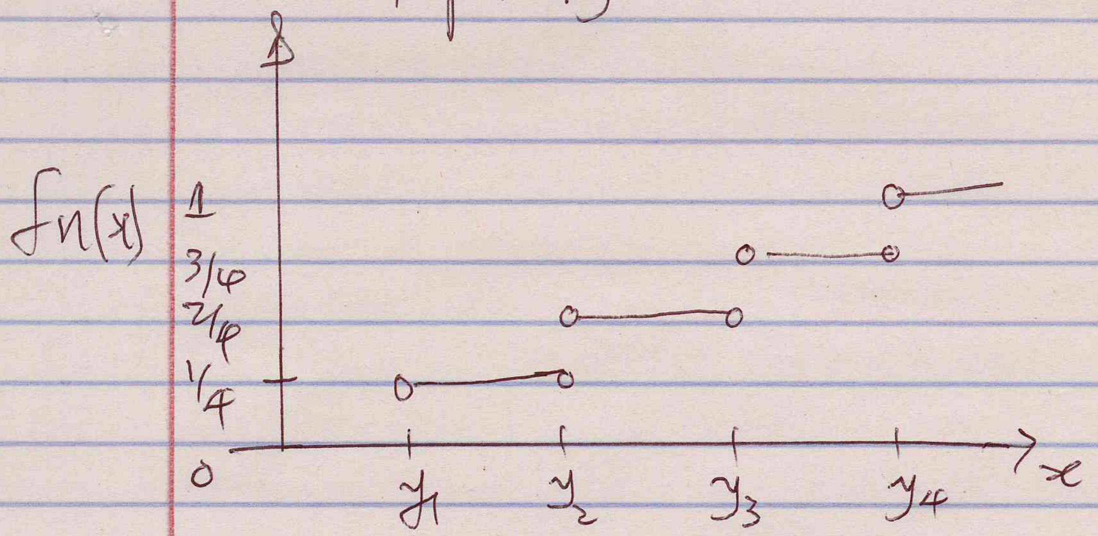
(H) ①

Kolmogorov-Smirnov test

- Suppose the random variables $x_1, x_2 \dots x_n$ is from a continuous random sample and let $x_1, x_2 \dots x_n$ be the observed values of $X_1, X_2 \dots X_n$.
- Since the observations come from a continuous distribution there is a probability of zero (0) that any of the observed values $x_1, x_2 \dots x_n$ will be equal. Thus we shall assume for simplicity that all the n-values are different.
- We consider a function $f_n(x)$ which ~~follows~~ is constructed from the values $x_1, x_2, \dots x_n$ as follows.
- For each number x $(-\infty < x < \infty)$ the value of $f_n(x)$ is defined as the proportion of observed values in the sample which are less than or equal to x
- In other words if exactly K of the observed values in the sample are less than or equal to x, then $f_n(x) = K/n$.
- The function $f_n(x)$ defined this way is called the sample distribution function (sdf).
- Sometimes $f_n(x)$ is called the empirical distribution function. If we let $y_1 < y_2 < \dots y_n$ denote the value of the order statistics of the sample, then $f_n(x) = 0$ for $x < y$
- $f_n(x)$ jumps to the value $1/n$ at $x = y$ and remains at $1/n$ for $y_1 \leq x \leq y_2$.
- $f_n(x)$ jumps to $2/n$ at $x = y_2$ and remains at $2/n$ for $y_2 \leq x \leq y_3$ and so on.

Graphically



- Now let $f(x)$ denote the distribution function of the distribution from which the sample was drawn.
- For a given number x $(-\infty < x < \infty)$ the probability that any particular x_i is less than or equal to x is $F(x)$.
- Therefore it follows that from the law of large numbers (LLN) that as $n \to \infty$ the proportion of $f_n(x)$ of the observations in the sample which are less than or equal to x will converge to $f(x)$ in probability.

$$P\left(\lim_{n \to \infty} f_n(x)\right) = f(x) \quad --- \quad 1$$

for $-\infty < x < \infty$

where n is the sample size of $x_1, x_2 \dots x_n$.
$f_n(x)$ is the sample distribution function.
$f(x)$ is the distribution function

- The relation 1 expresses the fact that at each point x, the sample distribution function $f_n(x)$ will converge to the actual $f(x)$ of the distribution from which the random sample was drawn.

Let $D_n = \sup_{-\infty < x < \infty} \left| F_n(x) - F(x) \right|$

- Suppose we now wish to test the simple hypothesis H_0: the unknown distribution function $f(x)$ is actually a partially continuous distribution $f^*(x)$ Vs the general alternative that the actual distribution function is not $f^*(x)$ i.e

$$\left.\begin{aligned} &H_0 : f_n(x) = f^*(x) \quad -\infty < x < \infty \\ &H_1 : \text{the hypothesis } H_0 \text{ is not true} \end{aligned}\right\} \quad (2)$$

- This is a non parametric problem because the unknown distribution from which the random sample was taken might be any continuous distribution.
- Let $f_n(x)$ denote the sample distribution function (sdf) and let

$$D_n = \sup_{-\infty < x < \infty} \left| f_n(x) - F^*(x) \right|$$

- In other words D_n is the maximum difference between the sdf $f_n(x)$ and the hypothesized $f^*(x)$
- When H_0 in eqn (2) is true the probability distribution of D_n will be a certain distribution which is the same for any possible continuous distribution
- $F^*(x)$ does not depend on a particular distribution function being studied in a specific problem.
- Tables of this distribution for various values of n (sample size) have been developed and are tabulated in collection of statistical tables. The tables gives the values of D_n^{α} such that;

$$P(D_n \le d_n^{\alpha}) = 1-\alpha$$

$$\Rightarrow P(D_n > d_n^{\alpha}) = \alpha$$

- Below are some critical values for the Kolmogorov–Smirnov test

n \ α	0.20	0.1	0.05	0.01
5	0.45	0.51	0.56	0.67
10	0.32	0.37	0.41	0.49
15	0.27	0.26	0.34	0.40
20	0.23	0.24	0.29	0.36
25	0.21	0.22	0.27	0.32
30	0.19	0.20	0.24	0.29
35	0.18	0.19	0.23	0.27
40	0.17	0.18	0.21	0.25
45	0.16	0.17	0.2	0.24
50	0.13	0.7	0.9	0.23
$n>50$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.03}{\sqrt{n}}$

Example

Test the hypothesis by the Kolmogorov Smirnov Test that the following sample come from a standard normal distribution

−1.23	1.64	−0.21	0.70	1.40	0.44	−0.07	−0.02
−0.15	1.76	1.62	0.40	−2.11	−0.99	−0.42	0.81
1.47	−2.46	0.88	1.39	0.42	0.27	−0.39	−0.10
1.07							

Solution

We want to test

$H_0: f(x) = f^*(x)$ for $-\infty < x < \infty$

$H_1:$ the hypothesis is not true.

where $f^*(x) = \Phi(x)$ the cdf of the standard normal distribution at $\alpha = 0.05$

$n = 25$, $D_{25}^{0.05} = 0.27$ (from tables)

We reject H_0 if $D_n \geq D_{25}^{0.05} = 0.27$

where $D_{25} = \sup\limits_{-\infty < x < \infty} \left| f_{25}(x) - f^*(x) \right|$

Arrange the sample values in ascending order and for each value determine $f_n(x)$ and $f^*(x)$.

i	x	$F_n(x)$	$f^*(x)$	$\lvert f_n(x) - f^*(x) \rvert$
1	−2.46	1/25 = 0.04	0.007	0.033
2	−2.11	2/25 = 0.08	0.017	0.063
3	−1.23	3/25 = 0.12	0.109	0.011
4	−0.99	4/25 = 0.16	0.161	0.001
5	−0.42	0.20	0.337	0.137 → sup
6	−0.39	0.24	0.348	0.108
7	−0.21	0.28	0.417	0.137 → sup
8	−0.15	0.32	0.440	0.120
9	−0.10	0.36	0.460	0.10
10	−0.07	0.40	0.472	0.072
11	−0.02	0.44	0.492	0.052
12	0.27	0.48	0.606	0.126
13	0.40	0.52	0.655	0.135
14	0.42	0.56	0.663	0.103
15	0.44	0.60	0.670	0.070
16	0.70	0.64	0.758	0.114
17	0.81	0.68	0.791	0.111

i	x	$F_n(x)$	$F^*(x)$	$F_n(x) - F^*(x)$
18	0.88	0.72	0.811	0.091
19	1.07	0.76	0.858	0.098
20	1.39	0.80	0.918	0.118
21	1.40	0.84	0.919	0.079
22	1.47	0.88	0.929	0.049
23	1.62	0.92	0.947	0.027
24	1.64	0.96	0.950	0.010
25	1.76	$\frac{25}{25} = 1.0$	0.961	0.039

Getting $f^*(x)$

$$f^*(x)(-2.46) = p(x \leq -2.46) = p\left(\frac{x-\mu}{\sqrt{\sigma^2}}\right)$$

$$= \frac{-2.46-\mu}{\sqrt{\sigma^2}}$$

$$\Rightarrow p\left(z \leq -\frac{2.46-0}{1}\right)$$ for ~~an~~ standard normal $z \sim N(0,1)$ $\mu = 0, \sigma^2 = 1$

$$\Rightarrow p(z \leq -2.46) = 1 - p(z \leq -2.46)$$

$$= 1 - \phi(2.46)$$

$$= 1 - 0.9931$$

$$= 0.0069$$

$$\simeq 0.007$$

$D_{25} = 0.137$ and $D_{25}^{0.05} = 0.27$

$D_{25} = 0.137 < D_{25}^{0.05} = 0.27$ hence we do not

reject H_0 and conclude that the sample came from a standard normal distribution.

Exercise

Test the hypothesis by Kolmogorov-Smirnov test that the following sample values came from a normal distribution with mean 2 and variance 4 at $\alpha = 0.01$

2.72	3.84	0.88	5.72	5.48	3.12	0.10	2.48
1.70	0.52	2.64	3.64	3.40	1.80	−0.52	−0.12
2.30	3.10	1.04	1.02				