Learning from a random sample (IID data)

- Suppose that $(X_1, \ldots, X_n)$ is a random sample and the distribution of X_i is $\mathcal{D}$, then we say that $(X_1, \ldots, X_n)$ is a random sample from $\mathcal{D}$ (or a random sample with population distribution $\mathcal{D}$).
- Suppose that $(X_1, \ldots, X_n)$ is a random sample from $\mathcal{D}$. Then we can learn from the sample about
  - quantities determined by $\mathcal{D}$ such as the mean and variance of $\mathcal{D}$, or
  - the PDF of $\mathcal{D}$ (if $\mathcal{D}$ has a PDF).
- Learning from sample mean and sample standard deviation. Suppose that X_1, ..., X_n are IID. Suppose that $E(X_1) = \mu$ and $\sqrt{Var(X_1)} = \sigma$ are both finite. Let
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$
and
$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}.$$
Then $\bar{X} \approx \mu$ and $S \approx \sigma$ with large probability when n is large enough.
- Generate IID data using R from $N(0,4)$, $U(-1,2)$, and the exponential distribution with mean 1/4:

```
data <- rnorm(100, mean=0, sd=2) #generate 100 IID data from N(0,4)
data <- runif(100, -1, 2)        #generate 100 IID data from U(-1,2)
data <- rexp(100, rate=4)
#generate 100 IID data from the exponential distribution with mean 1/4
```

- Example 1. Generate 10000 IID data from $U(0,1)$ using R. Compute the sample mean and sample standard deviation, and compare them with the mean and standard deviation of $U(0,1)$.

  Sol. The R commands are given below.

```
#generate 10000 data from U(0,1)
data <- runif(10000, 0, 1)

#compute the sample mean
mean(data)

#compute the sample standard deviation
sd(data)
```

  The mean for $U(0,1)$ is 0.5, which is close to the sample mean. The standard deviation for U(0,1) is $\sqrt{1/12} \approx 0.2886751$, which is close to the sample standard deviation.

- Example 2. Generate 10000 IID data from $N(0, 9)$ using R. Compute the sample mean and sample standard deviation, and compare them with the mean and standard deviation of $N(0, 9)$.

  Sol. The R commands are given below.

```
#generate 10000 data from N(0,9)
data <- rnorm(10000, mean=0, sd=3)

#compute the sample mean
mean(data)

#compute the sample standard deviation
sd(data)
```

  The mean and standardard deviation for $N(0, 9)$ are 0 and 3 respectively, which are close to the sample mean and the sample standard deviation respectively.

- A normalized histogram is a historgram whose bin heights are divided by a postive constant so that the areas of the bins sum up to one.

- Learning from a normalized histogram. Suppose that $X_1, \ldots, X_n$ are IID and X_1 has PDF f. Then the normalized histogram based on $X_1, \ldots, X_n$ can approximate f well for large n.

- The R command for drawing a normalized histogram with break points selected using Scott's rule is `hist(x, nclass="scott", freq=FALSE)`, where `x` is the data vector.

- Example 3. Generate 10000 IID data from the exponential distribution with mean 1/4 using R. Plot the normalized histogram and add the graph of the PDF of the exponential distribution for comparison.

  Sol. The R commands are given below.

```
#generate data
data <- rexp(10000, rate=4)

#draw the normalized histogram
hist(data, nclass="scott", freq=FALSE)

#add the graph of the exponential PDF
f <- function(x){ 4*exp(-4*x) }
curve(f,add=TRUE, col="red")
```

- Example 4. Suppose that $n = 1000$ and $X_1, \ldots, X_n$ are IID $U(0, 1)$. Let $\mu = E(X_1) = 0.5$ and $\sigma = \sqrt{Var(X_1)} = \sqrt{1/12}$. Generate 10000

IID data using R from the distribution of

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$$

and plot a normalized histogram of the data. Add the $N(0,1)$ PDF to the plot.

Sol. The R commands are given below.

```
n <- 1000
mu <- 0.5
sigma <- sqrt(1/12)
data <- rep(0, 10000)

#generate 10000 IID data
for (i in 1:10000){
 x <- runif(n,0,1)
 data[i] <- sqrt(n)*(mean(x)-mu)/sigma
}

#plot the histogram
hist(data, nclass="scott", freq=FALSE)

#add the N(0,1) PDF
f <- function(x){ exp(-x^2/2)/sqrt(2*pi) }
curve(f, add=TRUE, col="red")
```

Note that the above result supports the central limit theorem (中央極限定理)

- Central Limit Theorem. Suppose that $X_1, \ldots, X_n$ are IID with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\bar{X} = (X_1 + \cdots + X_n)/n$. Then the distribution of

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$$

is approximately $N(0,1)$.