Kolmogorov Smirnov One Sample Test

Sanchita Khan (STAT-24) Sumedha Guha (STAT-05) Shramana Guin (STAT-06) Suryadeep Ghosh (STAT-28)

Department of Statistics

Third Year, Semester-V

Contents

- 1. Goodness of Fit
- 2. Kolmogorov Smirnov One sample test
- 3. The Hypothesis
- 4. Empirical Distribution Function
- 5. Some statistical properties of $S_n(x)$
- 6. Glivenko Cantelli Theorem
- 7. Test Statistic
- 8. Results
- 9. Applications of the K-S One Sample Statistics
- 10. Strengths of The K-S Test
- 11. Drawbacks
- 12. Practical Method
- 13. Example
- 14. References
- 15. Acknowledgement



► There are many situations where experimenters need to know what is the distribution of the population of their interest. In classical statistics, information about the form generally must be postulated in the null hypothesis to perform an exact parametric type of inference.

- ▶ There are many situations where experimenters need to know what is the distribution of the population of their interest. In classical statistics, information about the form generally must be postulated in the null hypothesis to perform an exact parametric type of inference.
- ► For example if they want to use a parametric test it is often assumed that the population under investigation is normal. The compatibility of a set of observed sample values with a normal distribution or any other distribution can be checked by a goodness of fit type of test.

- ▶ There are many situations where experimenters need to know what is the distribution of the population of their interest. In classical statistics, information about the form generally must be postulated in the null hypothesis to perform an exact parametric type of inference.
- For example if they want to use a parametric test it is often assumed that the population under investigation is normal. The compatibility of a set of observed sample values with a normal distribution or any other distribution can be checked by a goodness of fit type of test.
- ▶ The goodness of fit test is used to test if sample data fits a distribution from a certain population (i.e. a population with a normal distribution or one with a Weibull distribution).

- ► There are many situations where experimenters need to know what is the distribution of the population of their interest. In classical statistics, information about the form generally must be postulated in the null hypothesis to perform an exact parametric type of inference.
- For example if they want to use a parametric test it is often assumed that the population under investigation is normal. The compatibility of a set of observed sample values with a normal distribution or any other distribution can be checked by a goodness of fit type of test.
- ► The goodness of fit test is used to test if sample data fits a distribution from a certain population (i.e. a population with a normal distribution or one with a Weibull distribution).
- ► Goodness of fit checks whether there is any significant differences between an observed frequency distribution and a given theoretical (expected) frequency distribution.



▶ There are two types of goodness of fit tests:

- ▶ There are two types of goodness of fit tests:
 - Designed for null hypothesis concerning a discrete distribution. Pearsonian chi square test.

- ▶ There are two types of goodness of fit tests:
 - Designed for null hypothesis concerning a discrete distribution. Pearsonian chi square test.
 - Designed for null hypothesis concerning a continuous distribution
 - Kolmogorov Smirnov test and Lilliefor's test.

Kolmogorov Smirnov One sample test

▶ It is named after Andrey Kolmogorov and Nikolai Smirnoff.

Figure: Nikolai Smirnov and Andrey Kolmogorov





The Hypothesis is,

Let $x_1, x_2, ..., x_n$ be observations on continuous i.i.d random variables $X_1, X_2, ..., X_n$ with cdf F_X

The Hypothesis is,

- Let $x_1, x_2, ..., x_n$ be observations on continuous i.i.d random variables $X_1, X_2, ..., X_n$ with cdf F_X
- We want to test the hypothesis

$$H_0: F(x) = F_0(x)$$
 for all x

VS

 $H_1: H_0$ is not true for at least one x

The Hypothesis is,

- Let $x_1, x_2, ..., x_n$ be observations on continuous i.i.d random variables $X_1, X_2, ..., X_n$ with cdf F_X
- We want to test the hypothesis $H_0: F(x) = F_0(x)$ for all x vs
 - $H_1: H_0$ is not true for at least one x
- ▶ The basis of this test is that it relates the distance between the hypothesized cumulative distribution function and the empirical distribution function of the sample as a number, *D*, which is then compared to the critical *D*-value for that data distribution.

The Hypothesis is,

- Let $x_1, x_2, ..., x_n$ be observations on continuous i.i.d random variables $X_1, X_2, ..., X_n$ with cdf F_X
- We want to test the hypothesis $H_0: F(x) = F_0(x)$ for all x vs

 $H_1: H_0$ is not true for at least one x

- ▶ The basis of this test is that it relates the distance between the hypothesized cumulative distribution function and the empirical distribution function of the sample as a number, *D*, which is then compared to the critical *D*-value for that data distribution.
- Before jumping into the test, let us take a look at some points.

▶ The empirical distribution function (denoted by $S_n(x)$) of the sample is defined as the proportion of sample observations that are less than or equal of x for all real numbers x, that is,

- ▶ The empirical distribution function (denoted by $S_n(x)$) of the sample is defined as the proportion of sample observations that are less than or equal of x for all real numbers x, that is,
- ► $S_n(x) = \frac{\text{number of sample values } \leq x}{n}$

- ▶ The empirical distribution function (denoted by $S_n(x)$) of the sample is defined as the proportion of sample observations that are less than or equal of x for all real numbers x, that is,
- $S_n(x) = \frac{\text{number of sample values } \leq x}{n}$
- Or in terms of order statistics

$$S_n(x) = \begin{cases} 0 & \text{if } x \leqslant X_{(1)} \\ \frac{i}{n} & X_{(i)} \leqslant x \leqslant X_{(i+1)} \ i = 1, 2, ..., n-1 \\ 1 & x \geqslant X_{(n)} \end{cases}$$

- ▶ The empirical distribution function (denoted by $S_n(x)$) of the sample is defined as the proportion of sample observations that are less than or equal of x for all real numbers x, that is,
- $S_n(x) = \frac{\text{number of sample values } \leq x}{n}$
- Or in terms of order statistics

$$S_n(x) = \begin{cases} 0 & \text{if } x \leqslant X_{(1)} \\ \frac{i}{n} & X_{(i)} \leqslant x \leqslant X_{(i+1)} \ i = 1, 2, ..., n-1 \\ 1 & x \geqslant X_{(n)} \end{cases}$$

▶ In case of tied observations, the edf is still a step function but it jumps only at the ordered sample values $X_{(j)}$ and the height of the jump is equal to k/n, where k is the number of values tied at $X_{(j)}$.

- ▶ The empirical distribution function (denoted by $S_n(x)$) of the sample is defined as the proportion of sample observations that are less than or equal of x for all real numbers x, that is,
- $S_n(x) = \frac{\text{number of sample values } \leq x}{n}$
- Or in terms of order statistics

$$S_n(x) = \begin{cases} 0 & \text{if } x \leqslant X_{(1)} \\ \frac{i}{n} & X_{(i)} \leqslant x \leqslant X_{(i+1)} \ i = 1, 2, ..., n-1 \\ 1 & x \geqslant X_{(n)} \end{cases}$$

- ▶ In case of tied observations, the edf is still a step function but it jumps only at the ordered sample values $X_{(j)}$ and the height of the jump is equal to k/n, where k is the number of values tied at $X_{(j)}$.
- In terms of indicator variables we can express it as $S_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \le x)$



Some statistical properties of $S_n(x)$

$$T_n(x) = n.S_n(x) \sim Bin(n, F_0(x)).$$

Some statistical properties of $S_n(x)$

- $ightharpoonup T_n(x) = n.S_n(x) \sim \text{Bin}(n, F_0(x)).$
- ► Therefore $E(S_n(x)) = F_X(x)$ and $Var(S_n(x)) = \frac{F_X(x)(1-F_X(x))}{n} \longrightarrow 0$ as $n \longrightarrow \infty$,

Some statistical properties of $S_n(x)$

- $ightharpoonup T_n(x) = n.S_n(x) \sim \text{Bin}(n, F_0(x)).$
- ► Therefore $E(S_n(x)) = F_X(x)$ and $Var(S_n(x)) = \frac{F_X(x)(1 F_X(x))}{n} \longrightarrow 0$ as $n \longrightarrow \infty$,
- which implies for any fixed value x, $S_n(x)$ is a consistent estimator of $F_X(x)$

2. Glivenko Cantelli Theorem

▶ $S_n(x)$ converges uniformly to $F_x(.)$ with probability 1,

2. Glivenko Cantelli Theorem

- ▶ $S_n(x)$ converges uniformly to $F_x(.)$ with probability 1,
- ▶ i.e

$$P\{\lim_{n\to\infty}\sup_{-\infty< x<\infty}[|S_n(x)-F(x)|]=0\}=1$$

According to Glivenko- Cantelli Theorem, as $n \longrightarrow \infty$, $S_n(x)$ approaches the cdf $F_X(x)$ for all x.

- According to Glivenko- Cantelli Theorem, as $n \longrightarrow \infty$, $S_n(x)$ approaches the cdf $F_X(x)$ for all x.
- ightharpoonup Therefore, for large n,

- According to Glivenko- Cantelli Theorem, as $n \longrightarrow \infty$, $S_n(x)$ approaches the cdf $F_X(x)$ for all x.
- ► Therefore, for large *n*,
- ▶ the deviations between the true function and the statistical image $|S_n(x) F_X(x)|$ should be small for all values of x.

- According to Glivenko- Cantelli Theorem, as $n \longrightarrow \infty$, $S_n(x)$ approaches the cdf $F_X(x)$ for all x.
- ► Therefore, for large *n*,
- ▶ the deviations between the true function and the statistical image $|S_n(x) F_X(x)|$ should be small for all values of x.
- ► This results that,

- According to Glivenko- Cantelli Theorem, as $n \longrightarrow \infty$, $S_n(x)$ approaches the cdf $F_X(x)$ for all x.
- ► Therefore, for large *n*,
- ▶ the deviations between the true function and the statistical image $|S_n(x) F_X(x)|$ should be small for all values of x.
- This results that,
- ▶ if H_0 is true, the statistic $D_n = \sup_{x} |S_n(x) F_0(x)|$ is, for any n ,a reasonable measure of our estimate.

- According to Glivenko- Cantelli Theorem, as $n \longrightarrow \infty$, $S_n(x)$ approaches the cdf $F_X(x)$ for all x.
- ► Therefore, for large *n*,
- ▶ the deviations between the true function and the statistical image $|S_n(x) F_X(x)|$ should be small for all values of x.
- This results that,
- ▶ if H_0 is true, the statistic $D_n = \sup_{x} |S_n(x) F_0(x)|$ is, for any n ,a reasonable measure of our estimate.
- ▶ This D_n statistic , called K-S one sample statistic.

$$D_n^+ = \sup_{\substack{x \\ N_n^- = \sup_{x} (F_0(x) - S_n(x))}} S_n^- = \sup_{x} (F_0(x) - S_n(x))$$

are called One-sided K-S Statistic.

$$D_n^+ = \sup_{x} (S_n(x) - F_0(x))$$

$$D_n^- = \sup_{x} (F_0(x) - S_n(x))$$

are called One-sided K-S Statistic.

▶ D_n is particularly useful in non-parametric statistical inference because the probability distribution of D_n does not depend on $F_0(x)$ as long as F_0 is continuous.

Let us prove the previously stated statement.

- Let us prove the previously stated statement.
- Let us define the inverse of F_0 by $F_0^{-1}(y) = min\{x : F_0(x) \ge y\}.$

- Let us prove the previously stated statement.
- Let us define the inverse of F_0 by $F_0^{-1}(y) = min\{x : F_0(x) \ge y\}.$
- ► Then by making change of variables $y = F_0(x)$ or $x = F_0^{-1}(y)$ we can write,

$$P(D_n \leqslant d)$$

$$= P(\sup_{x} |S_n(x) - F_0(x)| \leqslant d)$$

$$= P(\sup_{0 < y < 1} |S_n(F_0^{-1}(y)) - y| \leqslant d)$$

- Let us prove the previously stated statement.
- Let us define the inverse of F_0 by $F_0^{-1}(y) = min\{x : F_0(x) \ge y\}.$
- ► Then by making change of variables $y = F_0(x)$ or $x = F_0^{-1}(y)$ we can write,

$$P(D_n \leqslant d)$$

$$= P(\sup_{x} |S_n(x) - F_0(x)| \leqslant d)$$

$$= P(\sup_{0 < y < 1} |S_n(F_0^{-1}(y)) - y| \leqslant d)$$

▶ Therefore, D_n is a distribution free statistic.

Now using the definition of the empirical cdf S_n we can write,

$$S_n(F_0^{-1}(y)) = \frac{1}{n} \sum_{i=1}^n I(X_i \leqslant F_0^{-1}(y)) = \frac{1}{n} \sum_{i=1}^n I(F_0(X_i) \leqslant y)$$

Now using the definition of the empirical cdf S_n we can write,

$$S_n(F_0^{-1}(y)) = \frac{1}{n} \sum_{i=1}^n I(X_i \leqslant F_0^{-1}(y)) = \frac{1}{n} \sum_{i=1}^n I(F_0(X_i) \leqslant y)$$

► Therefore,

$$P(\sup_{0 < y < 1} |S_n(F_0^{-1}(y)) - y| \le d)$$

$$= P(\sup_{0 < y < 1} |\frac{1}{n} \sum_{i=1}^n I(F_0(X_i) \le y) - y| \le d)$$

▶ The distribution of $F_0(X_i)$ is uniform over [0,1] and therefore the random variables,

- ▶ The distribution of $F_0(X_i)$ is uniform over [0,1] and therefore the random variables,
- ▶ $U_i = F_0(X_i)$ for $i \le n$ are independent and have uniform distribution on [0,1]

- ▶ The distribution of $F_0(X_i)$ is uniform over [0,1] and therefore the random variables,
- ▶ $U_i = F_0(X_i)$ for $i \le n$ are independent and have uniform distribution on [0,1]
- So it is proved that

$$P(D_n \leqslant d)$$

$$= P(\sup_{0 < y < 1} |\frac{1}{n} \sum_{i=1}^n I(U_i \leqslant y) - y| \leqslant d)$$

- ▶ The distribution of $F_0(X_i)$ is uniform over [0,1] and therefore the random variables,
- ▶ $U_i = F_0(X_i)$ for $i \le n$ are independent and have uniform distribution on [0,1]
- So it is proved that

$$P(D_n \leqslant d)$$

$$= P(\sup_{0 < y < 1} |\frac{1}{n} \sum_{i=1}^{n} I(U_i \leqslant y) - y| \leqslant d)$$

ightharpoonup is clearly independent of F_0

Results

▶ In order to use Kolmogorov statistics for inference, their sampling distributions must be known.

Results

- ▶ In order to use Kolmogorov statistics for inference, their sampling distributions must be known.
- ▶ This is stated in the following two results:

For $D_n = \sup_{x} |S_n(x) - F_0(x)|$, where $F_0(.)$ is any specific continuous cdf, we have under H_0

For $D_n = \sup_{x} |S_n(x) - F_0(x)|$, where $F_0(.)$ is any specific continuous cdf, we have under H_0

$$P\{D_{n} < \frac{1}{2n} + \gamma\} = \begin{cases} 0 & \text{for } \gamma < 0 \\ \int_{\frac{1}{2n} - \gamma}^{\frac{1}{2n} + \gamma} \int_{\frac{1}{3n} - \gamma}^{\frac{1}{3n} + \gamma} ... \int_{\frac{2n-1}{2n} - \gamma}^{\frac{2n-1}{2n} + \gamma} f(u_{1}, u_{2}, ..., u_{n}) du_{n} ... du_{1} & \text{for } 0 < \gamma < \frac{2n-1}{2n} \\ 1 & \text{for } \gamma > \frac{2n-1}{2n} \end{cases}$$

- For $D_n = \sup_{x} |S_n(x) F_0(x)|$, where $F_0(.)$ is any specific continuous cdf, we have under H_0
- $P\{D_{n} < \frac{1}{2n} + \gamma\} = \begin{cases} 0 & \text{for } \gamma < 0 \\ \int_{\frac{1}{2n} \gamma}^{\frac{1}{2n} + \gamma} \int_{\frac{1}{3n} \gamma}^{\frac{1}{3n} + \gamma} ... \int_{\frac{2n-1}{2n} \gamma}^{\frac{2n-1}{2n} + \gamma} f(u_{1}, u_{2}, ..., u_{n}) du_{n} ... du_{1} & \text{for } 0 < \gamma < \frac{2n-1}{2n} \\ 1 & \text{for } \gamma > \frac{2n-1}{2n} \end{cases}$
- where

$$f(u_1, u_2, ..., u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < ... < u_n < 1 \\ 0 & o.w \end{cases}$$

Numerical values of $D_{n,\alpha}$ are given for $n \leq 40$ and selected tail probabilities α .

- Numerical values of $D_{n,\alpha}$ are given for $n \leq 40$ and selected tail probabilities α .
- For larger sample sizes, Kolmogorov derived the following convinient approx to the sample distribution of D_n

- Numerical values of $D_{n,\alpha}$ are given for $n \leq 40$ and selected tail probabilities α .
- For larger sample sizes, Kolmogorov derived the following convinient approx to the sample distribution of D_n
- ▶ If F_X is any continuous DF, then for any d > 0,

$$\lim_{n\to\infty} P\{D_n\leqslant \frac{d}{\sqrt{n}}\}=L(d)$$

- Numerical values of $D_{n,\alpha}$ are given for $n \leq 40$ and selected tail probabilities α .
- For larger sample sizes, Kolmogorov derived the following convinient approx to the sample distribution of D_n
- ▶ If F_X is any continuous DF, then for any d > 0,

$$\lim_{n\to\infty} P\{D_n\leqslant \frac{d}{\sqrt{n}}\}=L(d)$$

where,

$$L(d) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

$$\begin{array}{l} \blacktriangleright \ P_{H_0}\{D_n^+ < c\} = \\ \begin{cases} 0 & \text{for } c \leq 0 \\ \int_{1-c}^1 \int_{\frac{n-1}{n}-c}^{u_n} ... \int_{\frac{2}{n}-c}^{u_3} \int_{\frac{1}{n}-c}^{u_2} f(u_1, u_2, ... u_n) du_1 du_n & \text{for } 0 < c < 1 \\ 1 & \text{for } c \geq 1 \end{cases} \end{array}$$

$$\begin{array}{l} \blacktriangleright \ P_{H_0}\{D_n^+ < c\} = \\ \begin{cases} 0 & \text{for } c \leq 0 \\ \int_{1-c}^1 \int_{\frac{n-1}{n}-c}^{u_n} ... \int_{\frac{2}{n}-c}^{u_3} \int_{\frac{1}{n}-c}^{u_2} f(u_1,u_2,...u_n) du_1....du_n & \text{for } 0 < c < 1 \\ 1 & \text{for } c \geq 1 \\ \end{cases} \\ \blacktriangleright \ \text{where} \end{array}$$

where

$$f(u_1, u_2, ..., u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < ... < u_n < 1 \\ 0 & o.w \end{cases}$$

Proof.

Proof.

As before, we first assume wlog that F_0 is the uniform distribution on (0,1). So we denote $F_0(X_{(i)}) = U_{(i)}$. Then we can write,

Proof.

As before, we first assume wlog that F_0 is the uniform distribution on (0,1). So we denote $F_0(X_{(i)}) = U_{(i)}$. Then we can write,

$$D_{n}^{+} = \sup_{x} (S_{n}(x) - F_{0}(x))$$

$$= \max_{0 \le i \le n} [\sup_{X_{(i)} \le x \le X_{(i+1)}} (\frac{i}{n} - F_{0}(x))]$$

$$= \max_{0 \le i \le n} (\frac{i}{n} - U_{(i)})$$

$$= \max \{\max_{1 \le i \le n} (\frac{i}{n} - U_{(i)})$$

$$= \max \{\max_{1 \le i \le n} (\frac{i}{n} - U_{(i)}), 0\}$$

For all 0 < c < 1, we have,

$$\begin{split} P(D_n^+ < c) \\ &= P[\max_{1 \le i \le n} (\frac{i}{n} - U_{(i)}) < c] \\ &= P(\frac{i}{n} - U_{(i)} < c \text{ for all } i = 1, 2, ..., n) \\ &= P(U_{(i)} > \frac{i}{n} - c \text{ for all } i = 1, 2, ..., n) \\ &= \int_{1-c}^{\infty} \int_{(n-1)/n-c}^{\infty} ... \int_{2/n-c}^{\infty} \int_{1/n-c}^{\infty} f(u_1, u_2, ..., u_n) du_1 du_2 ... du_n \end{split}$$

For all 0 < c < 1, we have,

$$\begin{split} P(D_n^+ < c) \\ &= P[\max_{1 \le i \le n} (\frac{i}{n} - U_{(i)}) < c] \\ &= P(\frac{i}{n} - U_{(i)} < c \text{ for all } i = 1, 2, ..., n) \\ &= P(U_{(i)} > \frac{i}{n} - c \text{ for all } i = 1, 2, ..., n) \\ &= \int_{1-c}^{\infty} \int_{(n-1)/n-c}^{\infty} ... \int_{2/n-c}^{\infty} \int_{1/n-c}^{\infty} f(u_1, u_2, ..., u_n) du_1 du_2 ... du_n \end{split}$$

where,

$$f(u_1, u_2, ..., u_n) =$$

$$\begin{cases} n! & \text{for } 0 < u_1 < u_2 < ... < u_n < 1 \\ 0 & \text{o.w} \end{cases}$$

 \triangleright For all 0 < c < 1, we have,

$$\begin{split} P(D_n^+ < c) \\ &= P[\max_{1 \le i \le n} (\frac{i}{n} - U_{(i)}) < c] \\ &= P(\frac{i}{n} - U_{(i)} < c \text{ for all } i = 1, 2, ..., n) \\ &= P(U_{(i)} > \frac{i}{n} - c \text{ for all } i = 1, 2, ..., n) \\ &= \int_{1-c}^{\infty} \int_{(n-1)/n-c}^{\infty} ... \int_{2/n-c}^{\infty} \int_{1/n-c}^{\infty} f(u_1, u_2, ..., u_n) du_1 du_2 ... du_n \end{split}$$

where,

$$f(u_1, u_2, ..., u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < ... < u_n < 1 \\ 0 & \text{o.w} \end{cases}$$

▶ which is equivalent to the stated integral. (Proved)



 \triangleright $D_n^+\&D_n^-$ have identical distributions because of symmetry.

- $\triangleright D_n^+ \& D_n^-$ have identical distributions because of symmetry.
- ▶ For large n , (Remember $F_0(.)$ is continuous) \forall d \geq 0

$$\lim_{n\to\infty} P\{D_n^+\leqslant \frac{d}{\sqrt{n}}\} = 1 - e^{-2d^2}$$

- \triangleright $D_n^+\&D_n^-$ have identical distributions because of symmetry.
- ▶ For large n , (Remember $F_0(.)$ is continuous) \forall d \geq 0

$$\lim_{n\to\infty} P\{D_n^+\leqslant \frac{d}{\sqrt{n}}\} = 1 - e^{-2d^2}$$

▶ If F_0 is any specified continuous cdf , then for every $d \ge 0$, the limiting null distribution of $v = 4nD_n^{+2}$, as $n \to \infty$, is the $\chi^2_{(2)}$.



- $\triangleright D_n^+ \& D_n^-$ have identical distributions because of symmetry.
- ▶ For large n , (Remember $F_0(.)$ is continuous) $\forall d \ge 0$

$$\lim_{n\to\infty} P\{D_n^+\leqslant \frac{d}{\sqrt{n}}\} = 1 - e^{-2d^2}$$

▶ If F_0 is any specified continuous cdf , then for every $d \ge 0$, the limiting null distribution of $v = 4nD_n^{+2}$, as $n \to \infty$, is the $\chi^2_{(2)}$.

Proof.



- \triangleright $D_n^+ \& D_n^-$ have identical distributions because of symmetry.
- ▶ For large n , (Remember $F_0(.)$ is continuous) \forall d \geq 0

$$\lim_{n\to\infty} P\{D_n^+\leqslant \frac{d}{\sqrt{n}}\} = 1 - e^{-2d^2}$$

▶ If F_0 is any specified continuous cdf , then for every $d \ge 0$, the limiting null distribution of $v = 4nD_n^{+2}$, as $n \to \infty$, is the $\chi^2_{(2)}$.

Proof.

▶ We have $D_n^+ < d/\sqrt{n}$ if and only if $4nD_n^{+2} < 4d^2$ or $V < 4d^2$



Therefore

$$\lim_{n \to \infty} P(V < 4d^2)$$
= $\lim_{n \to \infty} P(D_n^+ < d/\sqrt{n})$
= $1 - e^{-2d^2}$
= $1 - e^{-4d^2/2}$

Therefore

$$\lim_{n \to \infty} P(V < 4d^2)$$

$$= \lim_{n \to \infty} P(D_n^+ < d/\sqrt{n})$$

$$= 1 - e^{-2d^2}$$

$$= 1 - e^{-4d^2/2}$$

ightharpoonup So, $\lim_{n \to \infty} P(V < c) = 1 - e^{-c/2}$ for all c > 0

Therefore

$$\lim_{n \to \infty} P(V < 4d^2)$$

$$= \lim_{n \to \infty} P(D_n^+ < d/\sqrt{n})$$

$$= 1 - e^{-2d^2}$$

$$= 1 - e^{-4d^2/2}$$

- ightharpoonup So, $\lim_{n \to \infty} P(V < c) = 1 e^{-c/2}$ for all c > 0
- ► The right-hand side is the cdf of a chi square distribution with 2 degrees of freedom. (Proved)

▶ The differences between $S_n(x)$ & $F_0(x)$ should be small for all x except for sampling variation , if the H_0 is true.

- ► The differences between $S_n(x)$ & $F_0(x)$ should be small for all x except for sampling variation , if the H_0 is true.
- ▶ For $H_1: F_X(x) \neq F_0(x)$ for some x, large absolute values of these deviations tend to discredit the H_0 .

- ▶ The differences between $S_n(x)$ & $F_0(x)$ should be small for all x except for sampling variation , if the H_0 is true.
- ▶ For $H_1: F_X(x) \neq F_0(x)$ for some x, large absolute values of these deviations tend to discredit the H_0 .
- ► Therefore , the K-S goodness of fit test with significance level α is to reject H_0 when $D_n > D_{n,\alpha}$.

► The following expression is considerably easier for algebraic calculations & applies when ties are present:

$$D_n = \sup_{x} |S_n(x) - F_0(x)|$$

$$= \max[|S_n(x) - F_0(x)|, |S_n(x - \epsilon) - F_0(x)|]$$

► The following expression is considerably easier for algebraic calculations & applies when ties are present:

$$D_n = \sup_{x} |S_n(x) - F_0(x)|$$

$$= \max[|S_n(x) - F_0(x)|, |S_n(x - \epsilon) - F_0(x)|]$$

 \blacktriangleright where ϵ denotes any small +ve number.

▶ suppose $H_1: F_X(x) > F_0(x) \ \forall x$

- ▶ suppose $H_1: F_X(x) > F_0(x) \ \forall x$
- the approximate rejection region is $D_n^+ > D_{n,\alpha}^+$

- ▶ suppose $H_1: F_X(x) > F_0(x) \ \forall x$
- the approximate rejection region is $D_n^+ > D_{n,\alpha}^+$
- ► Most test of goodness of fit are two-sided.

- ▶ suppose $H_1: F_X(x) > F_0(x) \ \forall x$
- the approximate rejection region is $D_n^+ > D_{n,\alpha}^+$
- Most test of goodness of fit are two-sided.
- ➤ The tail probabilities for the one sided statistic are approximately one-half of the corresponding tail probabilities for the two sided statistic.

Confidence Band

Now

$$P\{D_n > D_{n,\alpha}\} = \alpha$$

$$\iff P\{D_n < D_{n,\alpha}\} = 1 - \alpha$$

$$\iff P\{\sup_{x} |S_n(x) - F_0(x)| < D_{n,\alpha}\} = 1 - \alpha$$

$$\iff P\{S_n(x) - D_{n,\alpha} < F_0(x) < S_n(x) + D_{n,\alpha}\} = 1 - \alpha$$

Confidence Band

► Thus we define,

$$L_n(x) = \max(S_n(x) - D_{n,\alpha}, 0)$$
$$U_n(x) = \min(S_n(x) + D_{n,\alpha}, 1)$$

Confidence Band

Thus we define,

$$L_n(x) = \max(S_n(x) - D_{n,\alpha}, 0)$$

$$U_n(x) = \min(S_n(x) + D_{n,\alpha}, 1)$$

 \blacktriangleright as lower & upper confidence bands with associated confidence coefficient (1 - $\alpha)$

▶ The statistic D_n enables us to determine the minimum sample size occupied to guarantee with a certain probability $1-\alpha$, that the error in the estimate never exceeds a fixed value .

- ▶ The statistic D_n enables us to determine the minimum sample size occupied to guarantee with a certain probability $1-\alpha$, that the error in the estimate never exceeds a fixed value .
- i.e , we want to find the minimum value of n that satisfies,

$$P\{D_n < c\} = 1 - \alpha$$

$$\iff 1 - P\{D_n < c\} = P\{D_n > c\} = \alpha$$

$$\therefore c = D_{n,\alpha}$$

- ▶ The statistic D_n enables us to determine the minimum sample size occupied to guarantee with a certain probability $1-\alpha$, that the error in the estimate never exceeds a fixed value .
- i.e , we want to find the minimum value of n that satisfies,

$$P\{D_n < c\} = 1 - \alpha$$

$$\iff 1 - P\{D_n < c\} = P\{D_n > c\} = \alpha$$

$$\therefore c = D_{n,\alpha}$$

ightharpoonup' n' can be read directly from table as that sample size corresponding to $D_{n,\alpha}=c$.

- ▶ The statistic D_n enables us to determine the minimum sample size occupied to guarantee with a certain probability $1-\alpha$, that the error in the estimate never exceeds a fixed value .
- i.e , we want to find the minimum value of n that satisfies,

$$P\{D_n < c\} = 1 - \alpha$$

$$\iff 1 - P\{D_n < c\} = P\{D_n > c\} = \alpha$$

$$\therefore c = D_{n,\alpha}$$

- ightharpoonup' n' can be read directly from table as that sample size corresponding to $D_{n,\alpha}=c$.
- If no $n \le 40$ will meet the specified accuracy, the asymptotic distribution of D_n ($\lim_{n \to \infty} P\{D_n \le \frac{d}{\sqrt{n}}\} = L(d)$ where $L(d) = 1 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$ can be used by solving $c = d/\sqrt{n}$ for n, where d/\sqrt{n} is given in the last row of the table.



▶ Suppose error should be less than 0.25 with probability 0.98.

- ▶ Suppose error should be less than 0.25 with probability 0.98.
- ▶ We look down the 0.02=(1-0.98) column of table until we find the largest $c \le 0.25$.

- ▶ Suppose error should be less than 0.25 with probability 0.98.
- ▶ We look down the 0.02=(1-0.98) column of table until we find the largest $c \le 0.25$.
- ▶ This entry is 0.247 which corresponds to n = 36.

Strengths of The K-S Test

► *D*− value result will not change if *X* values are transformed to logs or reciprocals or any other transformation.

Strengths of The K-S Test

- ► D— value result will not change if X values are transformed to logs or reciprocals or any other transformation.
- ▶ Non-restriction of sample size.

Strengths of The K-S Test

- ► D— value result will not change if X values are transformed to logs or reciprocals or any other transformation.
- ▶ Non-restriction of sample size.
- ► The D− value is an easy to compute and the graph can be understood easily.

Drawbacks

► The K-S test if less sensitive when the differences between the curves is greatest at the beginning or the end of the distributions

Drawbacks

- ► The K-S test if less sensitive when the differences between the curves is greatest at the beginning or the end of the distributions
- ► It works best only when the CDF's deviate the most near the center of the distribution.

Drawbacks

- ► The K-S test if less sensitive when the differences between the curves is greatest at the beginning or the end of the distributions
- ► It works best only when the CDF's deviate the most near the center of the distribution.
- The situation in which normality tests are needed –small sample sizes—is also a situation when they perform poorly.

Steps in K-S test:

1. sort the data from smallest to largest.

- 1. sort the data from smallest to largest.
- 2. Compute the empirical distribution function.

- 1. sort the data from smallest to largest.
- 2. Compute the empirical distribution function.
- 3. Find the maximum absolute difference (D-value).

- 1. sort the data from smallest to largest.
- 2. Compute the empirical distribution function.
- 3. Find the maximum absolute difference (D-value).
- 4. If D is greater than critical D, then it can be concluded that the distribution are indeed different, otherwise there is not enough evidence to prove the difference between the two data-set.

- 1. sort the data from smallest to largest.
- 2. Compute the empirical distribution function.
- 3. Find the maximum absolute difference (D-value).
- 4. If D is greater than critical D, then it can be concluded that the distribution are indeed different, otherwise there is not enough evidence to prove the difference between the two data-set.
- 5. A P-value can also be calculated from the D value and the sample size of the two data sets.

Here is data for 100 observations.

```
-0.16
       -0.68
               -0.32
                       -0.85
                                0.89
                                       -2.28
                                                0.63
                                                       0.41
                                                               0.15
                                                                       0.74
1.30
       -0.13
                0.80
                       -0.75
                                0.28
                                       -1.00
                                                0.14
                                                      -1.38
                                                              -0.04
                                                                      -0.25
-0.17
        1.29
                0.47
                       -1.23
                                0.21
                                       -0.04
                                               0.07
                                                      -0.08
                                                               0.32
                                                                      -0.17
0.13
       -1.94
                0.78
                        0.19
                               -0.12
                                       -0.19
                                               0.76
                                                      -1.48
                                                              -0.01
                                                                       0.20
                                0.80
                                                1.32
-1.97
       -0.37
                3.08
                       -0.40
                                       0.01
                                                      -0.47
                                                               2.29
                                                                      -0.26
-1.52
       -0.06
               -1.02
                        1.06
                                0.60
                                       1.15
                                               1.92
                                                      -0.06
                                                              -0.19
                                                                       0.67
0.29
        0.58
                0.02
                        2.18
                               -0.04
                                       -0.13
                                               -0.79
                                                      -1.28
                                                              -1.41
                                                                      -0.23
0.65
       -0.26
               -0.17
                       -1.53
                               -1.69
                                       -1.60
                                               0.09
                                                      -1.11
                                                               0.30
                                                                       0.71
-0.88
       -0.03
                0.56
                       -3.68
                                2.40
                                       0.62
                                                0.52
                                                      -1.25
                                                               0.85
                                                                      -0.09
-0.23
       -1.16
                0.22
                       -1.68
                                0.50
                                       -0.35
                                               -0.35
                                                      -0.33
                                                              -0.24
                                                                       0.25
```

Do they come from N(0,1)?

▶ 1st we order the data.

```
-3.68
        -2.28
                -1.97
                        -1.94
                                -1.69
                                        -1.68
                                                -1.60
                                                         -1.53
                                                                 -1.52
                                                                         -1.48
-1.41
        -1.38
                -1.28
                        -1.25
                                -1.23
                                        -1.16
                                                -1.11
                                                         -1.02
                                                                 -1.00
                                                                         -0.88
-0.85
        -0.79
                -0.75
                        -0.68
                                -0.47
                                        -0.40
                                                -0.37
                                                         -0.35
                                                                 -0.35
                                                                         -0.33
-0.32
        -0.26
                -0.26
                        -0.25
                                -0.24
                                         -0.23
                                                 -0.23
                                                         -0.19
                                                                 -0.19
                                                                         -0.17
-0.17
        -0.17
                -0.16
                        -0.13
                                -0.13
                                         -0.12
                                                 -0.09
                                                         -0.08
                                                                 -0.06
                                                                         -0.06
-0.04
        -0.04
                -0.04
                        -0.03
                                -0.01
                                         0.01
                                                 0.02
                                                          0.07
                                                                  0.09
                                                                          0.13
0.14
        0.15
                 0.19
                         0.20
                                 0.21
                                         0.22
                                                 0.25
                                                          0.28
                                                                  0.29
                                                                          0.30
0.32
        0.41
                 0.47
                         0.50
                                 0.52
                                         0.56
                                                 0.58
                                                         0.60
                                                                  0.62
                                                                          0.63
0.65
        0.67
                 0.71
                         0.74
                                 0.76
                                         0.78
                                                 0.80
                                                         0.80
                                                                  0.85
                                                                          0.89
 1.06
         1.15
                 1.29
                         1.30
                                 1.32
                                         1.92
                                                 2.18
                                                          2.29
                                                                  2.40
                                                                          3.08
```

▶ 1st we order the data.

| -3.68 | -2.28 | -1.97 | -1.94 | -1.69 | -1.68 | -1.60 | -1.53 | -1.52 | -1.48 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| -1.41 | -1.38 | -1.28 | -1.25 | -1.23 | -1.16 | -1.11 | -1.02 | -1.00 | -0.88 |
| -0.85 | -0.79 | -0.75 | -0.68 | -0.47 | -0.40 | -0.37 | -0.35 | -0.35 | -0.33 |
| -0.32 | -0.26 | -0.26 | -0.25 | -0.24 | -0.23 | -0.23 | -0.19 | -0.19 | -0.17 |
| -0.17 | -0.17 | -0.16 | -0.13 | -0.13 | -0.12 | -0.09 | -0.08 | -0.06 | -0.06 |
| -0.04 | -0.04 | -0.04 | -0.03 | -0.01 | 0.01 | 0.02 | 0.07 | 0.09 | 0.13 |
| 0.14 | 0.15 | 0.19 | 0.20 | 0.21 | 0.22 | 0.25 | 0.28 | 0.29 | 0.30 |
| 0.32 | 0.41 | 0.47 | 0.50 | 0.52 | 0.56 | 0.58 | 0.60 | 0.62 | 0.63 |
| 0.65 | 0.67 | 0.71 | 0.74 | 0.76 | 0.78 | 0.80 | 0.80 | 0.85 | 0.89 |
| 1.06 | 1.15 | 1.29 | 1.30 | 1.32 | 1.92 | 2.18 | 2.29 | 2.40 | 3.08 |

▶ then we compute the empirical distribution.

Here

$$S_{100}(-3.68) = \frac{1}{100}$$

$$S_{100}(-2.28) = \frac{2}{100}$$

$$.$$

$$.$$

$$S_{100}(3.08) = 1$$

Here

$$S_{100}(-3.68) = \frac{1}{100}$$

$$S_{100}(-2.28) = \frac{2}{100}$$

$$\vdots$$

$$S_{100}(3.08) = 1$$

▶ If our data is ordered, x_1 being the least and x_n being the largest,

$$S_n(x_i) = \frac{i}{100}$$

| 0.000 | 0.011 | 0.024 | 0.026 | 0.045 | 0.047 | 0.055 | 0.064 | 0.064 | 0.070 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.080 | 0.084 | 0.101 | 0.107 | 0.110 | 0.123 | 0.133 | 0.154 | 0.158 | 0.189 |
| 0.198 | 0.215 | 0.226 | 0.249 | 0.321 | 0.343 | 0.356 | 0.362 | 0.363 | 0.369 |
| 0.375 | 0.396 | 0.399 | 0.400 | 0.407 | 0.409 | 0.410 | 0.423 | 0.425 | 0.432 |
| 0.432 | 0.434 | 0.437 | 0.447 | 0.449 | 0.453 | 0.464 | 0.468 | 0.476 | 0.477 |
| 0.484 | 0.484 | 0.485 | 0.490 | 0.496 | 0.505 | 0.508 | 0.526 | 0.535 | 0.553 |
| 0.557 | 0.560 | 0.577 | 0.577 | 0.582 | 0.588 | 0.597 | 0.610 | 0.614 | 0.617 |
| 0.627 | 0.658 | 0.680 | 0.692 | 0.698 | 0.711 | 0.720 | 0.727 | 0.732 | 0.735 |
| 0.743 | 0.748 | 0.761 | 0.771 | 0.777 | 0.783 | 0.788 | 0.789 | 0.803 | 0.812 |
| 0.854 | 0.874 | 0.902 | 0.903 | 0.907 | 0.973 | 0.985 | 0.989 | 0.992 | 0.999 |

▶ for each observation x_i compute $F_{exp}(x_i) = P(Z \le x_i)$

```
0.000
        0.011
                0.024
                        0.026
                                0.045
                                         0.047
                                                 0.055
                                                         0.064
                                                                  0.064
                                                                          0.070
0.080
        0.084
                0.101
                        0.107
                                 0.110
                                         0.123
                                                 0.133
                                                         0.154
                                                                  0.158
                                                                          0.189
0.198
        0.215
                0.226
                        0.249
                                0.321
                                         0.343
                                                 0.356
                                                         0.362
                                                                  0.363
                                                                          0.369
0.375
        0.396
                0.399
                                                         0.423
                                                                  0.425
                        0.400
                                 0.407
                                         0.409
                                                 0.410
                                                                          0.432
0.432
        0.434
                0.437
                        0.447
                                 0.449
                                         0.453
                                                 0.464
                                                         0.468
                                                                  0.476
                                                                          0.477
0.484
        0.484
                0.485
                        0.490
                                0.496
                                         0.505
                                                 0.508
                                                         0.526
                                                                  0.535
                                                                          0.553
0.557
        0.560
                0.577
                        0.577
                                0.582
                                         0.588
                                                 0.597
                                                                          0.617
                                                         0.610
                                                                  0.614
0.627
        0.658
                0.680
                        0.692
                                 0.698
                                         0.711
                                                 0.720
                                                         0.727
                                                                  0.732
                                                                          0.735
0.743
        0.748
                0.761
                        0.771
                                 0.777
                                         0.783
                                                 0.788
                                                         0.789
                                                                  0.803
                                                                          0.812
0.854
        0.874
                0.902
                        0.903
                                 0.907
                                         0.973
                                                 0.985
                                                         0.989
                                                                  0.992
                                                                          0.999
```

- ▶ for each observation x_i compute $F_{exp}(x_i) = P(Z \le x_i)$
- In this case the expected distribution function is standard normal so use the normal table.

► Then compute the absolute difference between the entries in the two tables.

| 0.010 | 0.009 | 0.006 | 0.014 | 0.004 | 0.014 | 0.015 | 0.017 | 0.026 | 0.031 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.031 | 0.036 | 0.030 | 0.034 | 0.041 | 0.037 | 0.037 | 0.026 | 0.031 | 0.011 |
| 0.012 | 0.005 | 0.003 | 0.008 | 0.069 | 0.085 | 0.086 | 0.083 | 0.073 | 0.071 |
| 0.064 | 0.077 | 0.067 | 0.061 | 0.055 | 0.049 | 0.039 | 0.045 | 0.035 | 0.033 |
| 0.023 | 0.013 | 0.006 | 0.008 | 0.002 | 0.008 | 0.006 | 0.012 | 0.014 | 0.024 |
| 0.026 | 0.036 | 0.046 | 0.052 | 0.054 | 0.056 | 0.062 | 0.052 | 0.054 | 0.048 |
| 0.054 | 0.060 | 0.055 | 0.061 | 0.067 | 0.073 | 0.071 | 0.070 | 0.076 | 0.082 |
| 0.084 | 0.061 | 0.049 | 0.049 | 0.052 | 0.048 | 0.051 | 0.054 | 0.058 | 0.064 |
| 0.068 | 0.071 | 0.069 | 0.070 | 0.074 | 0.078 | 0.082 | 0.092 | 0.088 | 0.087 |
| 0.055 | 0.045 | 0.029 | 0.037 | 0.043 | 0.013 | 0.015 | 0.009 | 0.002 | 0.001 |

► Then compute the absolute difference between the entries in the two tables.

| 0.010 | 0.009 | 0.006 | 0.014 | 0.004 | 0.014 | 0.015 | 0.017 | 0.026 | 0.031 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.031 | 0.036 | 0.030 | 0.034 | 0.041 | 0.037 | 0.037 | 0.026 | 0.031 | 0.011 |
| 0.012 | 0.005 | 0.003 | 0.008 | 0.069 | 0.085 | 0.086 | 0.083 | 0.073 | 0.071 |
| 0.064 | 0.077 | 0.067 | 0.061 | 0.055 | 0.049 | 0.039 | 0.045 | 0.035 | 0.033 |
| 0.023 | 0.013 | 0.006 | 0.008 | 0.002 | 0.008 | 0.006 | 0.012 | 0.014 | 0.024 |
| 0.026 | 0.036 | 0.046 | 0.052 | 0.054 | 0.056 | 0.062 | 0.052 | 0.054 | 0.048 |
| 0.054 | 0.060 | 0.055 | 0.061 | 0.067 | 0.073 | 0.071 | 0.070 | 0.076 | 0.082 |
| 0.084 | 0.061 | 0.049 | 0.049 | 0.052 | 0.048 | 0.051 | 0.054 | 0.058 | 0.064 |
| 0.068 | 0.071 | 0.069 | 0.070 | 0.074 | 0.078 | 0.082 | 0.092 | 0.088 | 0.087 |
| 0.055 | 0.045 | 0.029 | 0.037 | 0.043 | 0.013 | 0.015 | 0.009 | 0.002 | 0.001 |

► The Kolmogorov-Smirnov statistic $D_n = 0.092$ is the maximum shown here in blue.

▶ At the 95% level the critical value is approximately given by,

$$D_{crit,0.05} = \frac{1.36}{\sqrt{n}}$$

▶ At the 95% level the critical value is approximately given by,

$$D_{crit,0.05} = \frac{1.36}{\sqrt{n}}$$

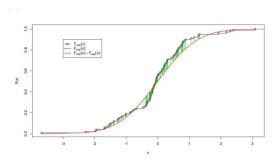
▶ Here we have a sample size of n = 100 so $D_{crit} = 0.136$.

At the 95% level the critical value is approximately given by,

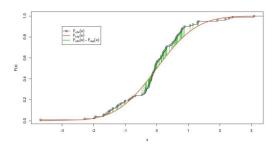
$$D_{crit,0.05} = \frac{1.36}{\sqrt{n}}$$

- ▶ Here we have a sample size of n = 100 so $D_{crit} = 0.136$.
- ► Since 0.092 < 0.136 do not reject the null hypothesis.

We have calculated the maximum absolute distance between the expected and observed distribution functions, in green in the plot below.



- We have calculated the maximum absolute distance between the expected and observed distribution functions, in green in the plot below.
- $\blacktriangleright \text{ Here } F_{obs}(x) = S_n(x)$



References

 Nonparametric Statistical Inference (Fourth Edition), by Jean Dickinson Gibbons and Subhabrata Chakraborti



Acknowledgement

We would like to express our special thanks of gratitude to our respected Professor N V KRISHNA CHAITANYA YERROJU who gave us the golden opportunity to do this wonderful presentation on the topic Kolmogrov-Smirnov One Sample Test, which also helped us in doing a lot of Research and we came to know about so many new things. We are really thankful to him.