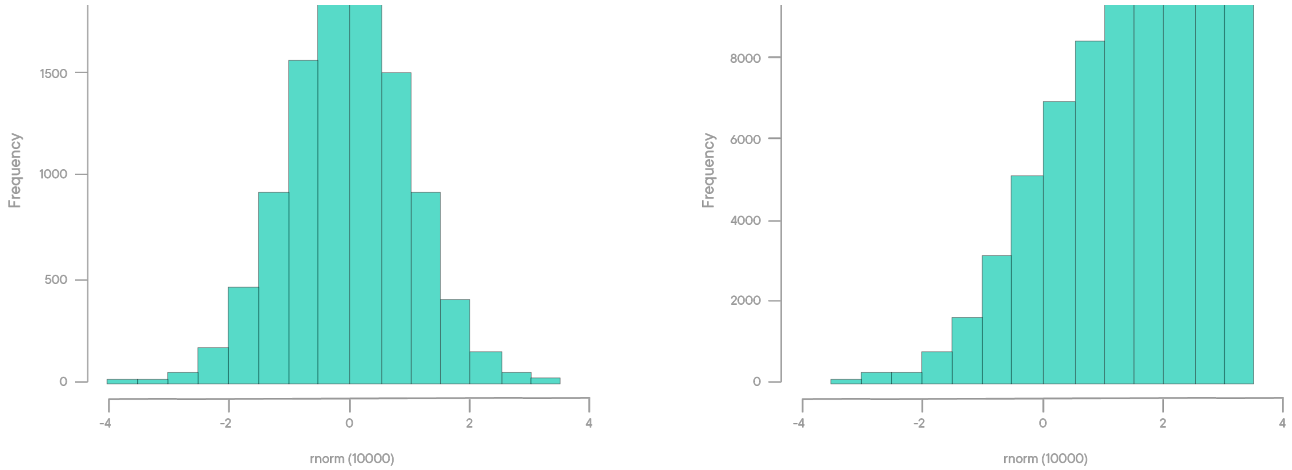
Kolmogorov-Smirnov Test

A K-S test provides a way of comparing distributions, whether two sample distributions or a sample distribution with a theoretical distribution - comparable to what we've already seen when we learned about one sample or two-sample t-tests. The distributions are compared in their cumulative form as **Empirical Cumulative Distribution Functions**. The test statistic in K-S test used to compare distributions is simply the maximum vertical distance between the two functions. Essentially, we are testing the sample data against another sample, to compare their distributions for similarities.

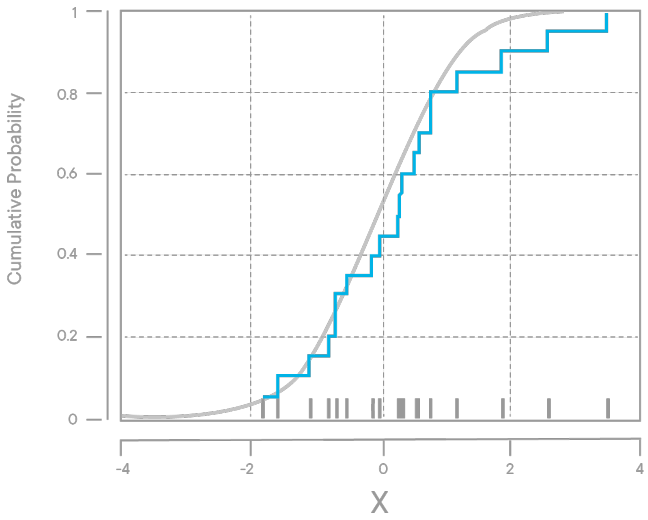
The Empirical Cumulative Distribution Function (ECDF)

> An empirical cumulative distribution function (CDF) is a non-parametric estimator of the underlying CDF of a random variable. It assigns a probability to each data point, orders the data from smallest to largest in value, and calculates the sum of the assigned probabilities up to and including each data point.

The most intuitive way to think about the empirical distribution function is that it relates to the cumulative distribution function (CDF) in a similar way to how a histogram relates to a probability density function. Let's look at the following figures to get this idea:



The left figure shows a regular histogram with samples looking like a normal distribution. The right figure shows the same samples except each bin in the histogram contains the cumulative count of samples up to that bin, which approximates the shape of the CDF for this random variable. Now the right figure doesn't exactly represent an empirical distribution function because the Y-axis is not normalized to 1 and the samples are binned instead of just plotted cumulatively. Nonetheless, the idea remains the same. An example of an empirical CDF is given below:



This image sums up the intuition for empirical distribution function. The blue line is our empirical CDF whereas the grey one is our theoretical CDF (i.e. plotted using parameters and fitting a probability function).

If X is a random variable with CDF $F(x) = P(X \leq x)$, and $x1, \ldots, xn$ are i.i.d. random variables sampled from X. Then, the empirical distribution function, $F(x)$, is a CDF:

$$\hat{F}(x) = \frac{\#\text{ of elements in sample} \leq x}{n} = \frac{1}{n}\Sigma_{i=1}^{n} I(x_i \leq x)$$

One-Sample K-S test

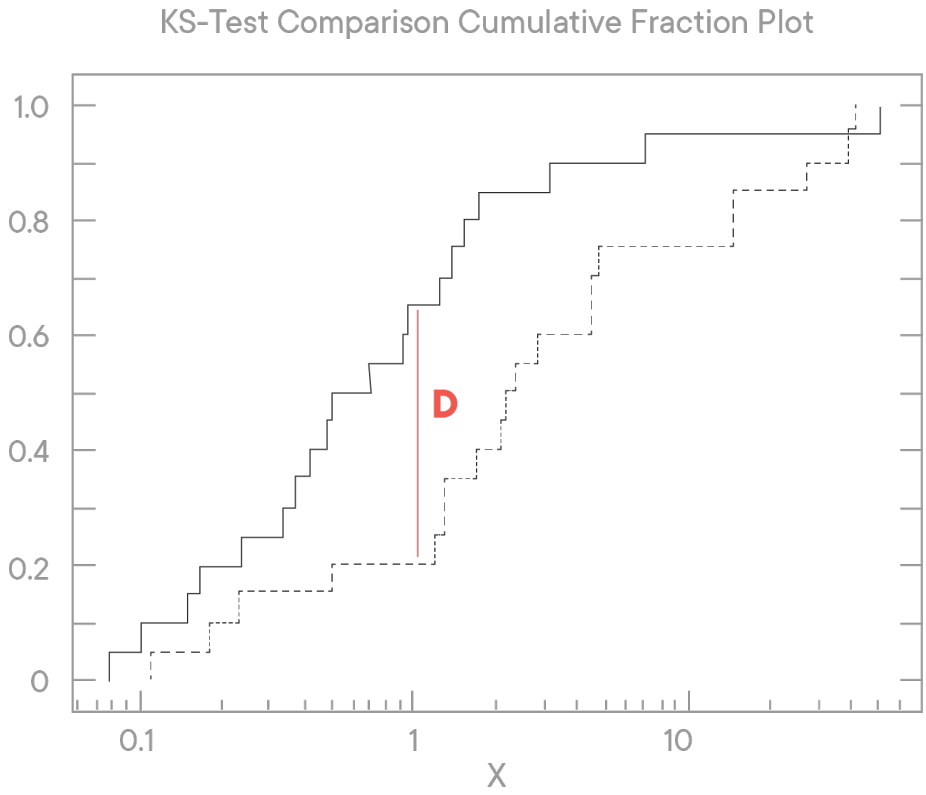
This is also known as the **Kolmogorov-Smirnov Goodness of Fit test**. It calculates the similarity between an observed (empirical) distribution and a completely specified theoretical continuous distribution. It is sensitive to all attributes of a distribution including mean, variance, and shape.

The key assumption of the one-sample test is that the theoretical distribution is fully defined continuous distribution, in terms of its parameters. This obviously means that its most common use case is that of testing normality. The test statistic, d, is simply the largest deviation between the observed cumulative function and the expected theoretical cumulative frequency distribution, i.e.

$$d = max(abs[F_0(X) - F_r(X)])$$

where

- **d** is the maximum deviation Kolmogorov statistic
- **$F_0(X)$** = (No.of observations ≤ X)/(Total no.of observations) i.e. the non parametric empirical distribution
- **$F_r(X)$** = The theoretical frequency distribution of X - parametric (e.g. based on mean value)



Null Hypothesis: There is no difference between the distribution of our sample and a normal distribution.

Acceptance Criteria: If the calculated value is less than the critical value, accept the null hypothesis.

Rejection Criteria: If the calculated value is greater than the critical value, reject the null hypothesis.

Example

Problem Statement:

In a study done from various modules of a data science course with 60 students, equal number of students are samples from each module. These students are interviewed and their intention to join the advanced machine learning module was noted. Following shows how many students showed a positive intention

- Python (5)
- Data Visualizations (9)
- SQL (11)
- Statistics (16)
- NLP (19)

It was expected that 12 students from each module would join advanced ML.

Let's use K-S test to find if there is any difference among student classes with regard to their intention of joining the advanced machine learning module.

First, we need to set up our null hypothesis.

> H_0: There is no difference among students of different modules with respect to their intention of joining advanced ML.

Streams	No. of students interested in joining		FO(X)	Fr(X)	|FO(X)−FT(X)|
	Observed(O)	Theoretical(T)			
Python	5	12	5/60	12/60	7/60
Viz.	9	12	14/60	24/60	10/60
SQL	11	12	25/60	36/60	11/60
Stats	16	12	41/60	48/60	7/60
NLP	19	12	60/40	60/60	60/60

```
Total       n=60
```

According to the formula above,

$$d = max(abs[F_0(X) - F_1(X)])$$

$$d = 11/60 = 0.183$$

Here's the Smirnov d-statistic for reference:

n\α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

The table value of d at 5% significance level is given by

$$d(0.05) = \frac{1.36}{\sqrt{n}} = \frac{1.36}{\sqrt{60}} = 0.175$$

Since the calculated d value (0.183) is greater than the critical value (0.175), hence we reject the null hypothesis and conclude that there is a difference among students of different modules in their intention of joining the advanced ML course.

Two-Sample K-S Test

The two-sample K-S test checks if two **independent** samples have been drawn from the same population, or, equivalently, from two identical populations (X = Y).

As with the one-sample test, it is moderately sensitive to all parameters of the distribution. The one-tailed version of this test has a specific purpose i.e .to test whether values of one population are larger than values of another population. Similar to one-sample test, cumulative distributions are compared, but here two sample distributions are compared instead of a sample distribution and a theoretical distribution as we saw above. For the two-tailed version of the test, the test statistic (d) is the largest absolute deviation between the two observed cumulative step functions, irrespective of the direction of the difference.

> The null hypothesis states for this test that there is no difference between the two distributions. The d-statistic is calculated in the same manner as we saw above.

$$d = max[abs[F_{n1}(X) - F_{n2}(X)]]$$

- n1 = Observations from first sample.
- n2 = Observations from second sample.

When the cumulative distribution shows large maximum deviation d, it is a reflection of the difference between the two sample distributions.

The critical value of d for samples where n1=n2 and is ≤ 40, the K-S table for two sample case is used. When n1 and/or n2 > 40 then the K-S table for large samples of two-sample test should be used. The null hypothesis is accepted if the calculated value is less than the table value and vice-versa.

Thus, the use of any of these nonparametric tests helps a researcher to test the significance of his results when the characteristics of the target population are unknown or no assumptions had been made about them.

Example

Given two samples, test if their distributions are the same.

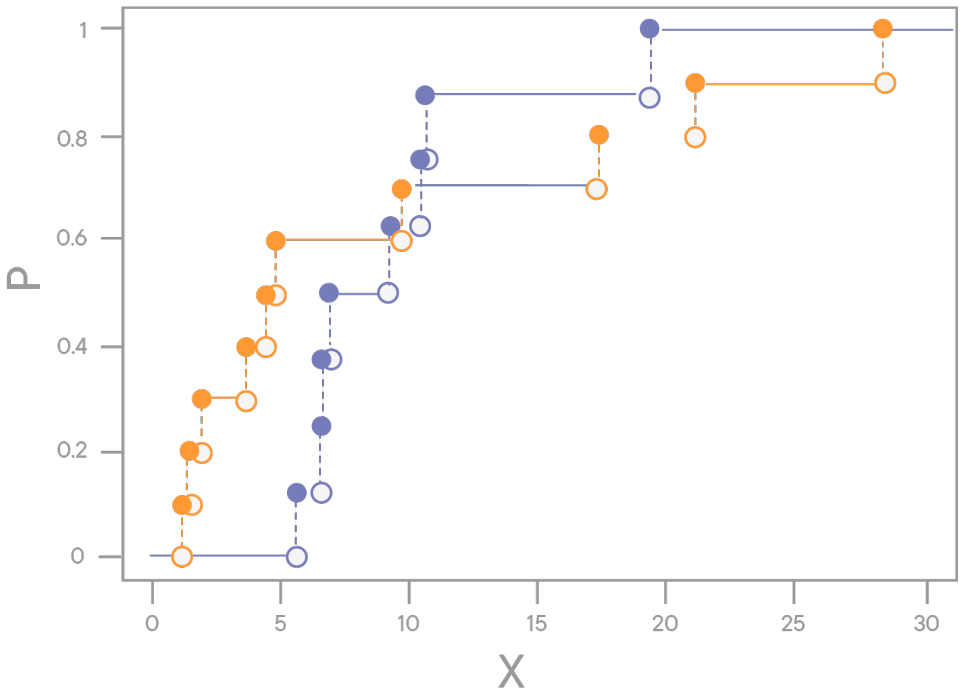
Compute the observed cumulative distribution functions of the two samples and compute their maximum difference.

```
X : 1.2, 1.4, 1.9, 3.7, 4.4, 4.8, 9.7, 17.3, 21.1, 28.4
Y : 5.6, 6.5, 6.6, 6.9, 9.2, 10.4, 10.6, 19.3
```

We sort the combined sample, in order to compute the empirical cdfs:

the combined sample, in order to compute the empirical cdf's:

```
   1.2 1.4 1.9 3.7 4.4 4.8 5.6 6.5 6.6 6.9 9.2 9.7 10.4 10.6 17.3 19.3 21.1 28.4
Fx 0.1 0.2 0.3 0.4 0.5 0.6 0.6 0.6 0.6 0.6 0.6 0.7 0.7  0.7  0.8  0.8  0.9  1.0
Fy 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.2 0.4 0.5 0.6 0.6 0.8  0.9  0.9  1.0  1.0  1.0
```



The Kolmogorov-Smirnov statistic is again the maximum absolute difference of the two observed distribution functions. From the above image, and also by feeding above values in the given formula, we get **d = 0.6**.

For two samples, the 95% critical value can be approximated by the formula:

$$d(0.05) = 1.36\sqrt{1/n_1 + 1/n_2} = 0.645$$

Since 0.6 < 0.645, we retain the null hypothesis in this case.

Kolmogorov-Smirnov tests have the advantages that:

- the distribution of statistic does not depend on cumulative distribution function being tested and
- the test is exact

They have the disadvantage that they are more sensitive to deviations near the center of the distribution than at the tails.

Summary

In this lesson, we looked at K-S test and how this test can be used to test for normality assumptions. We also looked at a one-sample K-S test and a two-sample K-S test with simple examples. Next, we'll see how to implement these tests in Python.