Empirical CDF and the Kolmogorov—Smirnov statistic

- If we want to determine whether data are generated from a specific distribution, we can use the empirical CDF (computed based on the data).
- Recall that a distribution can be characterized by its CDF (cumulative distribution function; 累積分布函數).
  - The CDF of a distribution $\mathcal{D}$ is the function F defined by

    $$F(x) = P(\mathcal{D} \le x) \text{ for } x \in (-\infty, \infty).$$

    For instance, the CDF of $N(0,1)$ is `pnorm`.
- If data are IID (independent and identically distributed; 獨立同分布) from the distribution with CDF F, then F can be estimated by the empirical CDF based on the data.
- The empirical CDF based on data $X_1, \ldots, X_n$ is the function F_n defined by

  $$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i \le x)}{n},$$

  where for $1 \le i \le n$,

  $$I(X_i \le x) = \begin{cases} 1 & \text{if } X_i \le x; \\ 0 & \text{otherwise.} \end{cases}$$

- Suppose that we have IID data (獨立同分布資料) $X_1, \ldots, X_n$ and each X_i has CDF F. Let F_n be the empirical CDF based on the data. Then it can be shown that $E(F_n(x)) = F(x)$ and $Var(F_n(x)) = F(x)(1-F(x))/n$, so the empirical CDF $F_n \approx F$ for large n.
- The R command for computing the empirical CDF is `ecdf`.

  Example 1. Define a function `ecdf1` to compute the empirical CDF for input data. Generate a random sample of size 30 from $N(0,1)$. Make a plot of the empirical CDF computed using `ecdf1` and add the plot of the empirical CDF computed using `ecdf`. Also add the plot of the CDF for $N(0,1)$.

  Sol.

```
ecdf1 <- function(data){
 n <- length(data)
 f <- function(x){
  m <- length(x)
  ans <- rep(0,m)
  for (i in 1:m){ ans[i] <- length(data[data<=x[i]])/n }
```

```
  return(ans)
  }
 return(f)
}

set.seed(1)
data <- rnorm(30); m <- min(data); M <- max(data)
f1 <- ecdf1(data)
f2 <- ecdf(data)
curve(f1, m, M)
curve(f2, m, M, add=T, col=2)
curve(pnorm, m, M, add=T, col=3)
```

- To measure the distance between the empirical CDF F_n and a specific CDF F_0, one can use the Kolmogorov—Smirnov statistic D, where

$$D = \sup_x |F_n(x) - F_0(x)|.$$

  To compute D, one needs to use the fact that F_n is a step function with jumps at data points.

- Suppose that F_n is the empirical CDF based on data X_1, ..., X_n. Let $X_{(1)}$, ..., $X_{(n)}$ be the sorted data in ascending order. Then

$$F_n(x) = \begin{cases} 0 & \text{if } x < X_{(1)}; \\ i/n & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for some } i \in \{1, \ldots, n-1\}; \\ 1 & \text{if } x \ge X_{(n)}. \end{cases}$$

  For the data in Example 1, we can plot the empirical CDF and mark the jump points of the empirical CDF:

```
n <- 30
set.seed(1)
data <- rnorm(n); m <- min(data); M <- max(data)
f2 <- ecdf(data)
curve(f2,  m, M)
data.s <- sort(data)
points(data.s, (1:n)/n, col=2)
```

- The Kolmogorov-Smirnov statistic is computed using the following formula

$$\begin{aligned} &\sup_x |F_n(x) - F_0(x)| \\ &= \max_{1 \le i \le n} \max \left\{ \left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{n} - F_0(X_{(i)}) \right| \right\}. \end{aligned} \tag{1}$$

- Example 2. Write a function `ks.stat` that computes the Kolmogorov-Smirnov statistic based on input data and a given CDF `F0` using (1). Generate a sample of size 30 from $N(0,1)$. Compute the Kolmogorov-Smirnov statistic that measures the distance between the the empirical CDF based on the data and each of the following CDF:

  (a) the CDF of $N(0,1)$

  (b) the CDF of $N(0,4)$

  (c) the CDF of $N(30,1)$

  Base on the Kolmogorov-Smirnov statistics, which distribution is most suitable for the data, $N(0,1)$, $N(0,4)$, or $N(30,1)$?

  Sol.

```
ks.stat <- function(data, F0){
 n <- length(data)
 x <- sort(data)
 a <- F0(x) - (0:(n-1))/n
 b <- F0(x) - (1:n)/n
 Dn <- max( abs(c(a,b)) )
 return(Dn)
}

set.seed(1)
x <- rnorm(30)

pnorm.b <- function(x){ pnorm(x, mean=0, sd=2) }
pnorm.c <- function(x){ pnorm(x, mean=30, sd=1) }

ks.stat(x, pnorm) #statistic for N(0,1)
ks.stat(x, pnorm.b) #statistic for N(0,4)
ks.stat(x, pnorm.c) #statistic for N(30,1)
#N(0,1) is the most suitable distribution for the data
#since the corresponding Kolmogorov-Smirnov statistic is the smallest.
```

- The Kolmogorov-Smirnov statistic can be computed using the R function `ks.test`. The usage is

```
ks.test(data, F0)$statistic
```

  where `data` is the data vector and `F0` is a given CDF.

```
set.seed(1)
data <- rnorm(30)
ks.stat(data, pnorm)
ks.test(data, pnorm)$statistic
```

- The Kolmogorov-Smirnov statistic is the test statistic for the Kolmogorov-Smirnov test. The null hypothesis H_0 is that the data are from a distribution with CDF F_0. The test rejects H_0 if the Kolmogorov-Smirnov statistic is large. The p-value for the Kolmogorov-Smirnov test can be obtained using the R function `ks.test`. The usage is

  ```
  ks.test(data, F0)$p.value
  ```

  where `data` is the data vector and `F0` is a given CDF. We can reject H_0 at level α if the p-value is less than α.

  Example 3. Running the following R commands to generate a random sample of size 1000 from $N(0,1)$:

  ```
  set.seed(1)
  data <- rnorm(1000)
  ```

  Can we conclude that the data are not from $N(0.6,1)$ at level 0.05?

  Sol.

  ```
  set.seed(1)
  data <- rnorm(1000)
  pnorm_test <- function(x){ pnorm(x, mean=0.6) }
  ks.test(data, pnorm_test)$p.value
  #p-value < 0.05 => reject the null hypothesis that
  #data are from N(0.6,1) at level 0.05
  ```

- Practice problem. Generate 20 observations in R from $U(0,1)$ (the uniform distribution on $[0,1]$) and store the data in a variable `data` by running the following R commands.

  ```
  set.seed(1)
  data <- runif(20, min =0, max=1)
  ```

  Compute the Kolmogorov-Smirnov statistic that measures the distance between the the empirical CDF based on the data and each of the following CDFs:

  (a) the CDF of $U(0,1)$

  (b) the CDF of $N(0,1)$

  (c) the CDF of $N(0.5,1/12)$

  Base on the Kolmogorov-Smirnov statistics, which distribution is most suitable for the data, $U(0,1)$, $N(0,1)$, or $N(0.5,1/12)$?