Estimation based on IID data

- Recall: IID = Independent and Identically Distributed (獨立同分布). Random vectors $X_1, \ldots, X_n$ are IID means they are independent and have the same distribution.

- When we have IID data $X_1, \ldots, X_n$, $(X_1, \ldots, X_n)$ is called a random sample and n is called the sample size of the random sample. We would like to learn about the distribution of X_1 (or some quantities related to the distribution such as mean or variance for the univariate case) based on the sample. Below we will focus on the univariate case where each X_i is a random variable.

- Strong Law of Large Numbers (SLLN, 強大數法則). Suppose that $X_1, \ldots, X_n, \ldots$ are IID random variables and $E(X_1)$ is finite. Let $\bar{X} = \sum_{i=1}^n X_i/n$, then
$$P\left(\lim_{n\to\infty} \bar{X} = E(X_1)\right) = 1. \tag{1}$$
Note.
  - (1) implies the following result:
$$\lim_{n\to\infty} P\left(\left|\bar{X} - E(X_1)\right| > \varepsilon\right) = 0 \text{ for all } \varepsilon > 0. \tag{2}$$

- A version of Weak Law of Large Numbers (WLLN, 弱大數法則) is given in Theorem 5.1.1 in the text.

  Fact 1 *(Theorem 5.1.1 in the text) Suppose that $X_1, \ldots, X_n, \ldots$ are IID random variables and $E(X_1)$ and $Var(X_1)$ are finite. Let $\bar{X} = \sum_{i=1}^n X_i/n$, then (2) holds.*

  Note.
  - The proof of Fact 1 can be based on Chebyshev's inequality or Markov's inequality (Homework Problem 65 last semester).
  - Since (1) implies (2), the assumption that $Var(X_1)$ is finite is not needed in Fact 1.
  - When (2) holds, we say that $\bar{X}$ converges to $E(X_1)$ in probability, denoted by $\bar{X} \xrightarrow{P} E(X_1)$.

- Convergence in probability. Suppose that $Y, Y_1, Y_2, \ldots$ are random vectors. If
$$\lim_{n\to\infty} P\left(\|Y_n - Y\| > \varepsilon\right) = 0 \text{ for all } \varepsilon > 0,$$
then we say that Y_n converges to Y in probability (as $n \to \infty$), denoted by $Y_n \xrightarrow{P} Y$. Here $\|\cdot\|$ denotes the Euclidean norm.

- Suppose that we use a statistic T_n to estimate some quantity θ, where n is the sample size. If $T_n \xrightarrow{P} \theta$ as $n \to \infty$, then T_n is called a consistent estimator of θ (T_n爲θ的一致估計量).

- Note that an estimator must be a statistic, which can be computed given the data.

- Example 1. Suppose that we have IID data $X_1, \ldots, X_n$ and $\mu = E(X_1)$ is finite. Which of the following statements are true?

  (a) $\sum_{i=1}^n X_i/n$ is a consistent estimator of μ.
  (b) $1 + \sum_{i=1}^n X_i/n$ is a consistent estimator of $1 + \mu$.
  (c) $\mu + \sum_{i=1}^n X_i/n$ is a consistent estimator of 2μ.

  Ans. (a)(b)

- Example 2. Suppose that we have IID data $X_1, \ldots, X_n$ and the possible values of X_1 are $a_1, a_2, \ldots, a_m$. Let n_1 be the number of X_is that are equal to a_1. Then n_1/n is a consistent estimator of $P(X_1 = a_1)$.

- Example 3. Suppose that we have IID data $X_1, \ldots, X_n$, X_1 takes values in an interval $(-\infty, \infty)$, and X_1 has a PDF f. Suppose that $|f'|$ is bounded above by a constant M on $(-\infty, \infty)$. Suppose that $\{h_n\}_{n=1}^\infty$ is a sequence of positive numbers such that

$$\lim_{n\to\infty} h_n = 0 \text{ and } \lim_{n\to\infty} nh_n = \infty.$$

  Divide $(-\infty, \infty)$ into disjoint segments of the form $(c, c + h_n]$ using break points in $\{kh_n\colon\ k \in \{0, \pm 1, \pm 2, \ldots\}\}$. For a point $x_0 \in (-\infty, \infty)$, let $(c, c + h_n]$ be the segment such that $x_0 \in (c, c + h_n]$. Let

$$Y_i = \begin{cases} 1 & \text{if } X_i \in (c, c + h_n]; \\ 0 & \text{if } X_i \notin (c, c + h_n]. \end{cases}$$

  Then $\sum_{i=1}^n Y_i/(nh_n)$ is a consistent estimator of $f(x_0)$.
  A sketch of proof. We will first show that

$$\lim_{n\to\infty} E\left(\frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0)\right)^2 = 0 \tag{3}$$

  and then apply Markov's inequality to show that $\sum_{i=1}^n Y_i/(nh_n)$ is a consistent estimator of $f(x_0)$. Note that

$$\begin{aligned} & E\left(\frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0)\right)^2 \\ &= Var\left(\frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0)\right) + \left(E\left(\frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0)\right)\right)^2 \\ &\le \frac{p_n}{nh_n^2} + \left(\frac{p_n}{h_n} - f(x_0)\right)^2, \end{aligned}$$

where $p_n = P(X_1 \in (c, c+h_n])$. By the mean value theorem for integration, there exists $\xi \in (c, c+h_n)$ such that

$$\left| \frac{p_n}{h_n} - f(x_0) \right| = |f(\xi) - f(x_0)| \le M h_n \to 0$$

as $n \to \infty$, so (3) holds. For $\varepsilon > 0$,

$$\begin{aligned} P\left(\left| \frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0) \right| > \varepsilon \right) &= P\left(\left(\frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0) \right)^2 > \varepsilon^2 \right) \\ &\le \frac{E\left(\frac{\sum_{i=1}^n Y_i}{nh_n} - f(x_0) \right)^2}{\varepsilon^2} \qquad \text{(Markov inequality)} \\ &\overset{(3)}{\to} 0 \text{ as } n \to \infty, \end{aligned}$$

so $\sum_{i=1}^n Y_i/(nh_n)$ is a consistent estimator of $f(x_0)$.

- In general, for T_n: an estimator of some quantity θ, if $\lim_{n\to\infty} E\|T_n - \theta\|^k = 0$ for some $k > 0$, then T_n is a consistent estimator of θ. The proof is left as an exercise.

- Generating IID data using R

  - The R command for generating n IID data from some distriubtion is `rxxxx`, where `xxxx` is the abbreviation of the distribution name in R. To find out the abbreviation, first run

    ```
    help.search("Distribution$", package="stats", fields="title")
    ```

    to obtain a list of distributions, and then run `help(...)`, where ... is the full name of a distribution. Then the abbreviation can be found. For example, run

    ```
    help("Normal")
    ```

    Then, we can find that the R command for generating n IID data from $N(m, s^2)$ is

    ```
    rnorm(n, mean=m, sd=s)
    ```

    and the (continuous) density for $N(m, s^2)$ evaluated at `x` is

    ```
    dnorm(x, mean=m, sd=s)
    ```

- Estimating the density using a normalized histogram.

  - For a normalized histogram, the height for each segment is count$/(nh_n)$, where n is the sample size, h_n is the segment length and count is the number of observations in the segment.

– The R command for drawing a histogram based on data vector `x` with break point vector `bks` is `hist(x, breaks=bks)`. One can use `hist(x, breaks="scott")` to determine break points based on Scott's suggestion. Running

```
hist(x, breaks="scott")
```

gives the same result as running the following R commands

```
nclass.scott <- function(x){
 n <- length(x)
 h <- 3.5*sd(x)*n^(-1/3)
 #sd(x) is the sample standard deviation of the sample x
 d <- max(x)-min(x)
 n.class <- ceiling(d/h)
 return(n.class)
}
hist(x, nclass=nclass.scott)
```

– To draw a normalized histogram using the R command `hist`, add the option `freq=FALSE`.

- Example 4. Generate $n = 1000$ IID data from the $N(1, (1.1)^2)$ distribution in R and plot the normalized histogram with break points determined based on Scott's suggestion.

  Sol. The R commands for generating the required data and drawing the histogram are given below:

```
n <- 1000
x <- rnorm(n, mean=1, sd=1.1)
hist(x, breaks="scott", freq=FALSE)
```

  We can also add the continuous density of $N(1, (1.1)^2)$ for comparison:

```
#define the density
f <- function(x){
 mu <- 1
 sigma <- 1.1
 z <- (x-mu)/sigma
 ans <- exp(-z^2/2)/(sqrt(2*pi)*sigma)
 return(ans)
}
#or
#f <- function(x){ dnorm(x, mean=1, sd=1.1) }
curve(f, add=TRUE, col=2)
```

- Suppose we have IID data $X_1, \ldots, X_n$ and the distribution of X_1 belongs to some known familty (such as the family of univariate normal distributions), then it is often possible to express the PDF or PMF of X_1 as

a function depending on some parameter vector θ. For example, if the distribution of X_1 is $N(\mu, \sigma^2)$, then $\theta = (\mu, \sigma)$. In such case, we would like to estimate θ based on the data to learn about the distribution of X_1. This is known as a parametric estimation problem.