Methods that assume a particular form for the distribution of a r.v. and try to estimate the parameters of that particular distribution from a "random sample," i.e. i.i.d. r.v.'s $X_1, \ldots, X_n$ can be called parametric methods.

Here, we focus on the so-called non-parametric methods, that are generic and do not assume parametric forms for probability distributions.

# Empirical Distribution

**Definition:** Assume that $X_1, X_2, \cdots, X_n$ are i.i.d. samples with common CDF $F_X$. The empirical distribution function of $X_1, X_2, \cdots, X_n$ is defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[X_i, +\infty)}(x)$$

where $I_A$ is the indicator function of set $A$.

Informally, the empirical cdf

is increased by $\frac{1}{n}$ at each $X_i$.

Goodness-of-fit

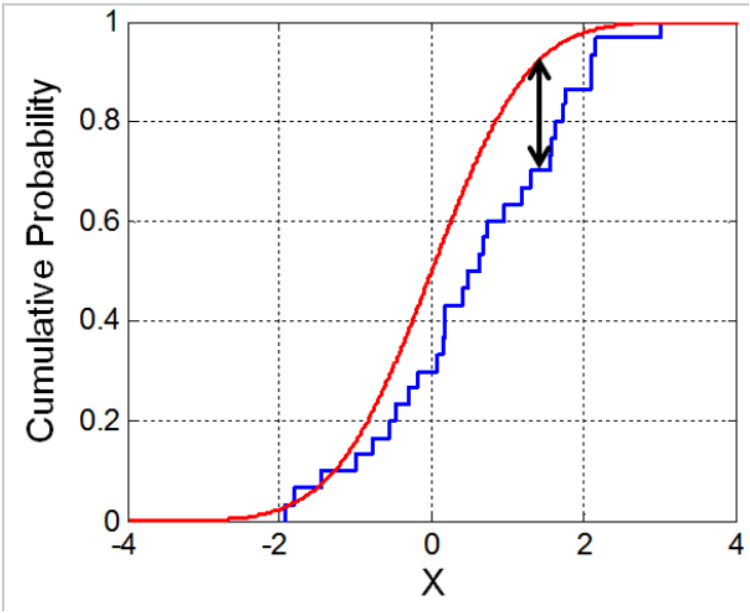
To measure how well a fitted distribution resembles the sample data, we can use the Kolmogorov-Smirnov (K-S) test statistic:

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|$$

$\hat{F}_n(x)$: empirical distribution

$F(x)$: fitted distribution

$$D_n = \sup_x |F_n(x) - F(x)|$$



## The Glivenko–Cantelli Theorem

Theorem: Assume that $X_1, X_2, \dots, X_n$ are i.i.d. samples with common cdf $F_X(x)$ and the empirical cdf $\hat{F}_n(x)$. Then the

Kolmogorov-Smirnov Statistic

$$D_n = \sup_x |\hat{F}_n(x) - F(x)|$$

converges to zero almost surely for all $x$, i.e.

$$\mathbb{P}\left(\lim_{n \to \infty} D_n \neq 0\right) = 0.$$

Remark: $\hat{F}_n(x)$ is a random variable at each $x$. $D_n$ is also a r.v.

Remark: The G-K Theorem is sometimes called the fundamental theorem of statistics, and is a basis for the Kolmogorov-Smirnov test.

Assume that multiple distributions are estimated from i.i.d samples using different (parametric or non-parametric) methods. Which distribution is "the best?"

The Kolmogorov–Smirnov Test Criterion says that the KS statistic has to be computed for each distribution and the distribution with the minimum $D_n$ has to be selected.

# Goodness of fit as hypothesis Testing

$H_0$: All samples $X_1, \cdots, X_n$ come from the cdf $F$

$H_1$: At least one of the samples $X_1, \cdots, X_n$ <u>does not</u> come from the cdf $F$

Under the null hypothesis (assuming $H_0$ is true)

the K-S test statistic $D_n$ follows a K-S distribution, which has a complicated form.

Therefore, to reject the null, we have to observe a sufficiently large $D_n$ so that we are in the rejection region.

K-S distribution has a complicated form.

$$P\left(D_n < \frac{1}{2n} + v\right) = n! \int_{\frac{1}{2n}-v}^{\frac{1}{2n}+v} \int_{\frac{3}{2n}-v}^{\frac{3}{2n}+v} \cdots \int_{\frac{2n-1}{2n}-v}^{\frac{2n-1}{2n}+v} g(u_1, u_2, \ldots, u_n)\, du_n \ldots du_2\, du_1 \qquad 0 \le v \le \frac{2n-1}{2n}$$

$$g(u_1, u_2, \ldots, u_n) = 1 \text{ over } 0 \le u_1 \le u_2 \le \ldots \le u_n \le 1,$$

And 0 otherwise

Birnbaum, Z.W. (1952), "Numerical Tabulation of the Distribution of Kolomogorov's Statistic for Finite Sample Size," Journal of the American Statistical Association, 47, 425–441

$$F_{D_6}(t) = \begin{cases} 0 & t < \frac{1}{12} \\ 46080t^6 - 23040t^5 + 4800t^4 - \frac{1600}{3}t^3 + \frac{100}{3}t^2 - \frac{10}{9}t + \frac{5}{324} & \frac{1}{12} \le t < \frac{1}{6} \\ 2880t^6 - 4800t^5 + 2360t^4 - \frac{1280}{3}t^3 + \frac{235}{9}t^2 + \frac{10}{27}t - \frac{5}{81} & \frac{1}{6} \le t < \frac{1}{4} \\ 320t^6 + 320t^5 - \frac{2600}{3}t^4 + \frac{4240}{9}t^3 - \frac{785}{9}t^2 + \frac{145}{27}t - \frac{35}{1296} & \frac{1}{4} \le t < \frac{1}{3} \\ -280t^6 + 560t^5 - \frac{1115}{3}t^4 + \frac{515}{9}t^3 + \frac{1525}{54}t^2 - \frac{565}{81}t + \frac{5}{16} & \frac{1}{3} \le t < \frac{5}{12} \\ 104t^6 - 240t^5 + 295t^4 - \frac{1985}{9}t^3 + \frac{775}{9}t^2 - \frac{7645}{648}t + \frac{5}{16} & \frac{5}{12} \le t < \frac{1}{2} \\ -20t^6 + 32t^5 - \frac{185}{9}t^3 + \frac{175}{36}t^2 + \frac{3371}{648}t - 1 & \frac{1}{2} \le t < \frac{2}{3} \\ 10t^6 - 38t^5 + \frac{160}{3}t^4 - \frac{265}{9}t^3 - \frac{115}{108}t^2 + \frac{4651}{648}t - 1 & \frac{2}{3} \le t < \frac{5}{6} \\ -2t^6 + 12t^5 - 30t^4 + 40t^3 - 30t^2 + 12t - 1 & \frac{5}{6} \le t < 1 \\ 1 & t \ge 1. \end{cases}$$

Fortunately, there are tables for the critical values of the distribution of K–S test.

| 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 | $\frac{1.94947}{\sqrt{n}}$ | $\frac{1.62762}{\sqrt{n}}$ | $\frac{1.51743}{\sqrt{n}}$ | $\frac{1.35810}{\sqrt{n}}$ | $\frac{1.22385}{\sqrt{n}}$ | $\frac{1.13795}{\sqrt{n}}$ | $\frac{1.07275}{\sqrt{n}}$ |

K-S is distribution free, i.e. it does not assume any particular distribution for the data

Drawback: Only fore one dimensional distributions

There are other goodness of fit tests including the Anderson-Darling test.

# The Chi-Squared Goodness of Fit Test

This test is based on the histogram of data. The idea is that if the pmf or pdf estimated for a sample is too different from its histogram, it probably is not a good fit for the data.

In particular, if $\mathbb{P}(e_j \leq X_i \leq e_{j+1})$ calculated according to the estimated pmf/pdf is not close to the histogram estimate of $\mathbb{P}(e_j \leq X_i \leq e_{j+1})$, which is $\frac{H_j}{n}$, the estimated pdf/pmf is not a good fit.

It can be shown that the quantity

$$\chi^2 = \sum_{j=1}^{m} \left( \frac{|H_j - np_j|^2}{np_j} \right)$$

has a Chi-squared distributia

with $m-1$ degrees of freedom when $n$ is large.

This is because

$$\frac{H_j - np_j}{\sqrt{np_j}}$$

has a standard normal distributia

when $n$ is large.

Therefore, if

$$\chi^2 > \chi^2_{m-1,\alpha}$$

(i.e. if the discrepancy is too large), we reject the null that the pdf dpmt we found is a good fit for the data.