Fitting distributions to a set of data

Davood Shahsavani
Dept. Of Mathematis, Division of Statistics
Shahrood Univ. Of Technlogy

overview

Common problem in risk analysis:

fitting a probability distribution to a set of observations for a variable, like the lifetime of a mechanical or electrical component

usage

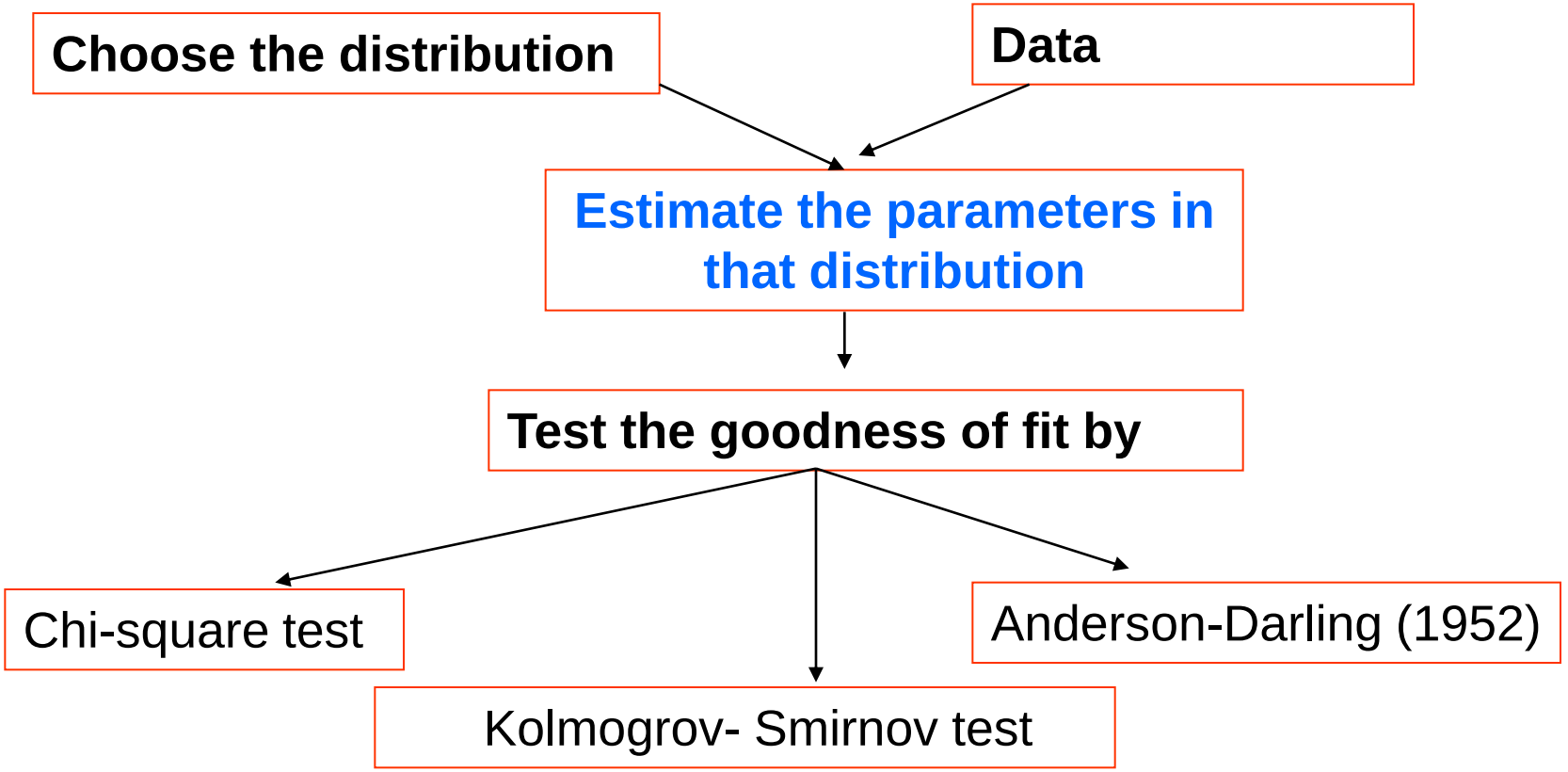
to be able to make forecasts about the future.

The principle behind fitting distributions:

finding the type of distribution (normal, lognormal, gamma, beta, etc) and the value of the parameters (mean, variance, etc) that give the highest probability of producing the observed data.

How to check the goodness of fit

How to check whether a predetermined distribution is a good fit to a set of data?



How to estimate the parameters of a distribution

What is the maximum likelihood estimator (MLE) of a parameter in a distribution?

Answer: The MLE of parameter α in a population in which the variable X has the density $f(x, \alpha)$ is a value of α **that** maximizes the joint probability density for the observed data $X1, X2, \ldots Xn$

Likelihood function or joint prob density of x1,,,,xn

$$L(\alpha) = f(x_1, \alpha) f(x_2, \alpha) \ldots\ldots\ldots f(x_n, \alpha)$$

α- hat is the estimator of α *if*

$$L(\hat{\alpha}) = \max_{\alpha} L(\alpha)$$

MLE for exponential distribution

Suppose that $X_1,\ldots,X_n$ have exponential distribution with the pdf

$$f(x_i,\alpha)=\frac{1}{\alpha}Exp(-x_i/\alpha)$$

$$L(\alpha)=f(x_1,\alpha)f(x_2,\alpha).....f(x_n,\alpha)$$

$$L(\alpha)=\frac{1}{\alpha^n}Exp(-\frac{\sum x_i}{\alpha})=\alpha^{-n}Exp(\frac{-n\bar{x}}{\alpha})$$

Maximization of L is the same as maximization of Ln(L)

$$Ln(L(\alpha))=-n\ln(\alpha)-\frac{n\bar{x}}{\alpha}$$

$$\frac{d}{d\alpha}Ln(L(\alpha))=0 \text{ ---> } \frac{-n}{\alpha}+\frac{n\bar{x}}{\alpha^2}=0$$

$$\longrightarrow \quad \boxed{\alpha=\bar{x}}$$

$$\frac{d^2}{d\alpha^2}Ln(L(\alpha))=0 \text{ ---> } \frac{n}{\alpha^2}-\frac{2n\bar{x}}{\alpha^3}$$

$$\frac{d^2}{d\alpha^2}Ln(L(\bar{x}))=\frac{n}{\bar{x}^2}-\frac{2n\bar{x}}{\bar{x}^3}=\frac{-1}{\bar{x}^2}<0$$

$$\longrightarrow \quad L(\bar{x})=\max_{\alpha} L(\alpha)$$

$\longrightarrow$ $\bar{X}$ Is the mL estimator of α

MLE for Parameters of Normal distribution

Suppose that $X_1,\ldots,X_n$ have N

$$f(x_i)=\frac{1}{\sigma\sqrt{2\pi}}Exp(-\frac{1}{2\sigma^2}(x_i-\mu)^2)$$

$$L(\mu,\sigma^2)=f(x_1)f(x_2).....f(x_n)$$

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}f(x_i)=\prod_{i=1}^{n}(2\pi\sigma^2)^{-0.5}\,Exp(-\frac{1}{2\sigma^2}(x_i-\mu)^2)$$

$$L(\mu,\sigma^2)=(2\pi\sigma^2)^{-n/2}Exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2)$$

$$\ln(L)=-(n/2)\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\frac{\partial\ln(L(\mu,\sigma^2))}{\partial\mu}=\sum_{i=1}^{n}(x_i-\mu)=0\longrightarrow\ \hat{\mu}=\bar{x}$$

$$\frac{\partial\ln(L)}{\partial\sigma^2}=-(n/2)(\frac{1}{\sigma^2})+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=0$$

$$\hat{\mu}=\bar{x}$$

$$\hat{\sigma}^2=s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$L(\bar{x},s^2)=\max\ L(\mu,\sigma^2)$$

Gamma and Ki-square distribution

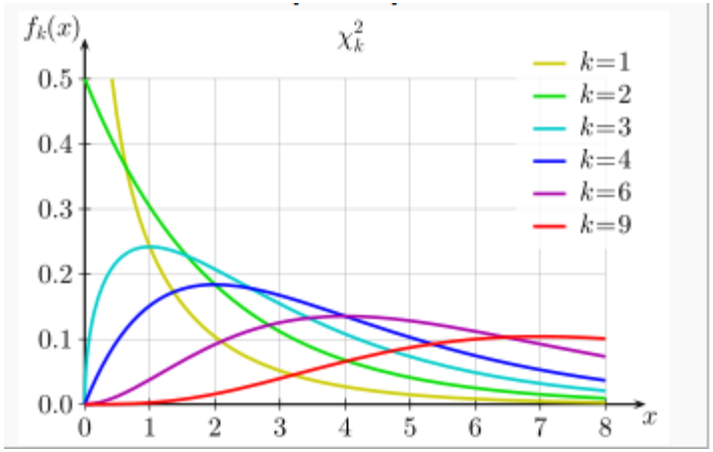
A continues random variable X has ***gamma distribution*** with parameters α and β if

$$f(x)=\begin{cases}\dfrac{1}{\Gamma(\alpha)\beta^{\alpha}}\,x^{\alpha-1}\,e^{-\frac{x}{\beta}} & x>0\\ 0 & x<0\end{cases}$$

If α=k/2 and β=2 then it is said that the random variable X has ***chi-square distribution*** with *k* degree of freedom

$$f(x)=\frac{1}{\Gamma(k/2)2^{k/2}}\,x^{k/2-1}\,e^{-\frac{x}{2}} \qquad x>0$$

$$X \approx \chi_k^2$$



Chi-square goodness of fit statistic

This statistic measures how well the expected frequency of the fitted distribution compares with the observed frequency of a histogram of the observed data.

assumptions:

1. The observed data consists of a random sample of *n* independent data points.
2. The measurement scale can be nominal (i.e. non-numeric) or numerical.
3. The *n* data points can be arranged into histogram form with *N* contiguous classes that cover the entire possible range of the variable.

Chi-square statistic

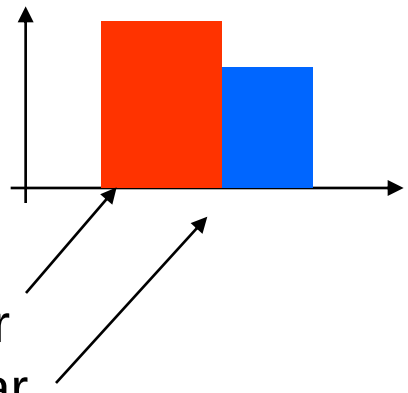
The data can be grouped into intervals of *equal probability* or *equal width*.

K=is the number of bins
suggestion: $k = 1 + \log_2 N$

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

O_i : The observed frequency of the i-th histogram class or bar
E_i : The expected frequency from the fitted distribution of *x*-values falling within the *x*-range of the i-th histogram bar

$E_i = \{F(i_{max}) - F(i_{min})\} * n$



F(*x*) = distribution function of the fitted distribution
(i_{min}) = the *x*-value lower bound of the i-th histogram bar
(i_{max}) = the *x*-value upper bound of the i-th histogram bar

Chi-square test

Hypothesis Testing

- H_0: the data follow the specified distribution;
  H_A: the data do not follow the specified distribution

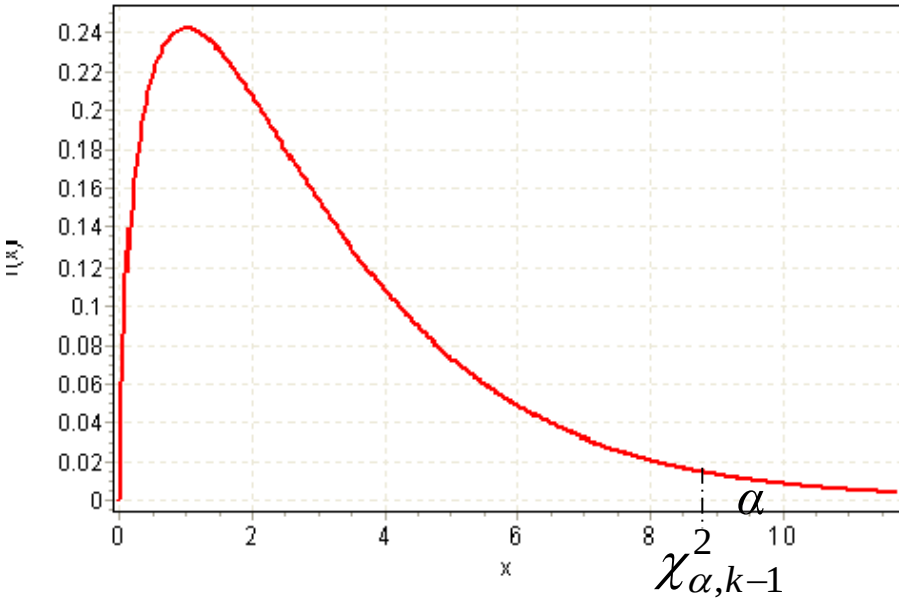
Decision rule:

1- Determine the level of error ($\alpha=0.05$)

2- Determine the corresponding value from chi square distribution $\chi^2_{\alpha,k-1}$

3- determine the chi-sqaure statistic

$$\chi^2_{observed} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$



The H_0 is rejected at the chosen significance level (1- α) if the test statistic is greater than the critical value defined as

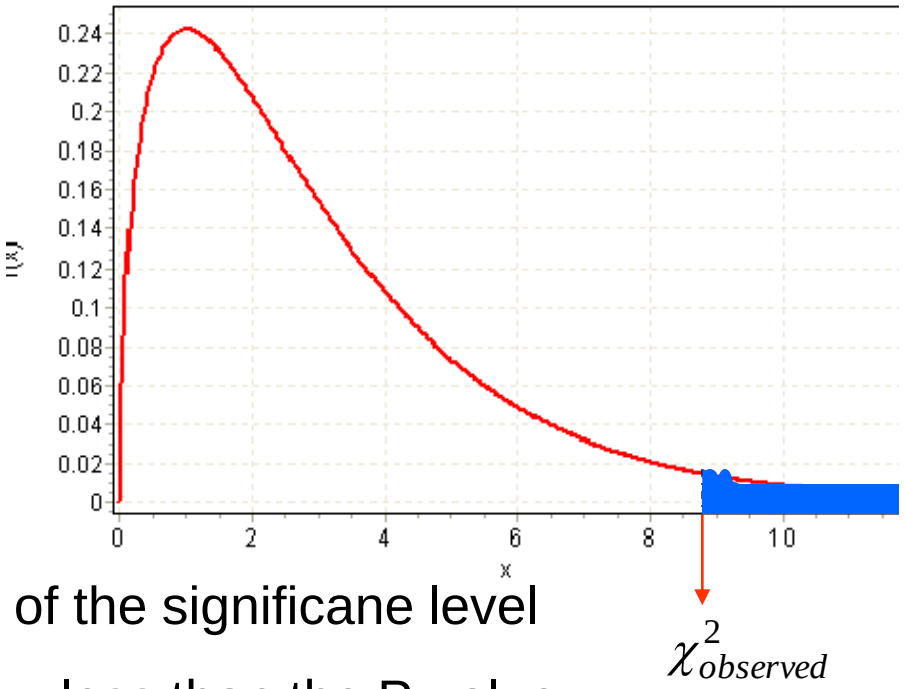
$$\text{Re } ject \;\; H_0 \; if \quad \chi^2_{\alpha,k-1} < \chi^2_{observed}$$

P- value

Suppose that X is distributed by chi-square distribution with k-1 degree of freedom

$$X \approx \chi^2_{k-1}$$

$$P-value = P(X > \chi^2_{observed})$$



P-value denotes the threshold value of the significane level

H_0 will be accepted for all values of α less than the P-value.

For example, if P=0.025, the null hypothesis will be accepted at all significance levels less than P (i.e. 0.01 and 0.02), and rejected at higher levels, including 0.05 and 0.1.

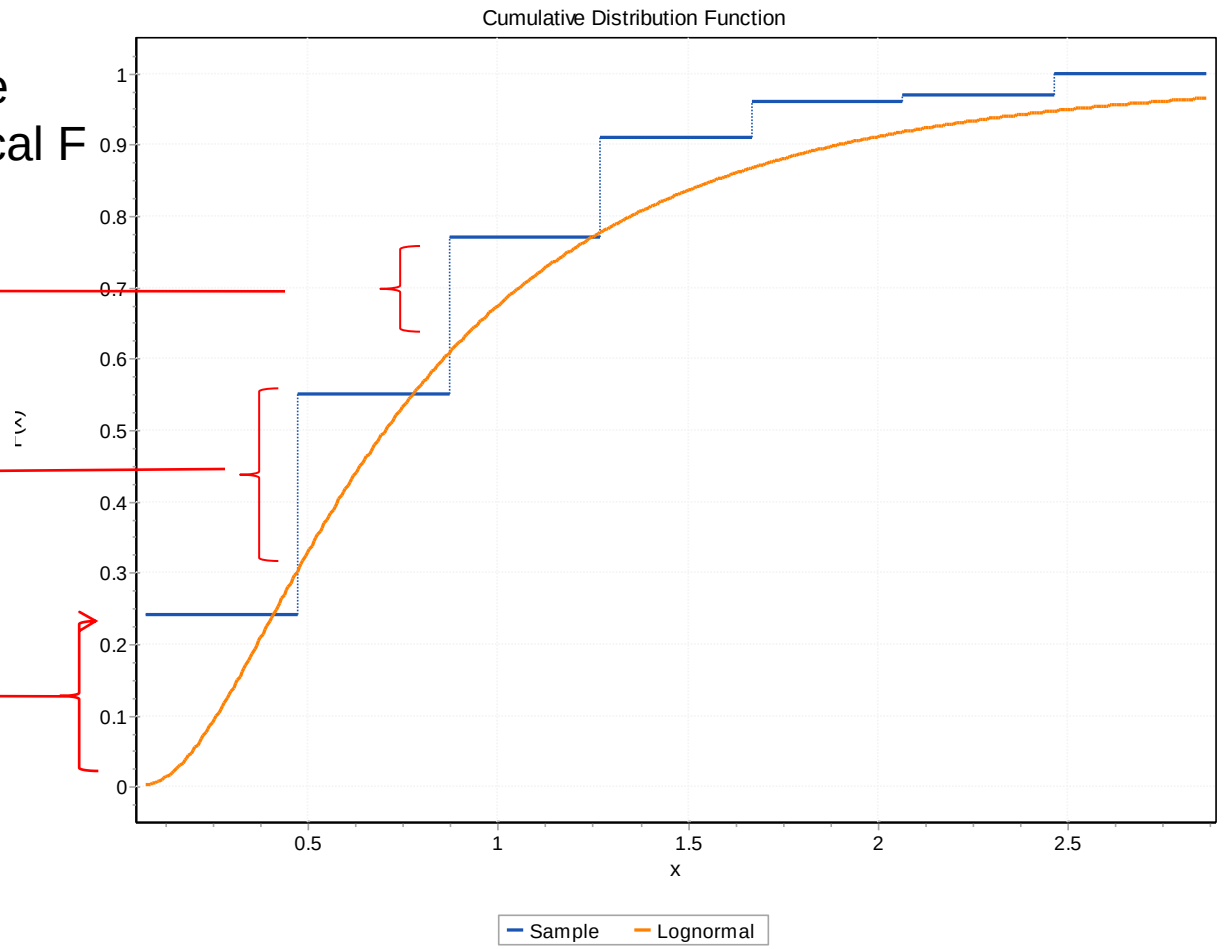
Kolmogrov - Smirnov test

K-S considers the maximum value of the difference between the empirical and theoretical F

$$F_n(x_3) - F(x_3)$$

$$F_n(x_2) - F(x_2)$$

$$F_n(x_1) - F(x_1)$$



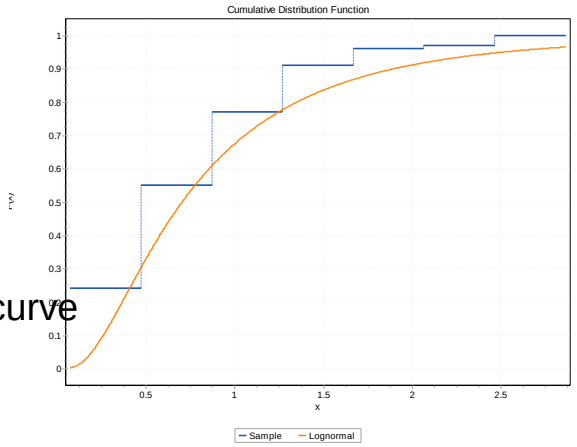
Kolmogrov - Smirnov statistic

The K–S statistic *Dn* is defined as:

$$Dn = \max[|Fn(x) - F(x)|]$$

stair

Continuous curve



Where *Dn* is known as the K–S distance

n = total number of data points

F(*x*) = cumulative distribution function of the fitted distribution

Fn(*x*) = *i*/*n*

i = the cumulative rank of the data point

Kolmogrov - Smirnov test

The hypothesis

H0: the data follow the specified distribution;

is rejected at significance level (1- α) if

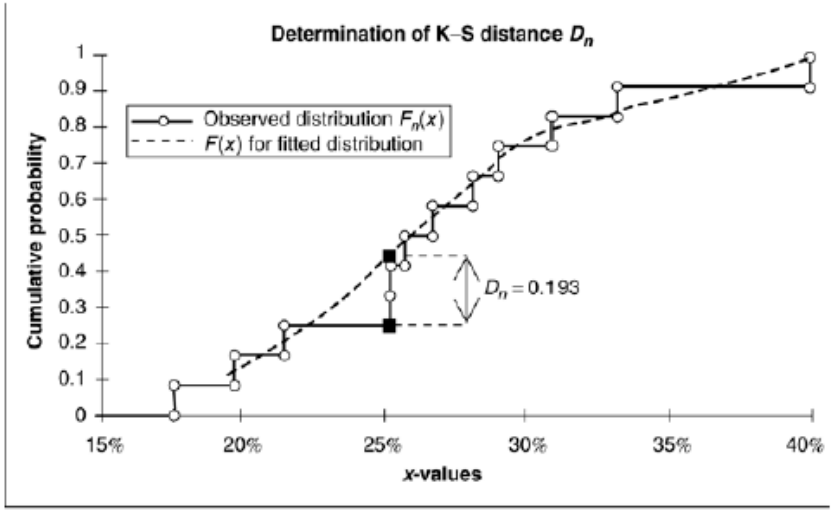
$$Dn = \max[|Fn(x) - F(x)|]$$

is greater than the critical value obtained from K_S table.

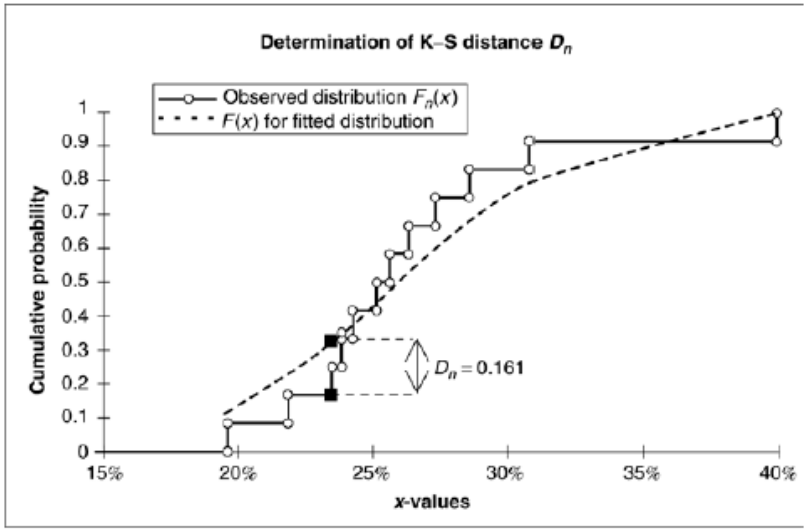
The fixed values of α (0.01, 0.05 etc.) are generally used to evaluate the null hypothesis (H_0) at various significance levels. A value of 0.05 is typically used for most applications, however, in some critical industries, a lower value may be applied.

.

Kolmogrov - Smirnov test



(a) Distribution is generally a good fit except in one particular area



(b) Distribution is generally a poor fit but with no single large discrepancies

Good fit

Poor fit

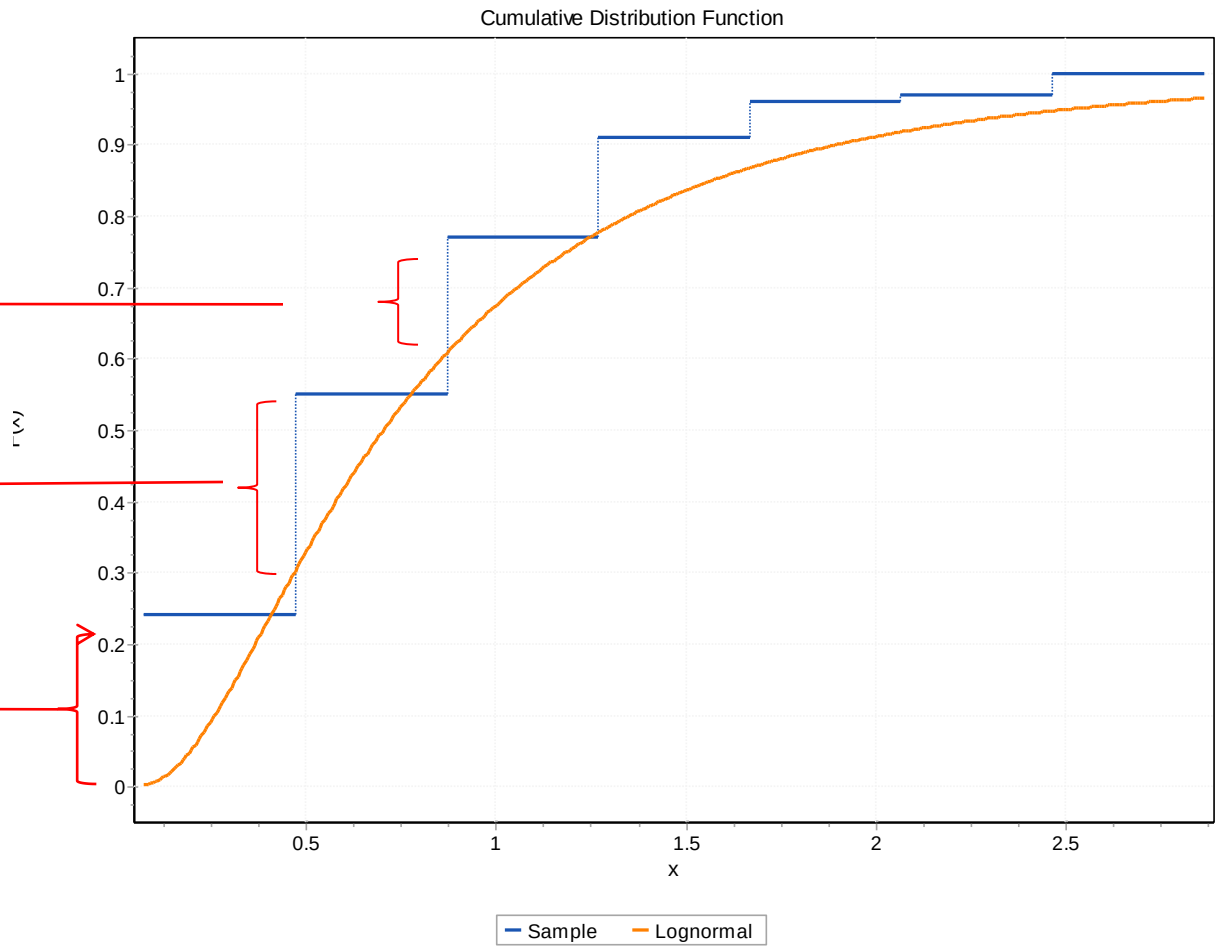
Anderson Darling Test

It consider all differences not only the maximum

$$F_n(x_3) - F(x_3)$$

$$F_n(x_2) - F(x_2)$$

$$F_n(x_1) - F(x_1)$$



Anderson Darling Test Statistics

$$A_n^2 = \int_{-\infty}^{\infty} |F_n(x) - F(x)|^2 \, \Psi(x) f(x) \, dx$$

It assign some special weights to each of the differences. The weight function is

$$\Psi(x) = \frac{n}{F(x)\{1 - F(x)\}}$$

where

n = total number of data points

$F(x)$ = distribution function of the fitted distribution

$f(x)$ = density function of the fitted distribution

$F_n(x) = i/n$

i = the cumulative rank of the data point

Anderson Darling in Practice

$$A^2 = -n - S\,,$$

$$S = \sum_{i=1}^{N} \frac{(2i-1)}{N} \left[\ln F(Y_i) + \ln\left(1 - F(Y_{N+1-i})\right)\right]$$

$$\{Y_1 < \cdots < Y_n\}$$