CH 11: Nonparametric Methods

- Estimation of CDF

Let $X_1, X_2, \cdots, X_n$ iid $\sim F$, we need to estimate F.

Empirical distri.:

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

- Step function

(1). $EF_n(x) = F(x)$,

$$Var(F_n(x)) = F(x)(1 - F(x))/n$$

(2). If $F(\cdot)$ is continuous, then

$$\sup_{x \in R^1} |F_n(x) - F(x)| \longrightarrow 0 \; a.s.$$

(3). If $0 < F(\cdot) < 1$, then

$$\frac{\sqrt{n}[F_n(x) - F(x)]}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d.} N(0, 1)$$

(4). If $F(\cdot)$ is continuous, then

$$\sqrt{n} \sup_{x \in R^1} |F_n(x) - F(x)| \xrightarrow{d.} X_0,$$

where $X_0 \sim Q$ with

$$Q(x) = \sum_{j=-\infty}^{+\infty} (-1)^j \exp\{-2j^2x^2\}$$

for $x > 0$. It can be used to test

$$H_0 : F = F_0, \qquad H_1 : F \neq F_0$$

首都师范大学 数学科学学院
统计学系
Department of Statistics

• Quantile estimation

the pth quantile of F $(0 < p < 1)$:

$$x_p = \inf\{x : F(x) \geq p\}$$

We use the sample order statistic $X_{([np]+1)}$ to estimate x_p.

• If x_p is unique for $F(x) = p$, then $X_{([np]+1)}$ is consistent and

$$\frac{\sqrt{n}(X_{([np]+1)} - x_p)}{\sqrt{p(1-p)}f(x_p)} \longrightarrow^{d.} N(0,1)$$

for $f(x_p) > 0$.

- Density estimation:

F_n estimates F,

F_n' estimates $f = F'$?

(1). Kernel estimation

When $h > 0$ is small, then

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}.$$

Take

$$\hat{f}_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}$$
$$= \frac{1}{nh}\sum_{i=1}^{n} K_0((X_i - x)/h),$$

where $K_0(x) = 0.5 * I\{|x| \leq 1\}$.

Rosenblatt (1956).

In general

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K((X_i - x)/h),$$

K may be taken as a density.

(1). If $f(x)$ is continuous at x, K is a bounded density and $\lim_{x\to\infty}|x|K(x)=0$, $h\to 0$, $nh\to\infty$, then

$$E\hat{f}_n(x) \longrightarrow f(x)$$

$$f_n(x) \longrightarrow f(x) \ \ a.s.$$

and

$$\sqrt{nh}(f_n(x) - E\hat{f}_n(x)) \longrightarrow^{d.} N(0, f(x)\int K^2(x)dx)$$

(2). If $f(x)$ is continuous uniformly, K has bounded variation, $h \to 0$, $\sum_{n=1}^{\infty} \exp\{-cnh_n^2\} < +\infty$ $(c > 0)$, then

$$\sup_x |f_n(x) - f(x)| \longrightarrow 0 \ \ a.s.$$

- Least neighborhood method:

Given $a_n \in Z^+$, let

$$\alpha_n(x) = \min\{t : \sum_{i=1}^{n} I_{(x-t,x+t]}(X_i) \geq a_n\}$$

$$f_n^*(x) = \frac{a_n}{2n\alpha_n(x)}.$$

- Test $H_0 : F = G, \quad H_1 : F \neq G$

Let

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)|$$

then, under H_0,

$$\sqrt{\frac{nm}{m+n}} D_{m,n} \longrightarrow^{d.} X_0.$$

- Skewness and Kurtosis Test for Normality
- Wilk Test for Normality

$$W = \frac{\sum_{k=1}^{[n/2]} \alpha_k (X_{(n+1-k)} - X_{(k)})}{\sum_{i=1}^{n} (X_{(i)} - \bar{X})^2}.$$

α_k are some constants.

- R^2 and D Test for Normality
- Count Statistic
- Rank Statistic
- U-Statistic

- Regression function estimation

$$m(x) = E(Y|X = x) = \frac{\int yf(x,y)dy}{f(x)}$$

and

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n Y_i K((X_i - x)/h)}{\sum_{i=1}^n K((X_i - x)/h)}.$$