統計計算與模擬

政治大學統計系余清祥

2024年5月7日～5月14日

第七單元 ：估計密度函數

http://csyue.nccu.edu.tw

Data Analysis

- Data analysis can be separated into two parts: **Exploratory Data Analysis (EDA)** and **Confirmatory Data Analysis (CDA)**.

→EDA is to explore the basic characteristics and CDA is to apply proper methods/models for the study goal.

→We tend to use tables & graphs to summarize basic properties of data, which are often crucial in the stage of CDA.

Data Distribution

- Distribution probably is the most essential information of data, we can use it to derive important quantities (e.g., mean & variance).

→Density estimation can be treated as an EDA approach of exploring distribution.

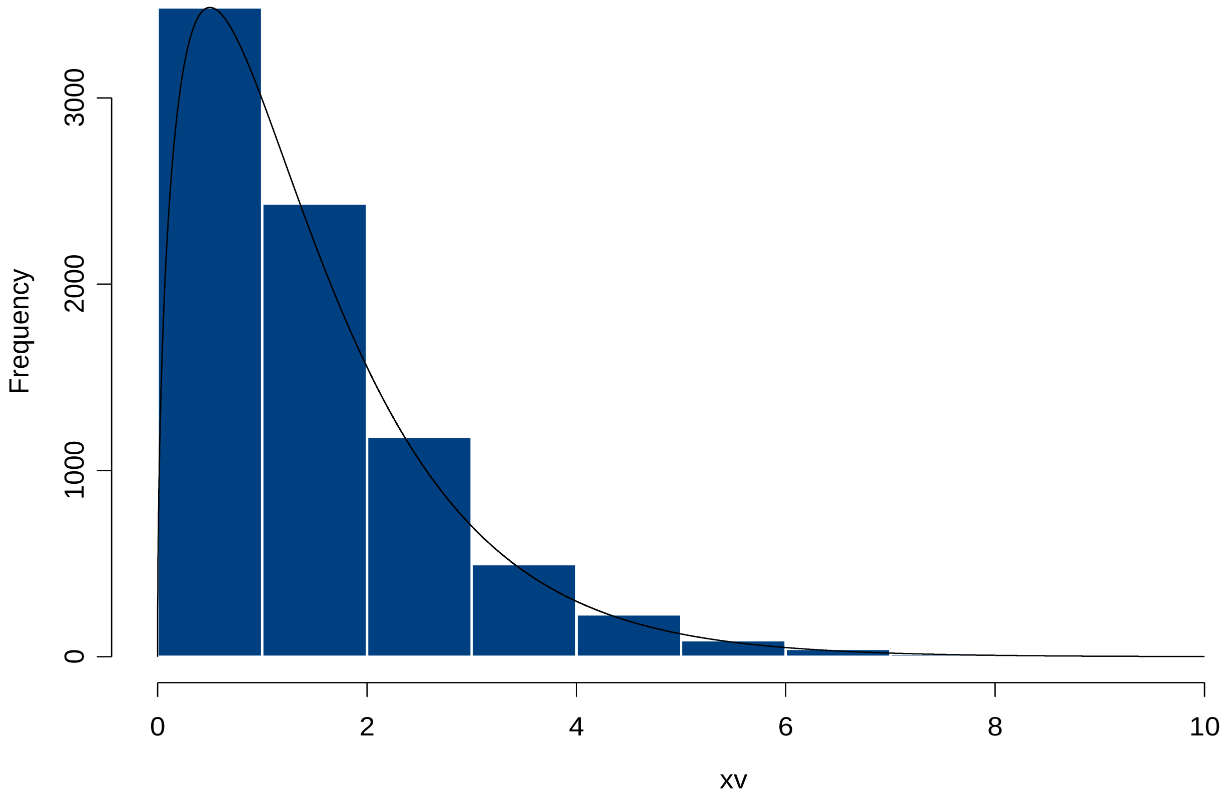
→Bayesian computing (e.g., MCMC) is another approach and it is like Monte Carlo Integration, requiring further assumption about the data.

Density Estimation

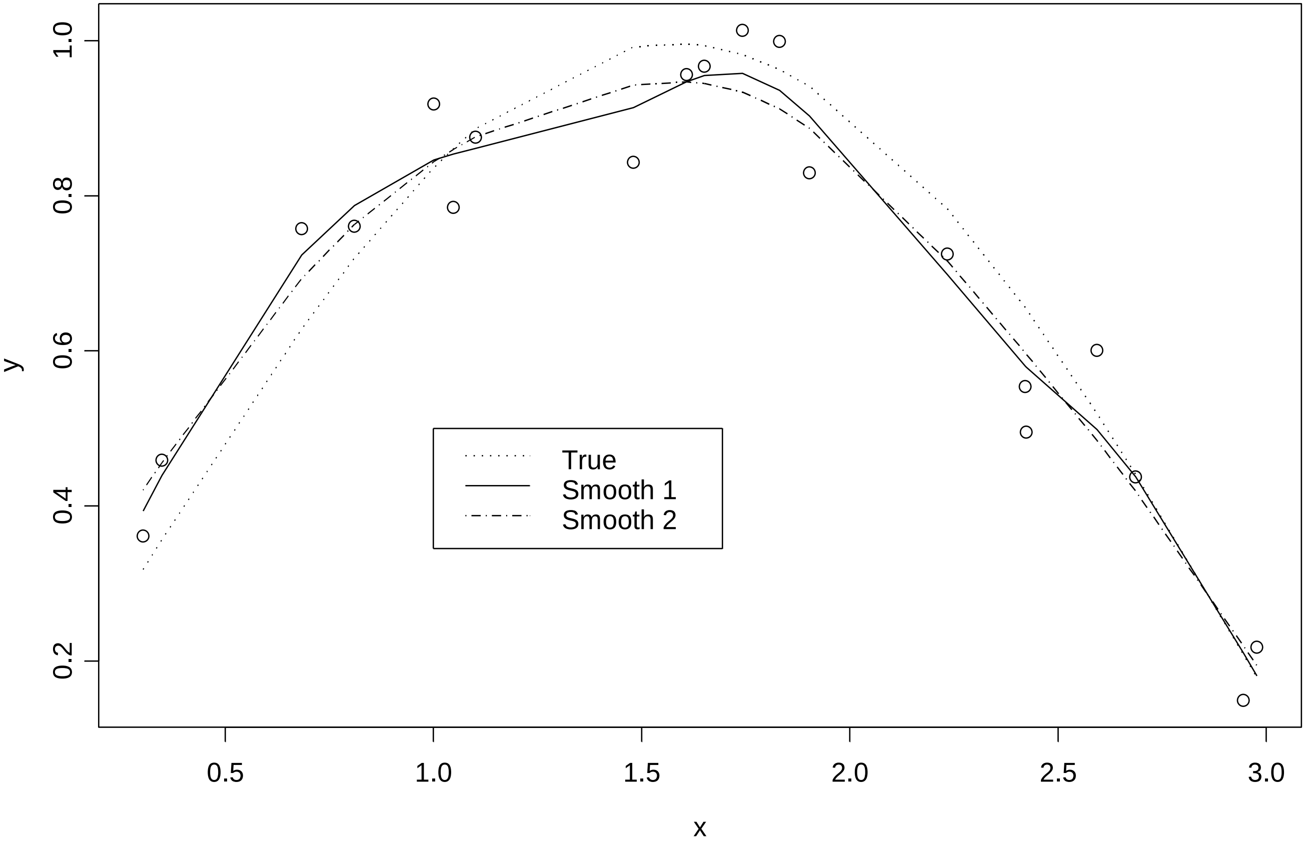
- Estimate the density functions without the assumption that the p.d.f. has a particular form.

→ The idea of estimating c.d.f. (i.e., *F(x₀)*) is to count the number of observations not larger than x_0. Since the p.d.f. is the derivative of c.d.f., the natural estimate of the p.d.f. would be $\hat{f}(x_0) = \sum I\{x_i = x_0\} / n$. However, this is likely not a good estimate since most points have zero density.

Therefore, we may want to assign a nonzero weight to points near points with observations. Intuitively, the weight should be larger if a point is close to a observation, but this is not necessary to be true.



- Smoothing, a process of obtaining a corresponding smooth set of values from irregular set of observed values, is closely linked computationally to density estimation.



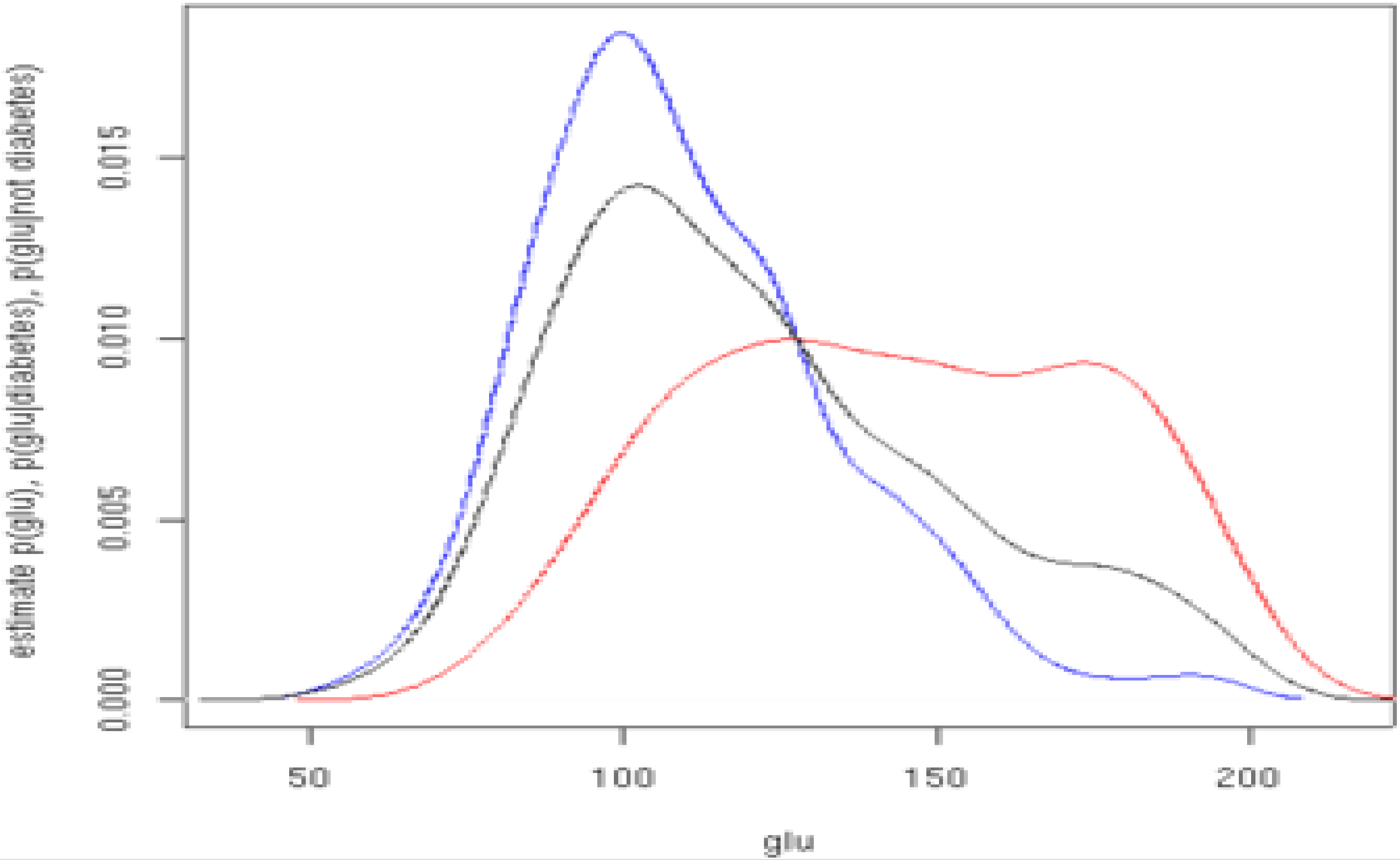
Empirical Examples

- A population of women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes mellitus according to WHO criteria.

→Three density estimates are constructed for “glu” (plasma glucose concentration), based on 532 complete records collected by the *US National Institute of Diabetes and Digestive and Kidney Diseases*.

註：「維基百科」& MASS package

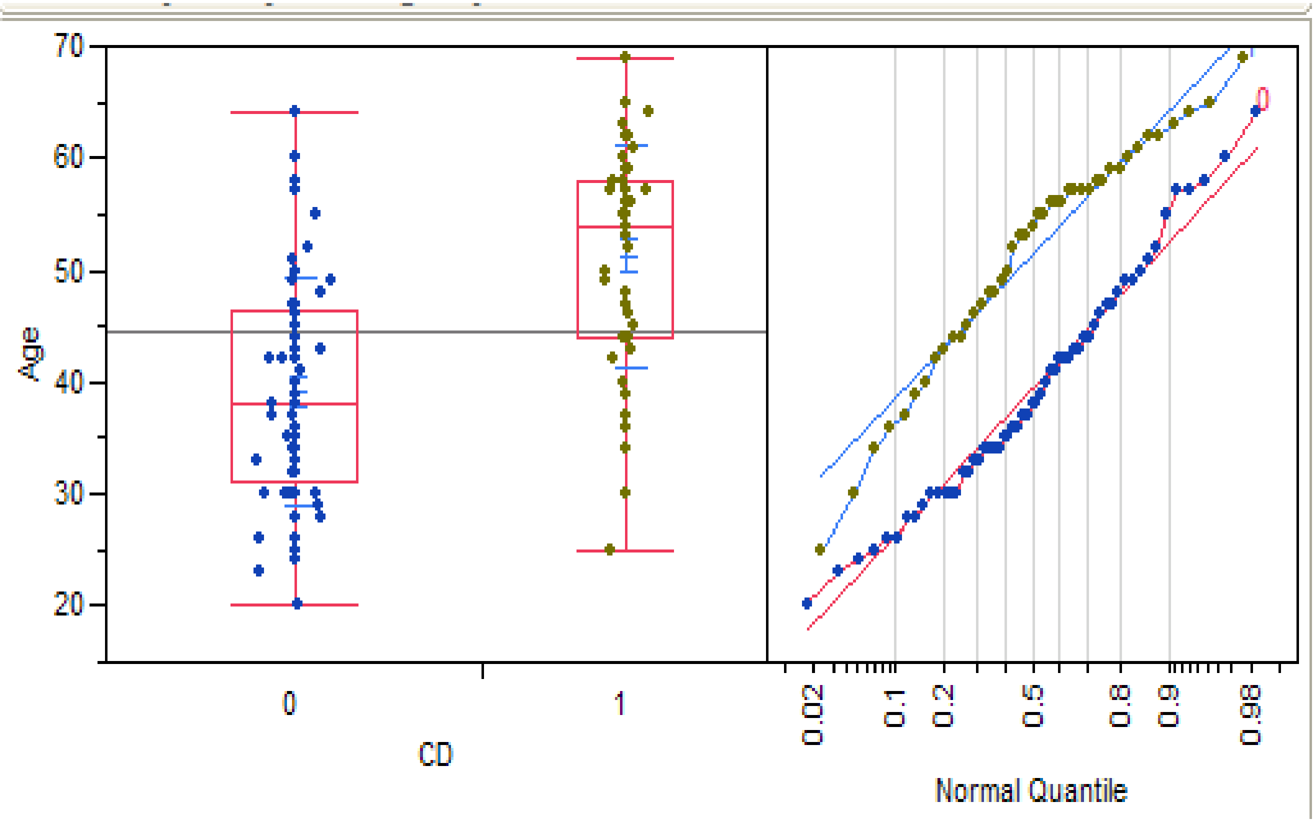
https://en.wikipedia.org/wiki/Density_estimation#Example_of_density_estimation

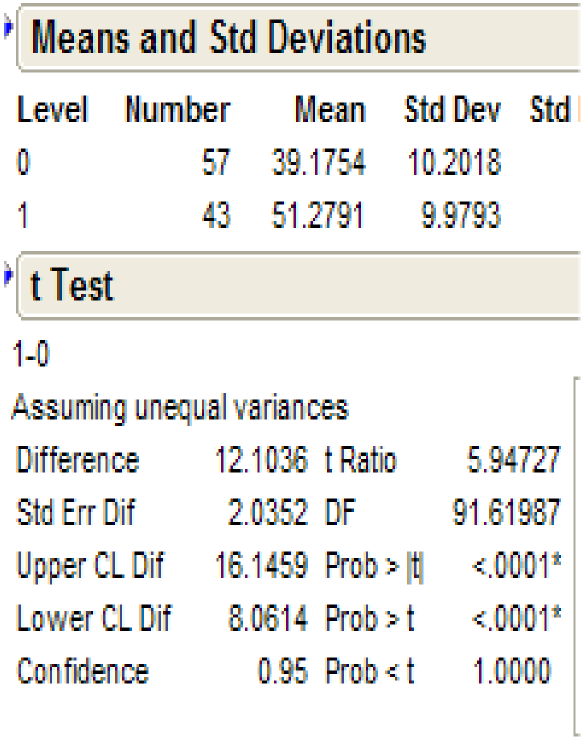


Estimated density of p (glu | diabetes=1),
p (glu | diabetes=0), and **p (glu)**

Logistic Regression

- How can we analyze these data?





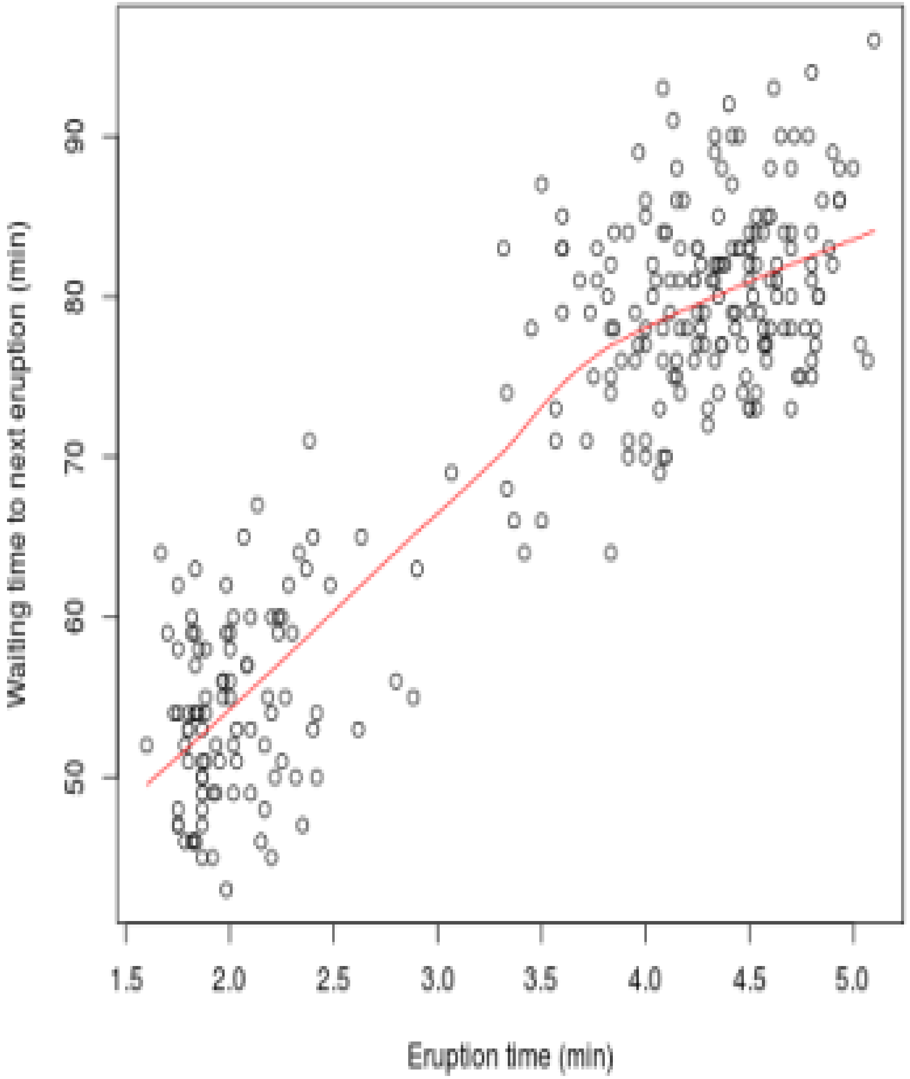
```
Means and Std Deviations

Level   Number     Mean    Std Dev   Std
0           57   39.1754   10.2018
1           43   51.2791    9.9793

t Test

1-0
Assuming unequal variances
Difference      12.1036  t Ratio      5.94727
Std Err Dif      2.0352  DF          91.61987
Upper CL Dif    16.1459  Prob > |t|   <.0001*
Lower CL Dif     8.0614  Prob > t     <.0001*
Confidence         0.95  Prob < t      1.0000
```

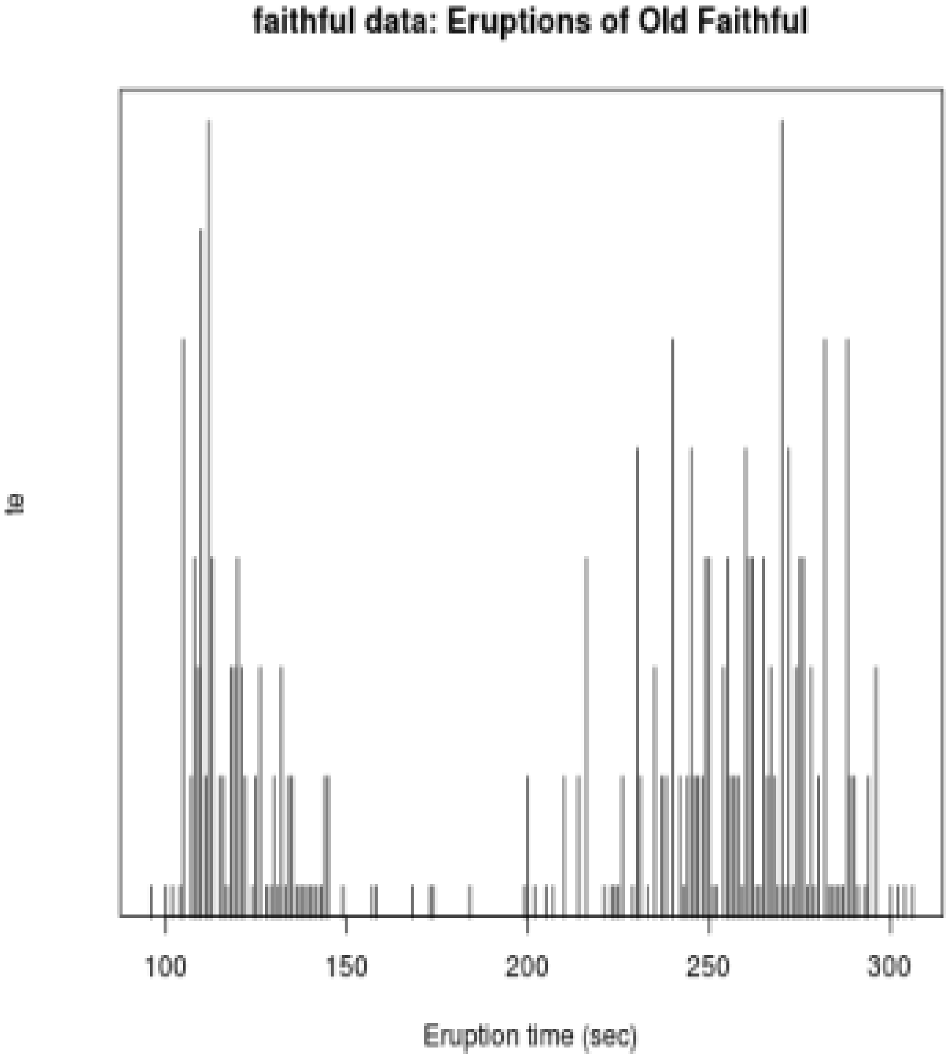
Non-pooled t-test

The mean age of the individuals with some signs of coronary heart disease is 51.28 years vs. 39.18 years for individuals without signs (t = 5.95, p < .0001).

■ Old Faithful Geyser (Yellowstone National Park)

→ It is a highly predictable geothermal feature, & has erupted every 44 minutes to two hours since 2000.




faithful data: Eruptions of Old Faithful

Path of Tropical Storms

(Typhoon, Cyclone, …)

→We want to know the hot spot areas!

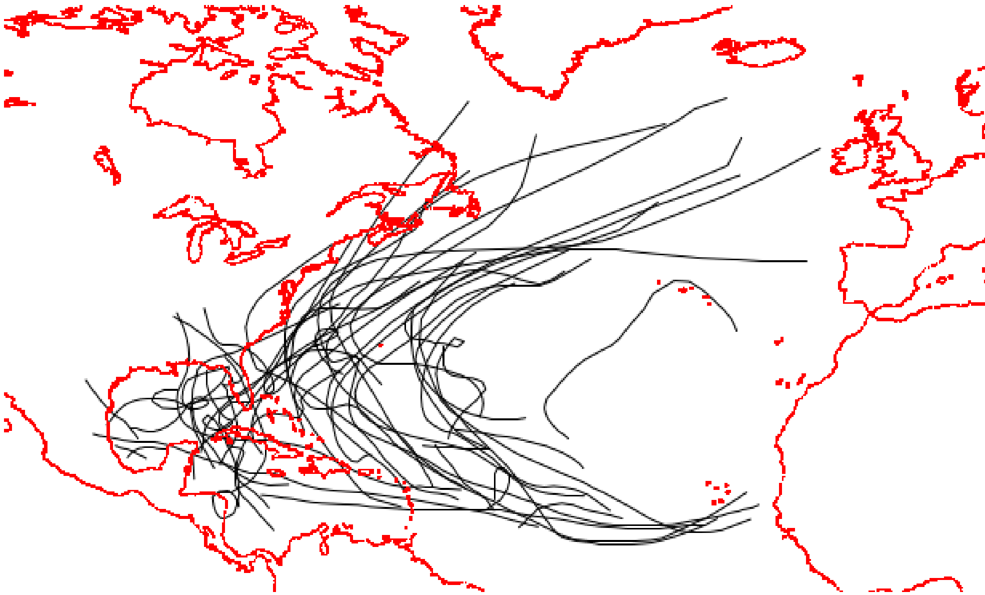
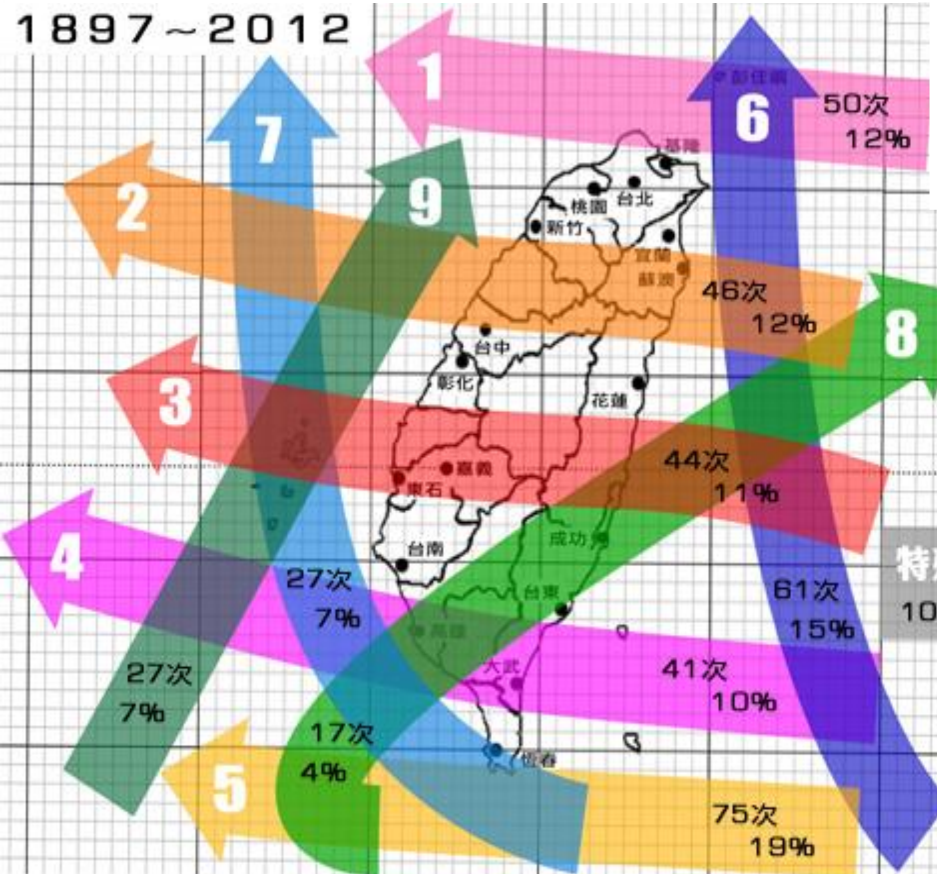
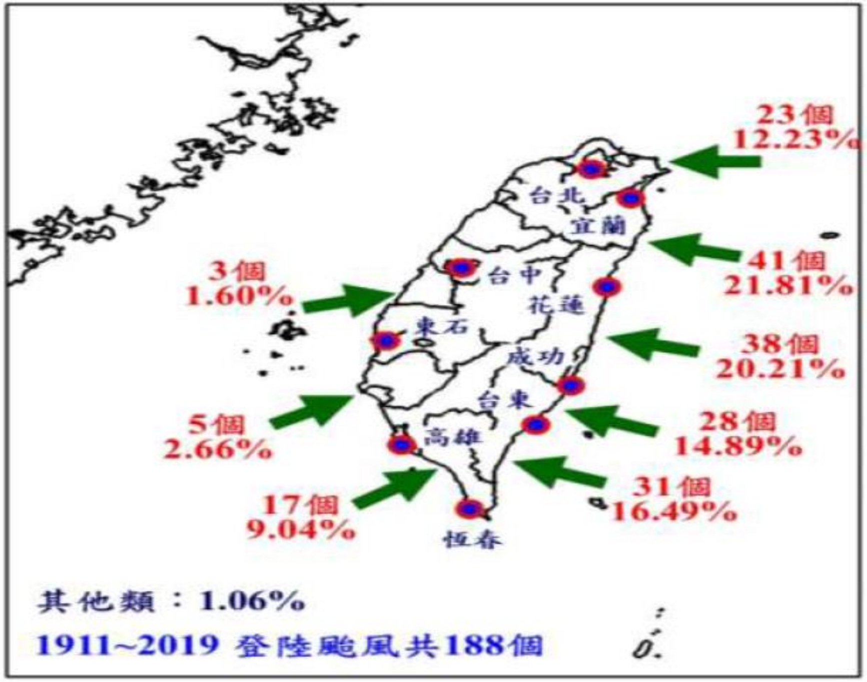


Figure 6: Paths of 40 tropical cyclones in the North Atlantic.

https://www.stat.cmu.edu/~larry/=sml/densityestimation.pdf



http://epaper.wra.gov.tw/Article_Detail.aspx?s=9BDE169D87C509CE



- Histogram

→ The histogram is the simplest and most familiar method of a density estimator.

→ Break the interval [*a*,*b*] into m bins of equal width *h*, say $a = a_0 < a_1 < \cdots < a_{m-1} < a_m = b.$ Then the density estimate of $x \in [a,b]$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{j=1}^{m}\frac{n_j}{h}\cdot I\{x \in [a_{j-1}, a_j]\},$$

where n_j is the number of observations in the interval $x \in [a_{j-1}, a_j]$.

(Note: *h* = *(a-b)/m*)

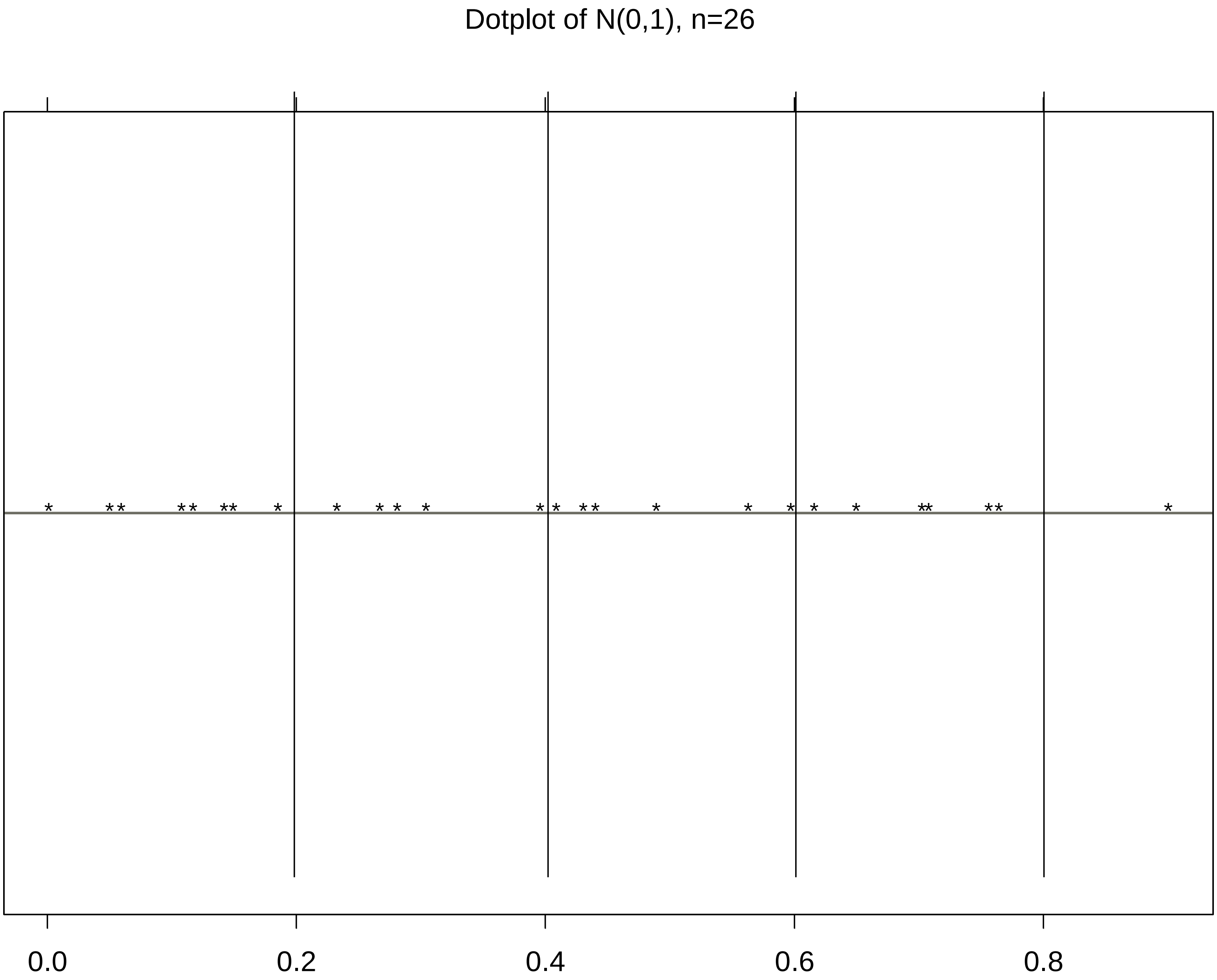
Notes:

(1) The histogram density estimate looks like the sample c.d.f. and is a step function.

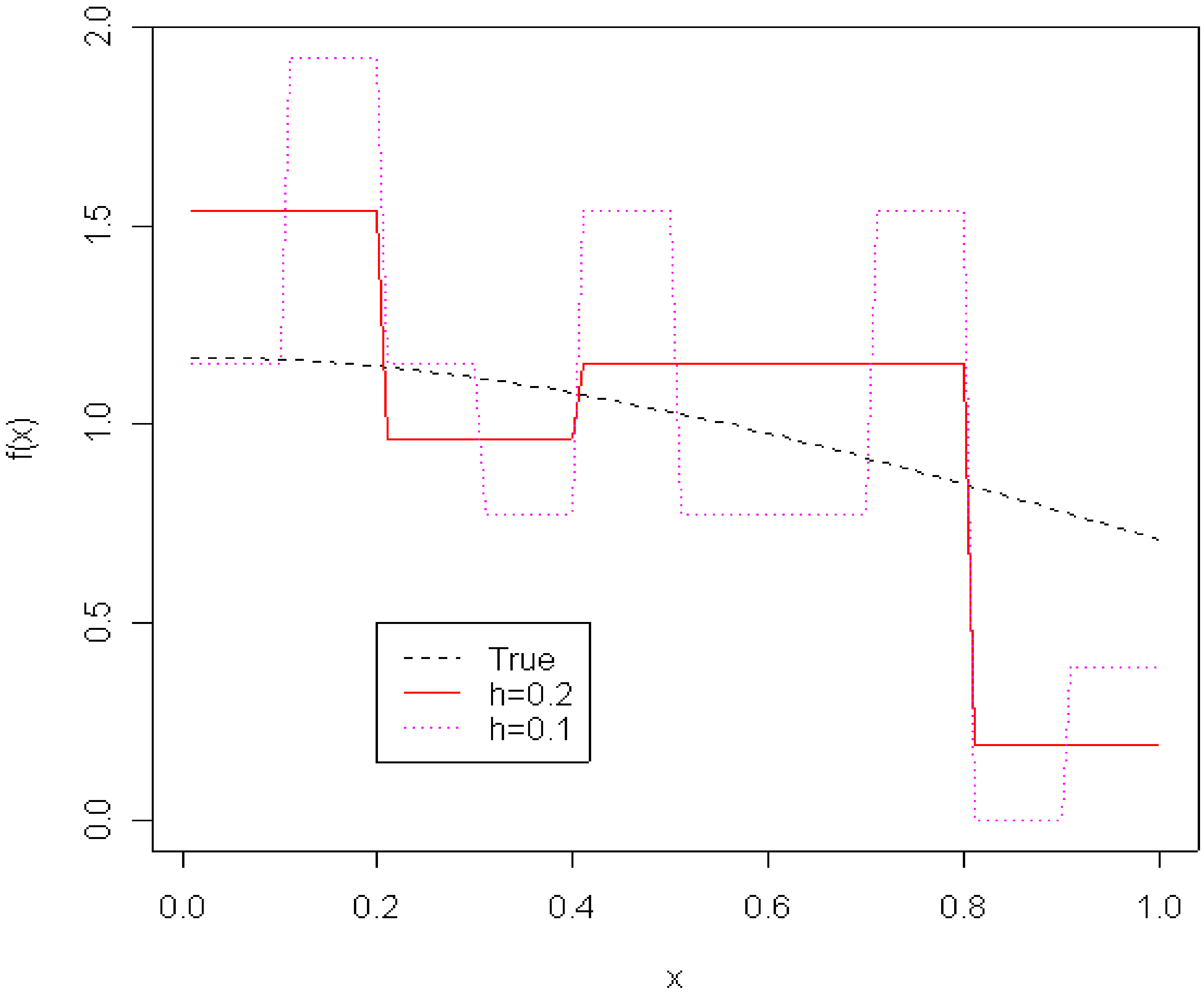
(2) The smaller *h* is, the smoother the density estimate will be. However, given a finite number of observations, when *h* is smaller than most of the distances between two points, the density estimate would become more discrete.

Q: What is the “optimal” choice of *h*?

(Optimal bin = 1.322*logN →Sturge’s Rule)


Dotplot of N(0,1), n=26
0.0
0.2
0.4
0.6
0.8

Histogram Estimate (n=26, N(0,1))



- The Naïve Density Estimator

→ Instead of rectangle, allow the weight is centered on *x*. From Silverman (1986),

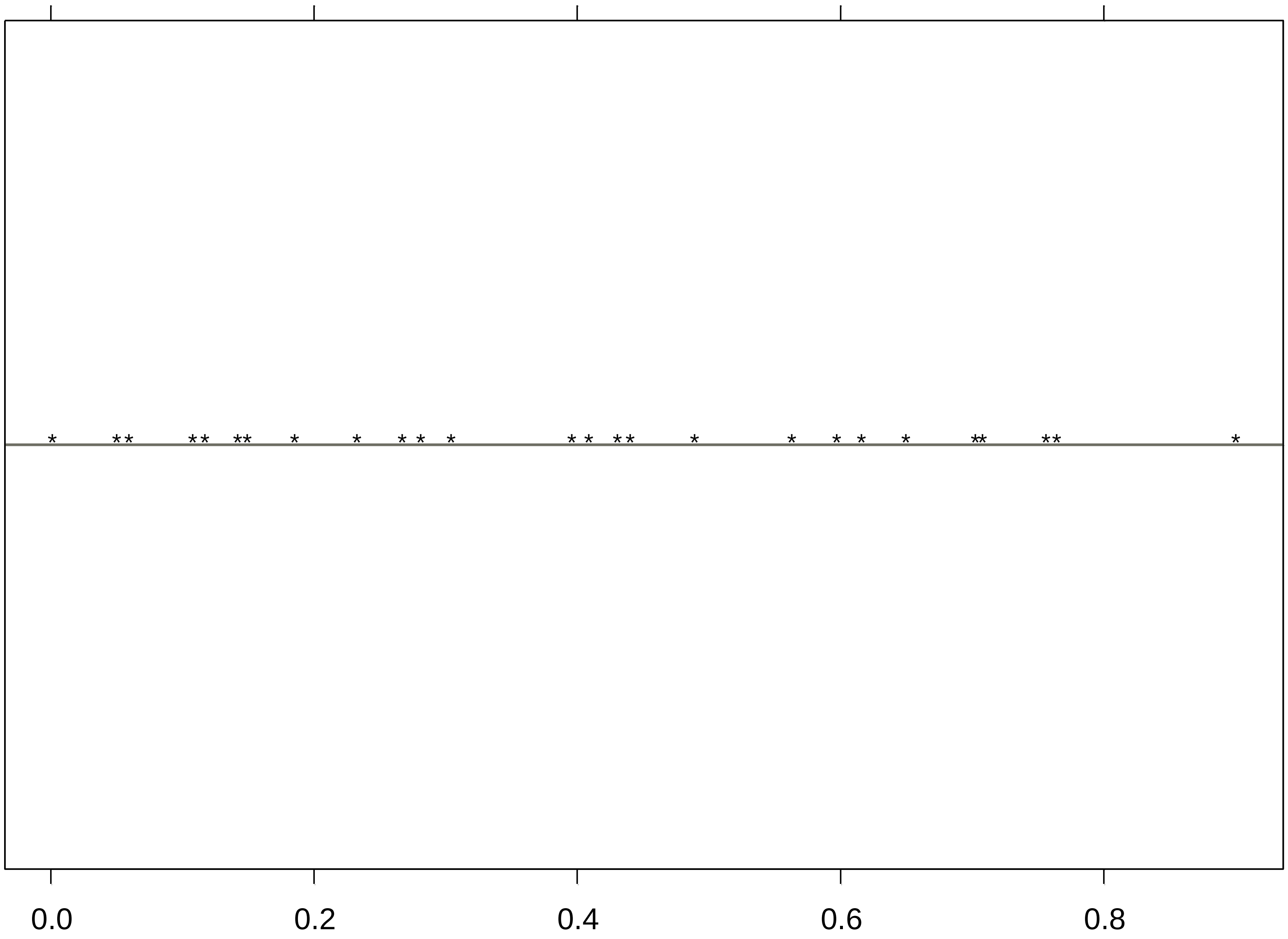
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - x_i}{h} \right),$$

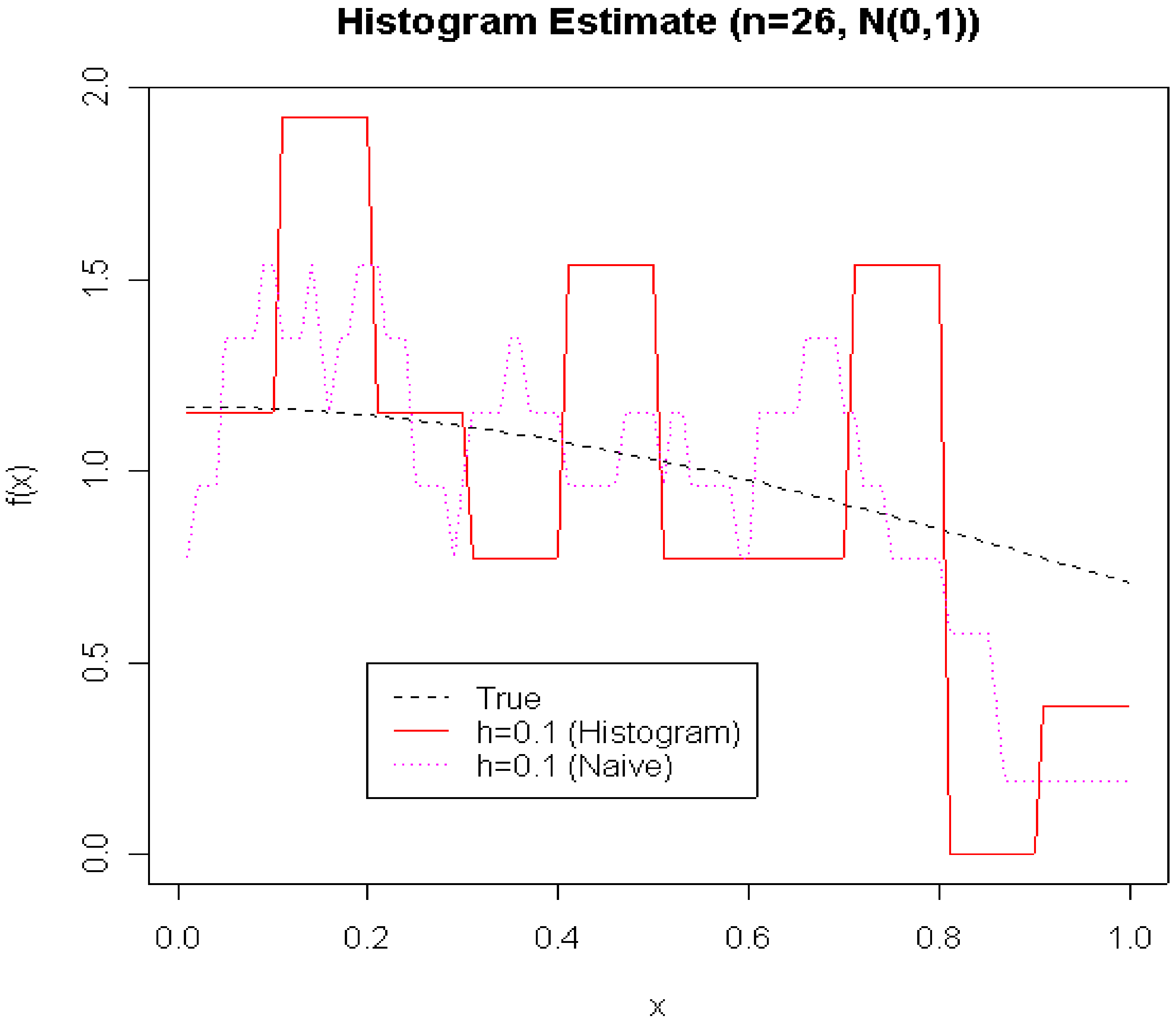
where

$$w(x) = \begin{cases} \frac{1}{2}, & |x| < 1; \\ 0, & Otherwise. \end{cases}$$

Because the estimate is constructed from a moving window of width 2*h*, it is also called a *moving-window histogram*.

Dotplot of N(0,1), n=26




Histogram Estimate (n=26, N(0,1))
f(x)
0.0
0.5
1.0
1.5
2.0
x
0.0
0.2
0.4
0.6
0.8
1.0
True
h=0.1 (Histogram)
h=0.1 (Naive)

- Kernel Estimator:

→ The naïve estimator is better than the histogram, since weight is based on distance between observations and *x*. However, it also has jumps (similar to the histogram estimate) at the observation points. By modifying the weight function w(.) to be more continuous, the raggedness of the naïve estimate can be overcome.

→ The kernel estimate is as following:

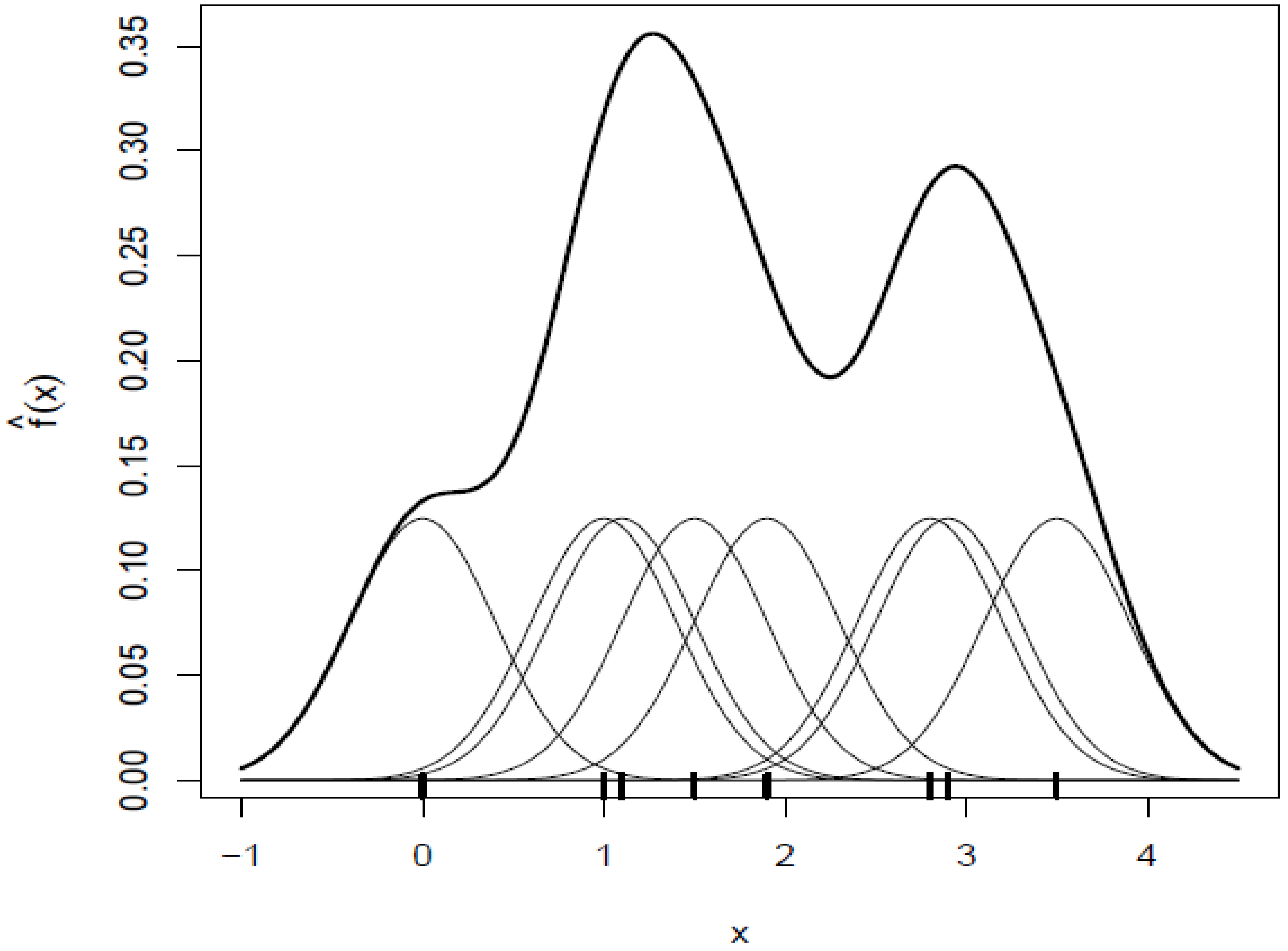
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right),$$

where $\int_{-\infty}^{\infty} K(t)dt = 1$ is the kernel of the estimator.

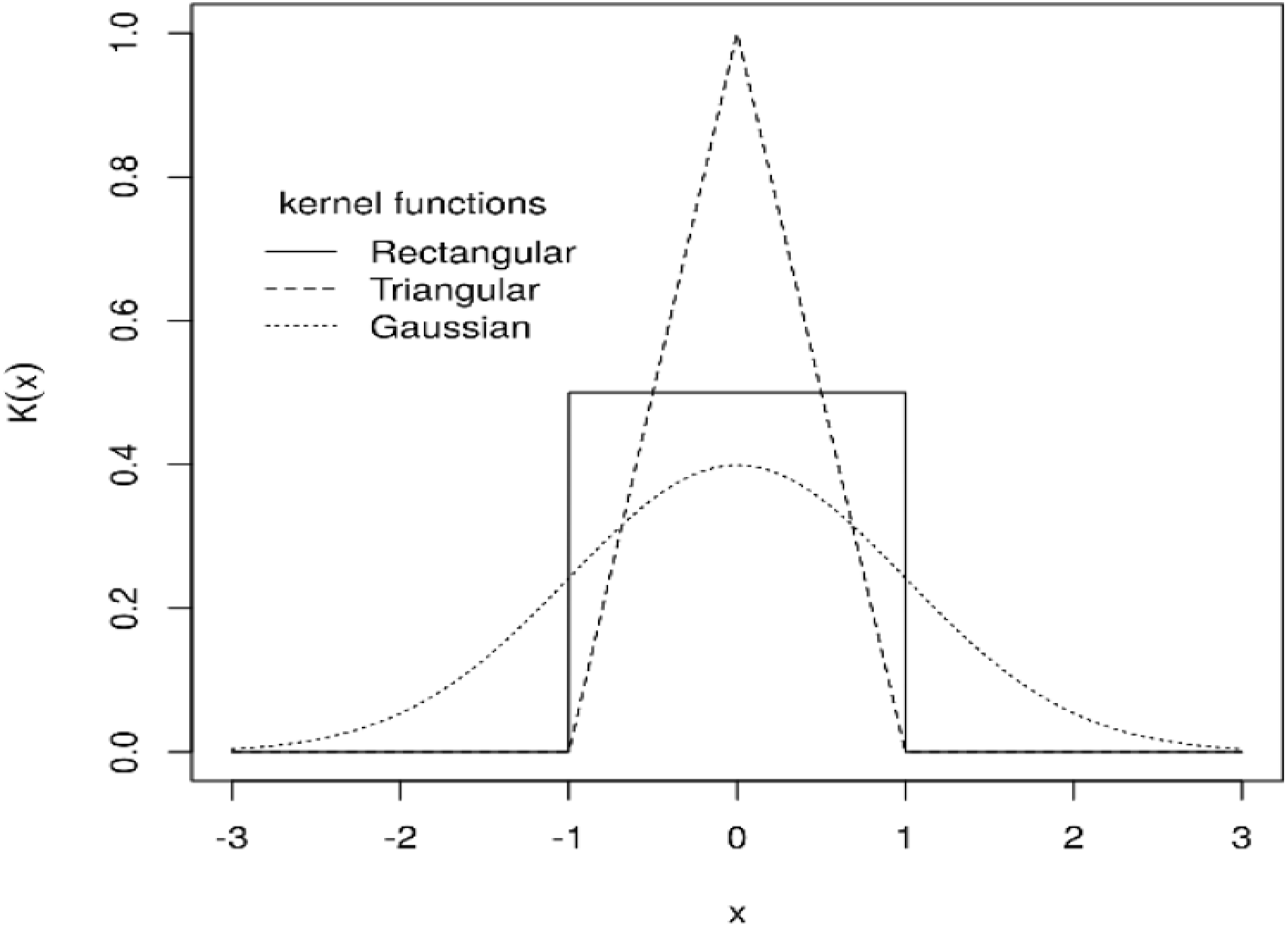
→ Usual choices of kernel functions:

Guassian (i.e., normal), Cosine, Rectangular, Triangular, Lapalce.

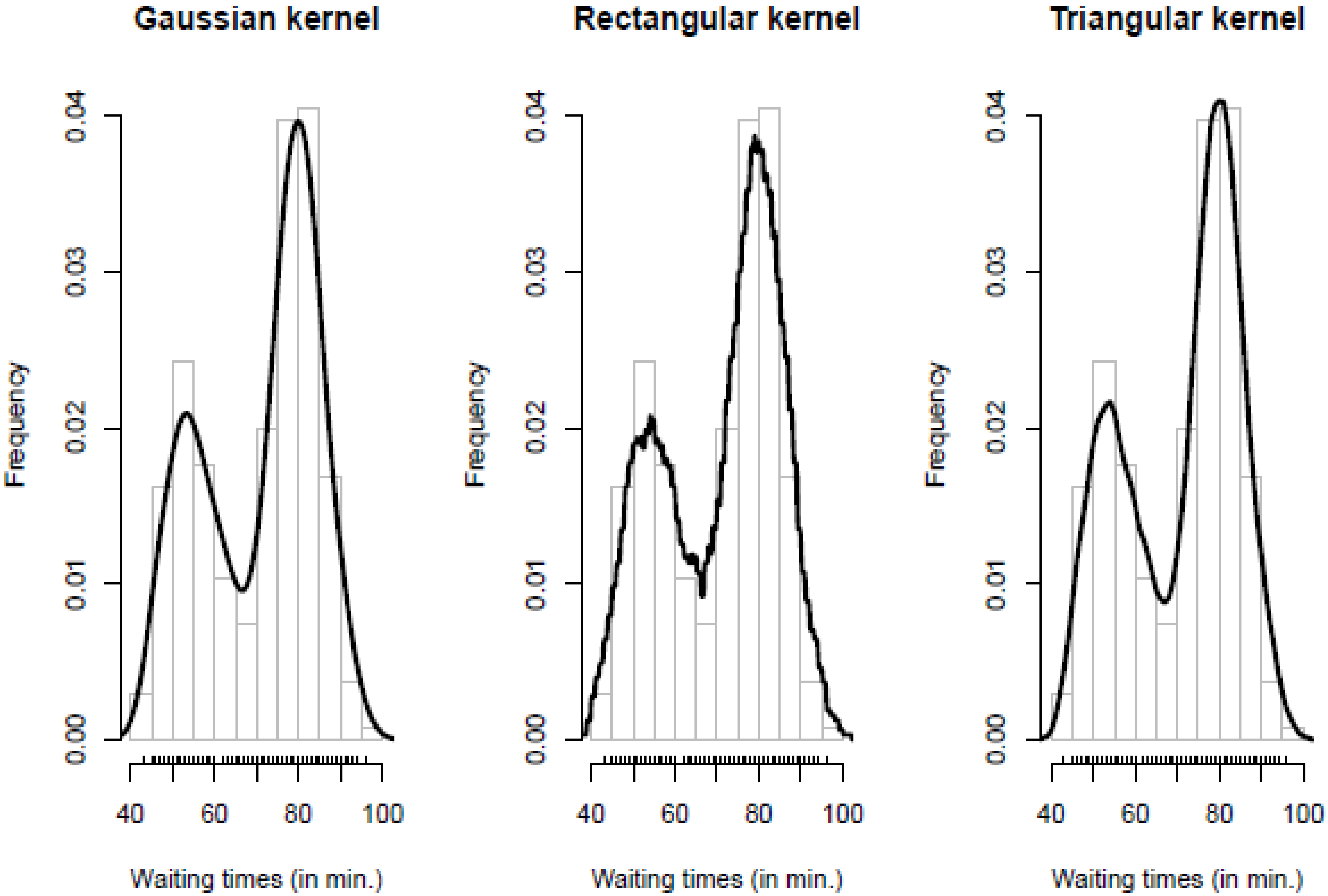
Note: The choice of the bandwidth (i.e., *h*) is more critical than the kernel function.



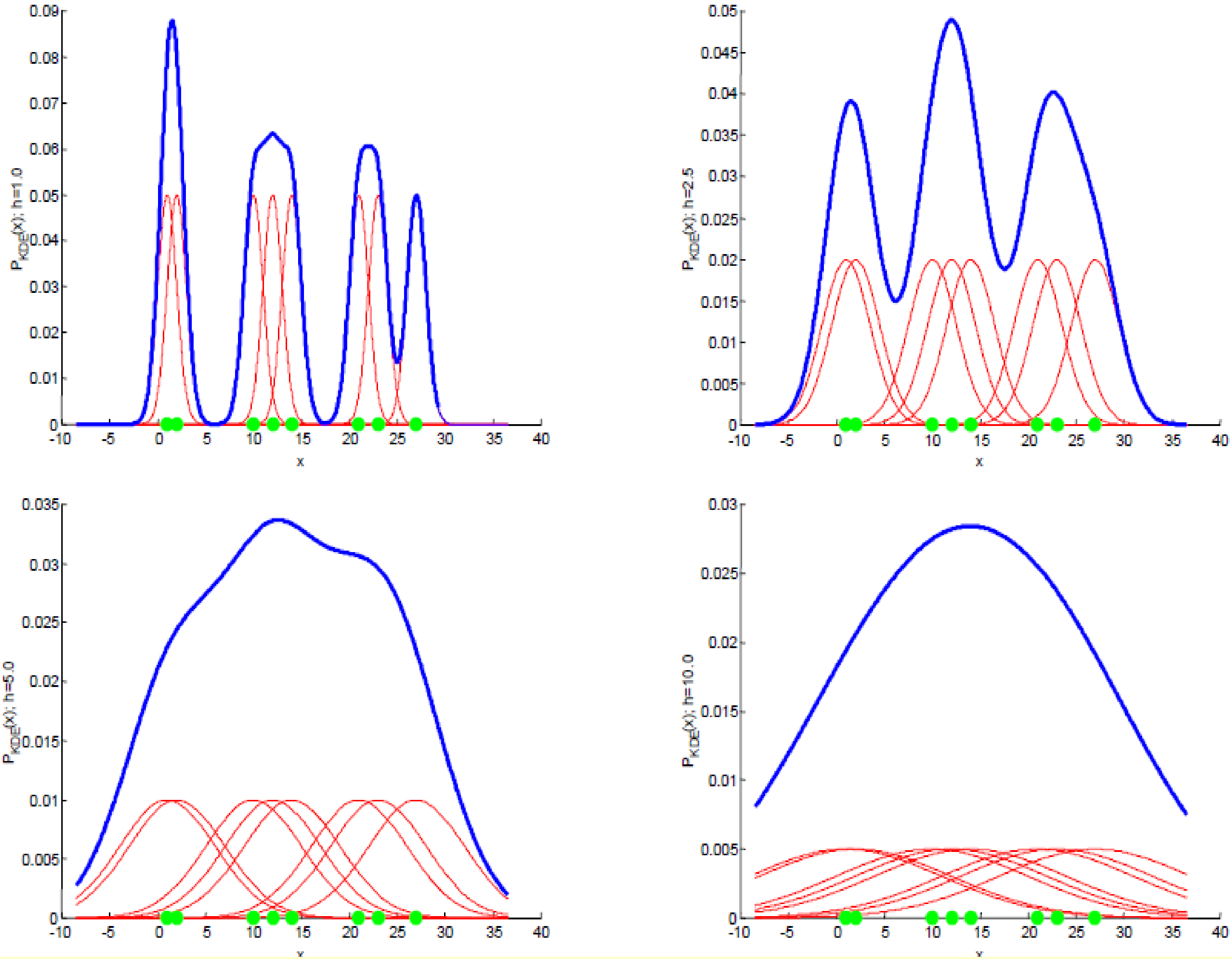
Example of a Kernel Estimator



不同核函數(Kernel Function)的比較

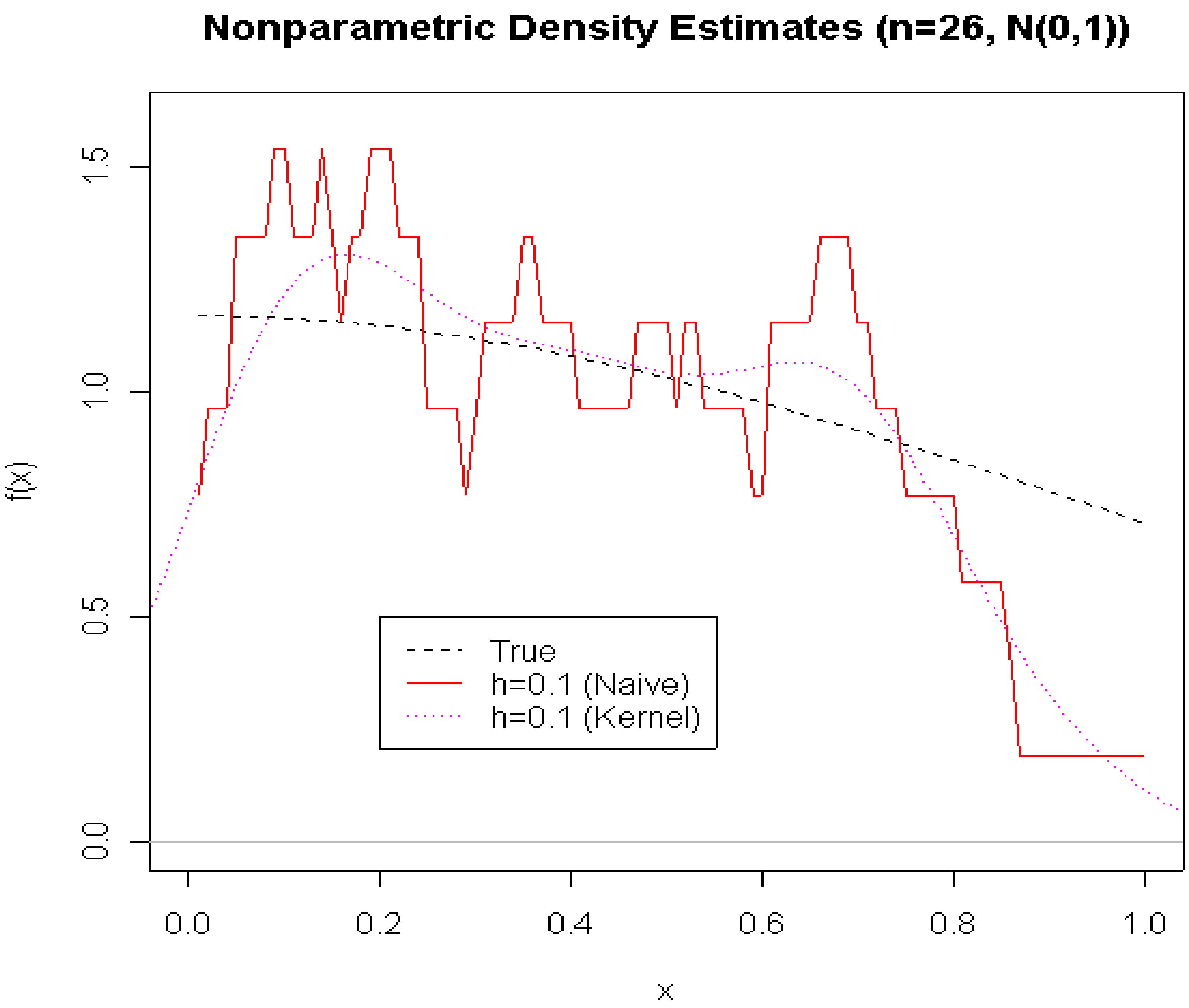

Gaussian kernel
Rectangular kernel
Triangular kernel
Frequency
0.00
0.01
0.02
0.03
0.04
40
60
80
100
Waiting times (in min.)

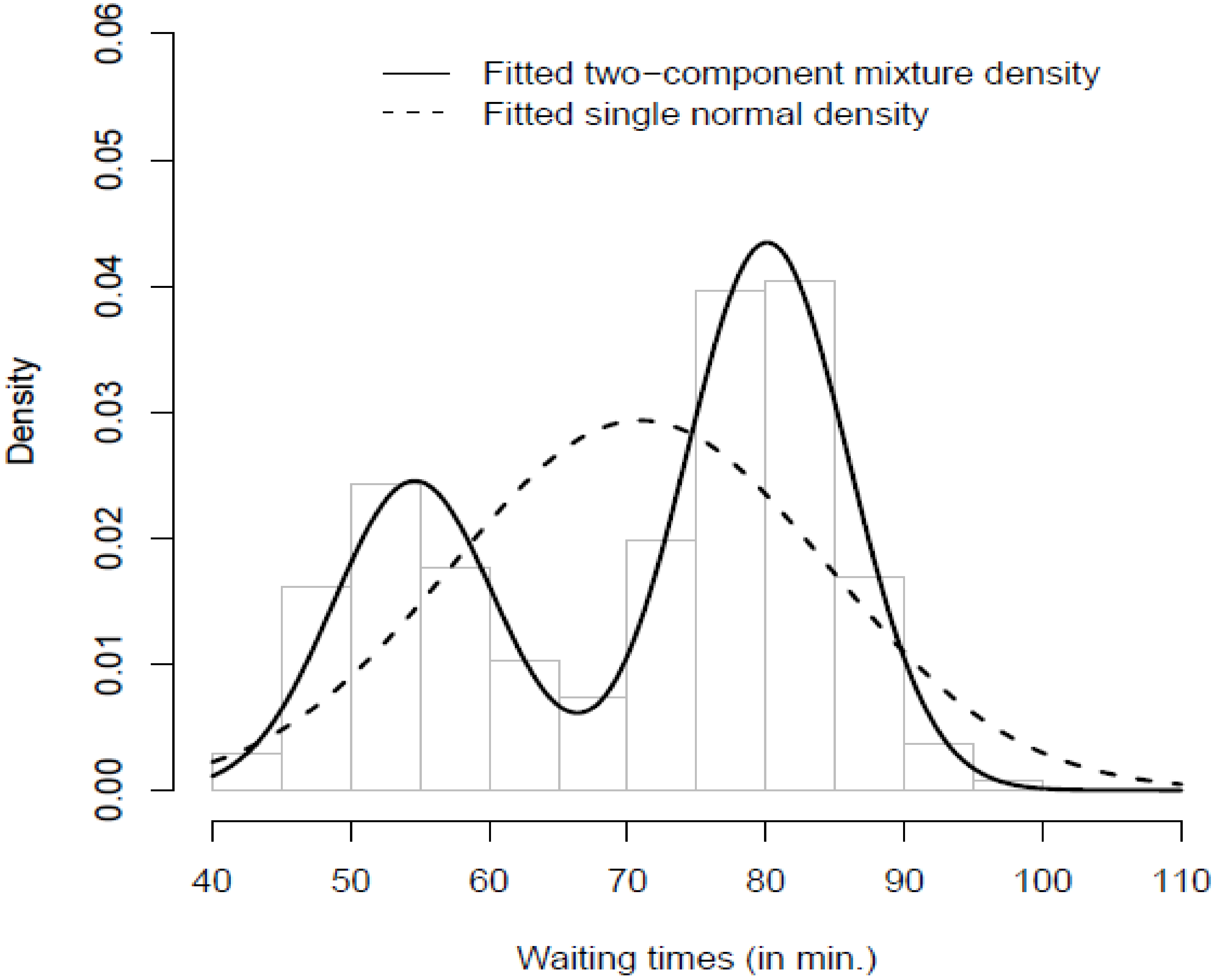
不同核函數的比較(續)



P_KDE(x); h=1.0
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
-10
-5
0
5
10
15
20
25
30
35
40
x
P_KDE(x); h=2.5
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-10
-5
0
5
10
15
20
25
30
35
40
x
P_KDE(x); h=5.0
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-10
-5
0
5
10
15
20
25
30
35
40
x
P_KDE(x); h=10.0
0.03
0.025
0.02
0.015
0.01
0.005
0
-10
-5
0
5
10
15
20
25
30
35
40
x

不同核修勻的環寬比較


Nonparametric Density Estimates (n=26, N(0,1))
f(x)
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
x
True
h=0.1 (Naive)
h=0.1 (Kernel)



Some Variations of Kernel Estimation

- Nearest-neighbor Estimator: (NNE)

→ Another usage of observations is to use the concept of "nearness" between points and observations. But, instead of distance, the nearness is measured according to the number of other observations between a point and the specified observation.

For example, if x_0 and x are adjacent in the ordering, then x is said to be 1-neighbor of x_0. If another observation between x_0 and x, then x is said to be 2-neighbor of x_0.

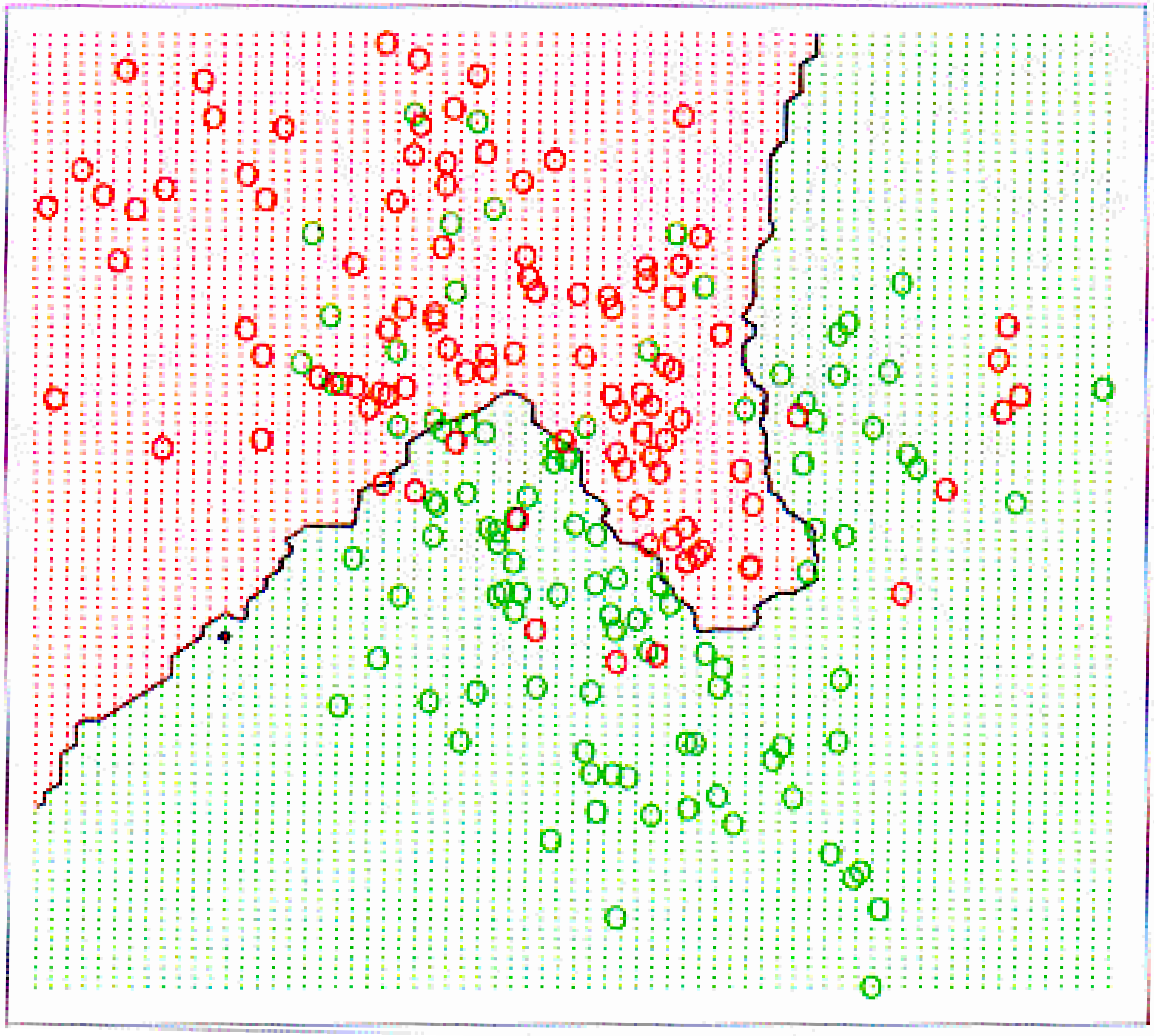
→ The nearest-neighbor density estimates are based on averages of the *k* nearest neighbors in the sample to the point *x*:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_k(x)}K\left(\frac{x-x_i}{h_k(x)}\right),$$

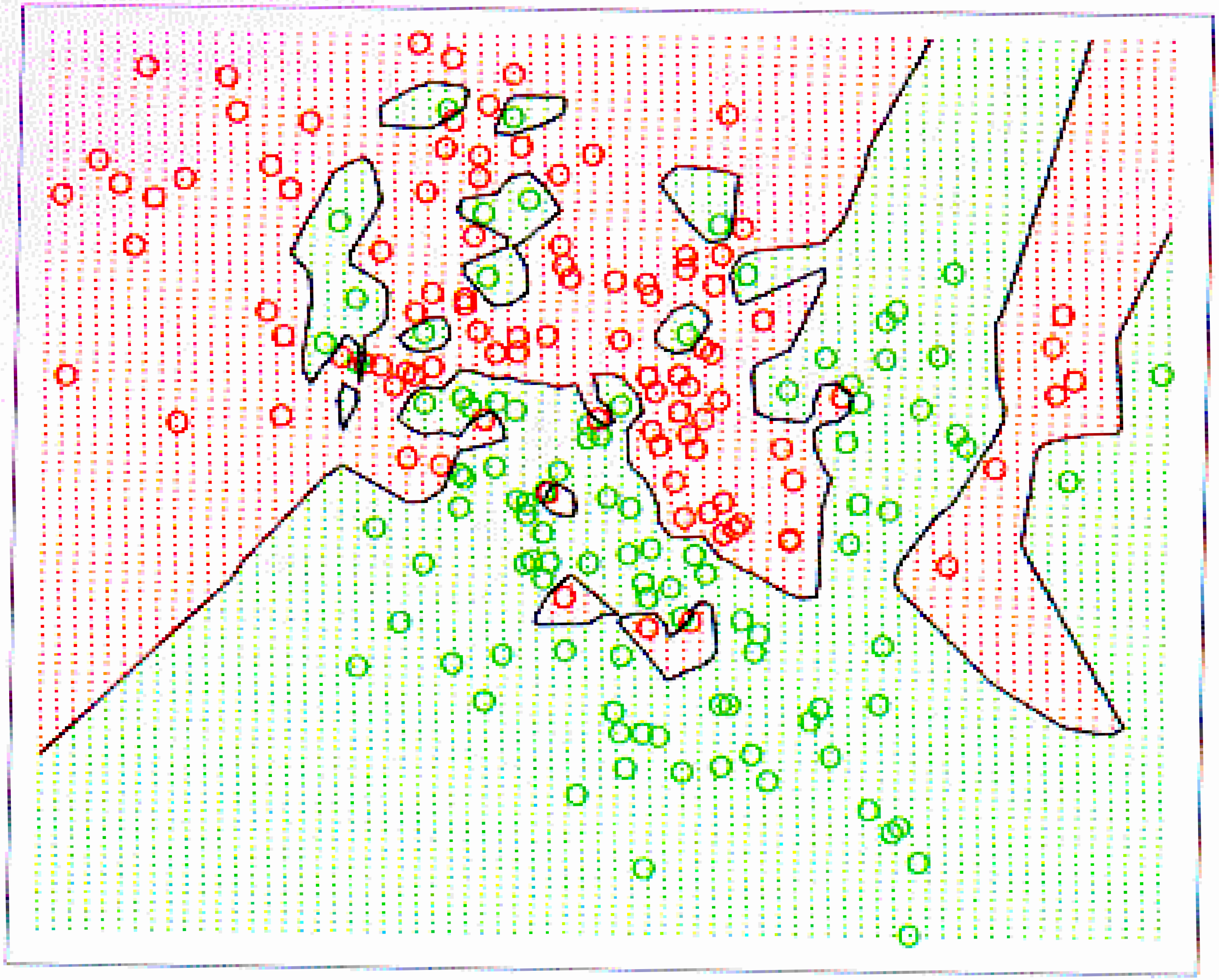
where $h_k(x)$ is the half-width of the smallest interval centered at *x* containing *k* data points.

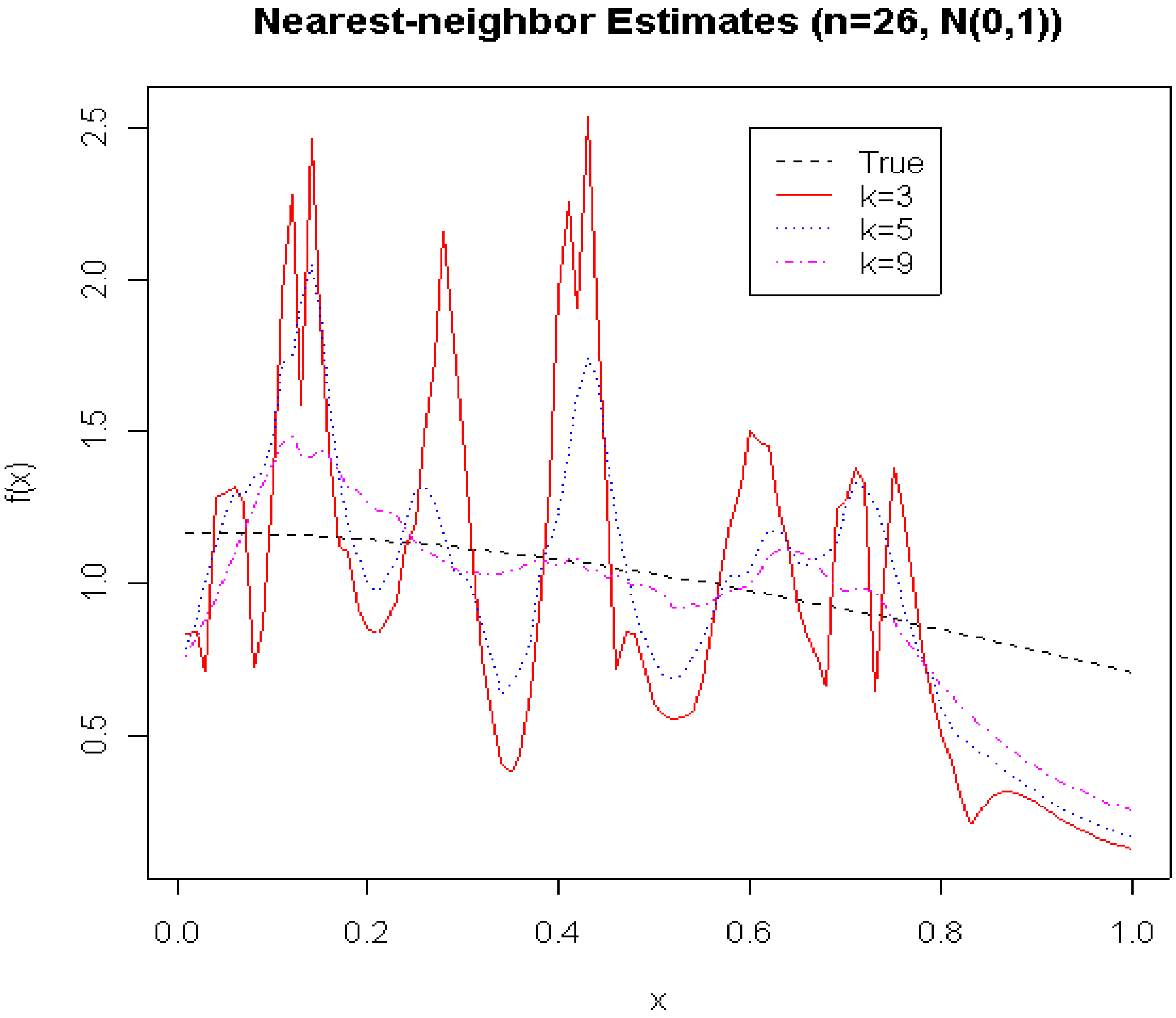
Note: Unlike kernel estimates, the NNE use variable-width window.

15-Nearest Neighbor Classifier



1-Nearest Neighbor Classifier




Nearest-neighbor Estimates (n=26, N(0,1))
True
k=3
k=5
k=9
f(x)
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
x

About Density Estimation

- Basically, there are two major concern in density estimation:

→Local vs. Global

Using more observations might incur bias but fewer observations have larger variance, i.e., a dilemma between $(\text{Bias})^2$ vs. Variance.

→Edge Effect

The end points have very few observations and the estimates may have larger variance.

- Linear Smoother:

→ The goal is the smooth estimates $\hat{M}$ of a regression function $M(x) = E(Y \mid X = x)$. A well-known example is the ordinary linear regression, where the fitted values are

$$\hat{y} = Hy, \ where \ H = X(X'X)^{-1}X'.$$

→ *A Linear Smoother* is the one which the smooth estimate satisfies the following form:

$$\hat{y} = S\,y,$$

where S is an $n \times n$ matrix depending on X.

- Running Means:

→ The simplest case is the running-mean smoother which computes $\hat{y}_i$ by averaging y_j's for which x_j falls in a neighborhood of x_i.

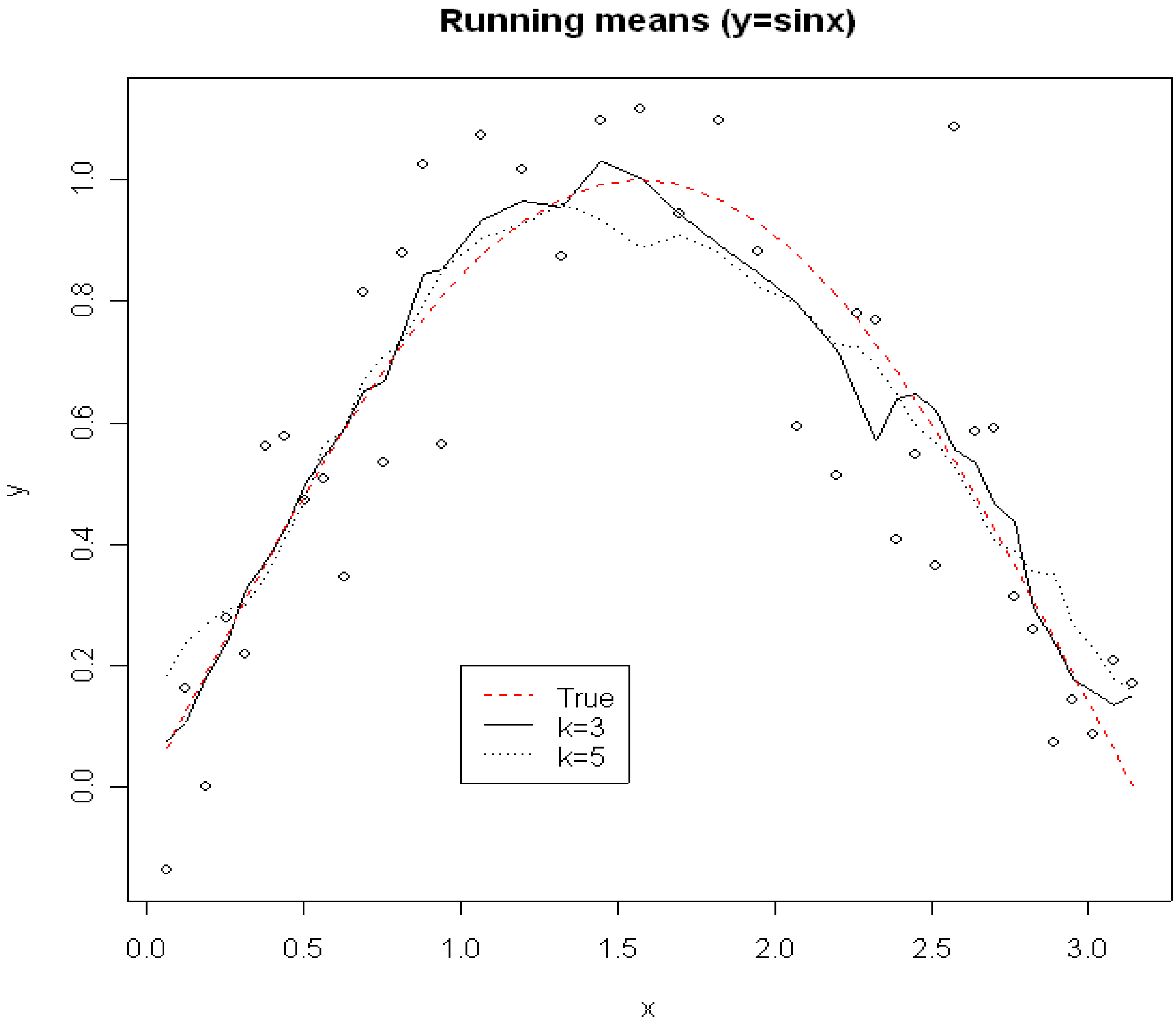
→ One possible choice of the neighborhood N_i is to adapt the idea in *Nearest-neighbor* where N_i is the one with points x_j for which $|i-j| \leq k$. Such a neighborhood contains k points to the left and k points to the right. (Note: The two tails have fewer points and could be less smooth.)

Note: The parameter *k*, called the *span* of the smoother, controls the degree of smoothing.

- Example 2. We will use the following data to demonstrate the linear smooth methods introduced in this handout. Suppose that

$$Y_i = \sin X_i + \varepsilon_i, \quad 0 \le X_i \le \pi,$$

where the noise ε_i is normally distributed with mean 0 and variance 0.04. Also, the setting of X is 15 points on [0,0.3π], 10 points on [0.3 π,0.7π] and 15 points on [0.7π,π].


Running means (y=sinx)
True
k=3
k=5
y
x
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

- Kernel Smoothers

→ The product of a running-mean smoother is usually quite unsmooth, since observations are getting equal weight regardless their distance to the point to be estimated. The kernel smoother with kernel *K* and window 2*h* uses

$$\hat{y}_i = \sum_{j \in N_i} w_{ij} y_j ,$$

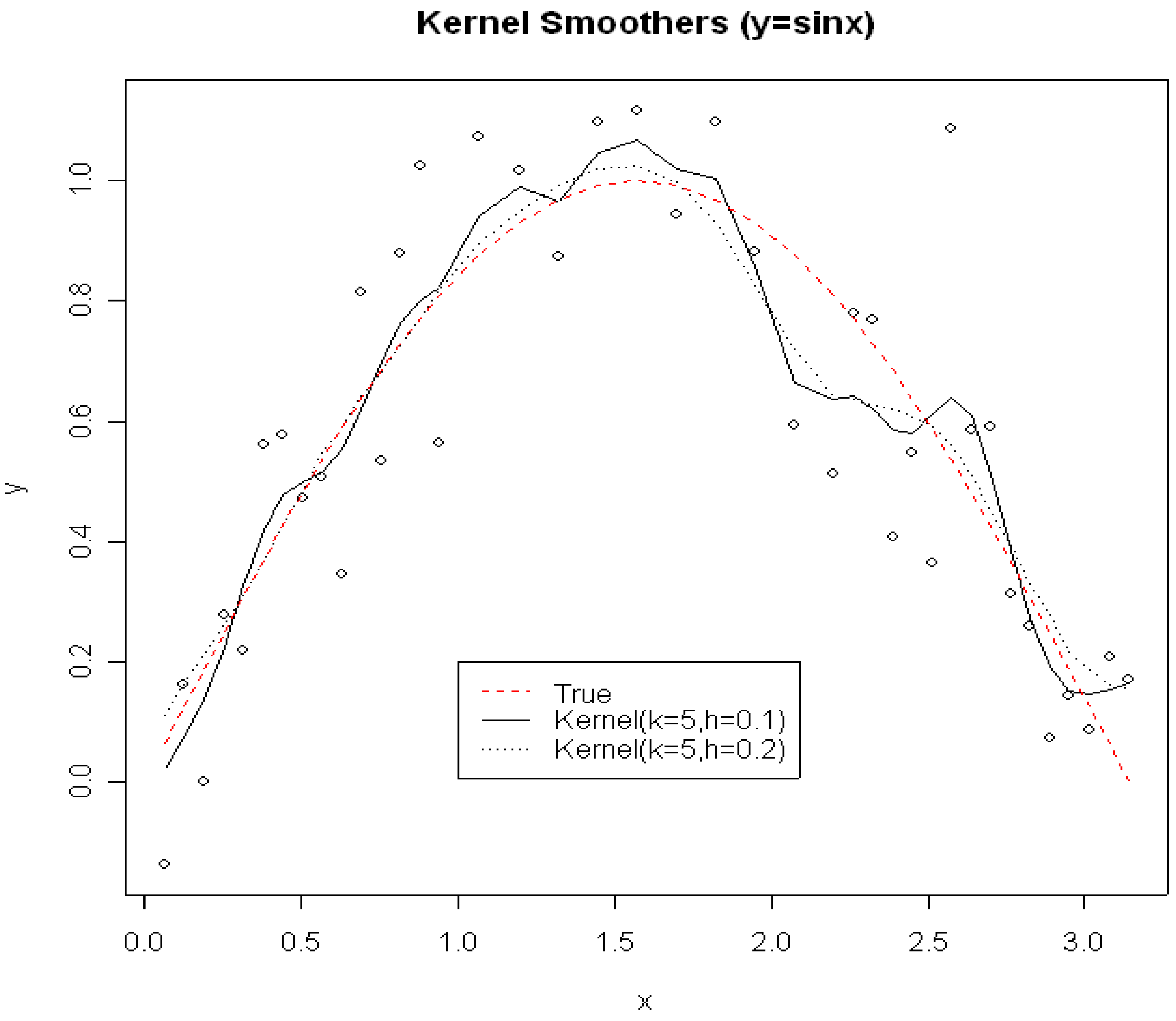
where

$$w_{ij} = K\left(\frac{x_i - x_j}{h}\right) / \sum_{j \in N_i} K\left(\frac{x_i - x_j}{h}\right)$$

Notes:

(1) If the kernel is smooth, then the resulting output will also be smooth. The kernel smoother estimate can thus be treated as a weighted sum of the (smooth) kernels,

(2) The kernel smoothers also cannot correct the problem of bias in the corners, unless the weight of observations can be negative.


Kernel Smoothers (y=sinx)
y
x
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
True
Kernel(k=5,h=0.1)
Kernel(k=5,h=0.2)

- Spline Smoothing:

→ For the linear smoothers discussed previously, the smoothing matrix S is symmetric, has eigenvalues no greater than unity, and produce linear functions.

→ The smoothing spline is to select $\hat{M}$ so as to minimize the following objective function:

$$S_\lambda(M) = \frac{1}{n}\sum_{i=1}^{n}[y_i - M(x_i)]^2 + \lambda\int_a^b [M''(t)]^2\,dt,$$

where $\lambda \geq 0$ and $M \in C^3$.

- What are *Splines*?

→ Spline functions, often called splines, are smooth approximating functions that behave very much like polynomials.

→ Splines can be used for two purposes:

(1) Approximate a given function (Interpolation)

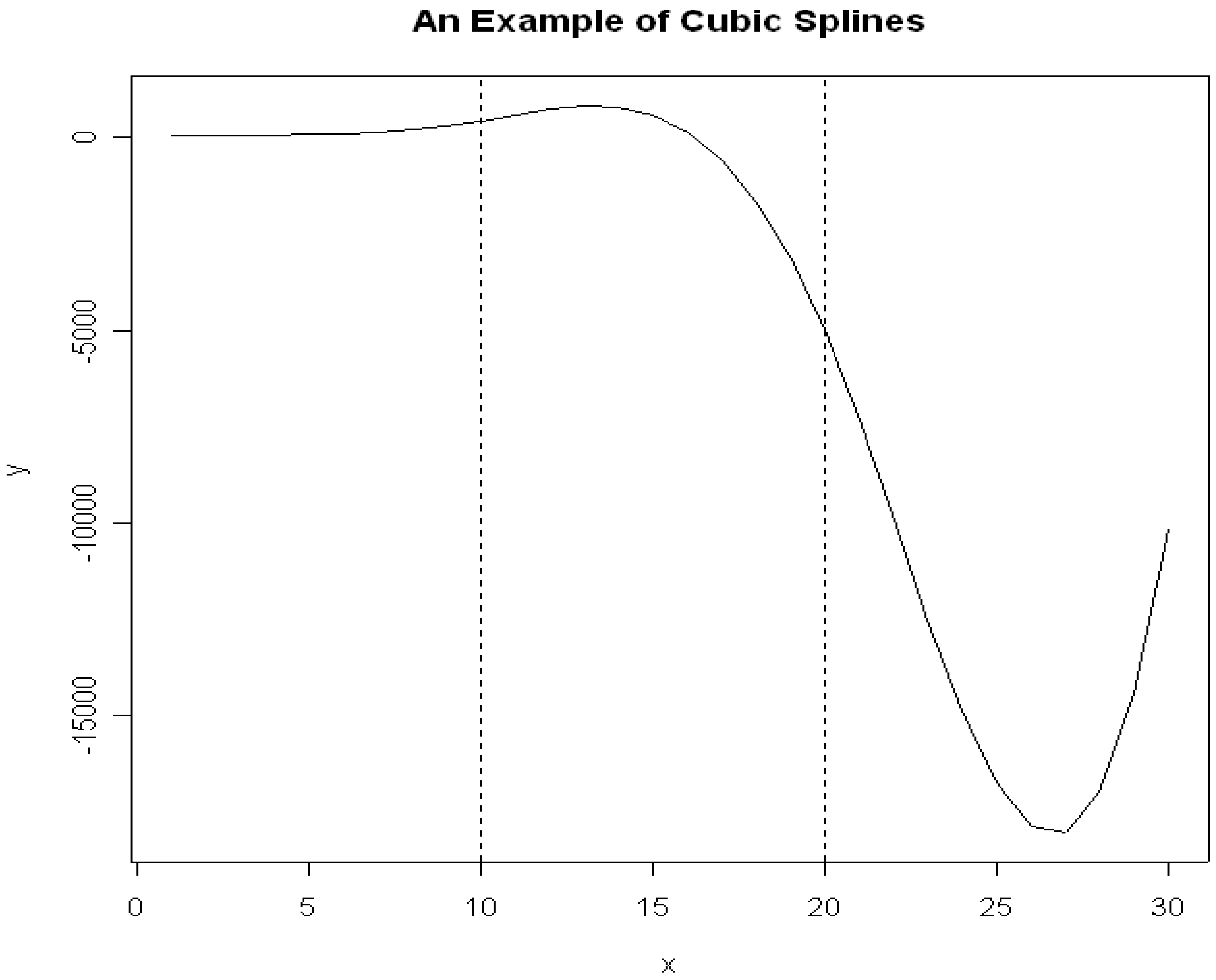
(2) Smooth values of a function observed with noise

Note: We use terms “interpolating splines” and “smoothing splines” to distinguish.

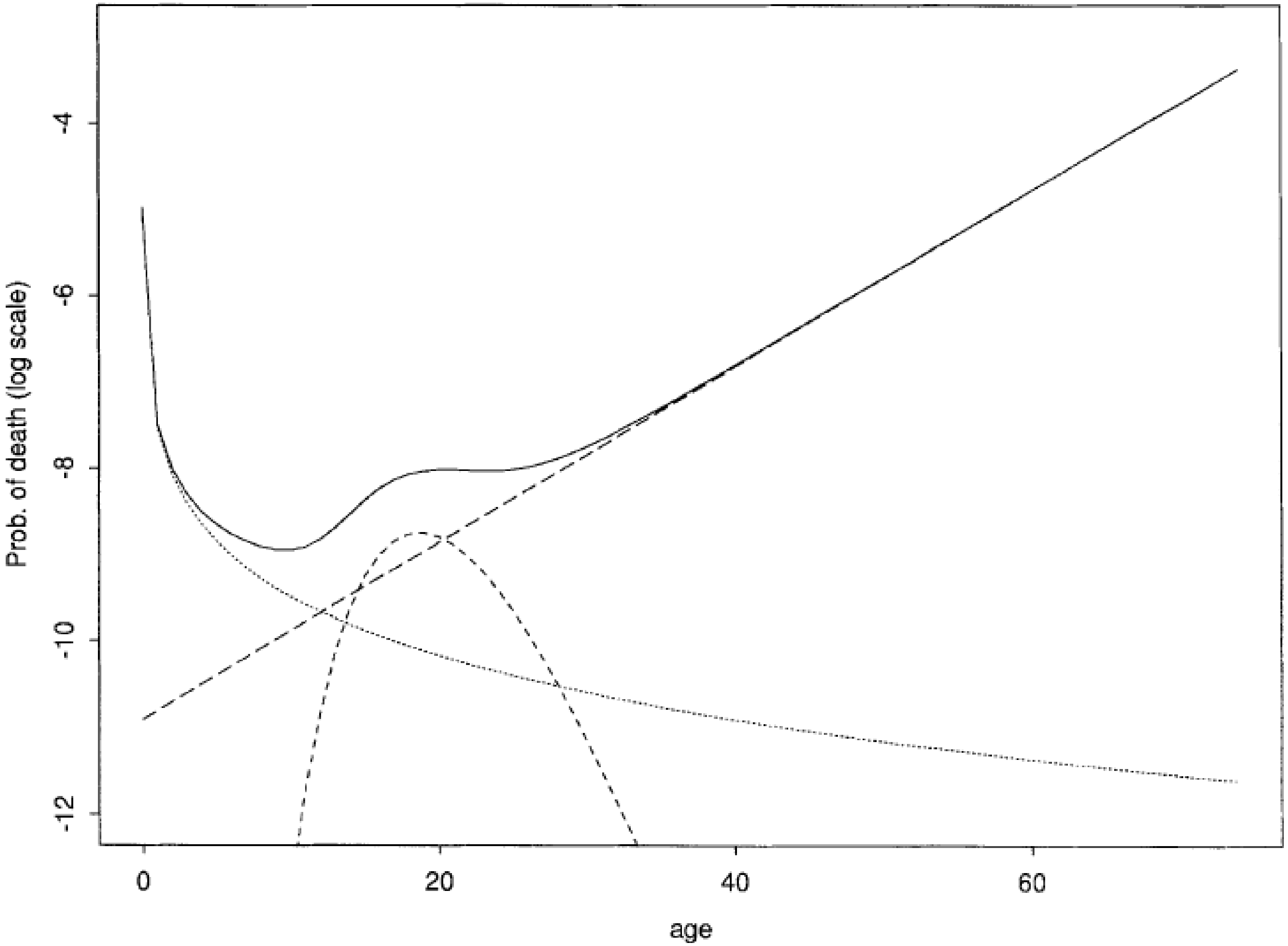
→ Loosely speaking, a spline is a piecewise polynomial function satisfying certain smoothness at the joint points. Consider a set of points, also named the set of knots, $K = \{x_1, x_2, \ldots, x_m\}$ with $x_1 < x_2 < \cdots < x_m$.

→ Piecewise-polynomial representations:

$$s(x) = \begin{cases} p_0(x) = p(x) & x < x_1 \\ p_1(x) & x_1 \le x < x_2 \\ \cdots & \cdots \\ p_m(x) & x_{m-1} \le x < x_m \\ p_{m+1}(x) & x_m \le x \end{cases}$$



Q: Is it possible to use a polynomial to do the job?



Heligman-Pollard 模型中各年齡組的死亡曲線

(A,B,C,D,E,F,G,H) = (.000544,.0170,.101,.000158.10.72,18.67,.0000183,1.11)

- 若將原始觀察值分成k+1個區間，其中k_1, k_2, …, k_k為這 k+1個區段內多項式的交接點；令每個區段內的多項式為三次，通常會假設：

$$\begin{cases} p_i(k_i) = p_{i+1}(k_i), \\ p_i^{'}(k_i) = p_{i+1}^{'}(k_i), \quad i = 1,2,\ldots,k \\ p_i^{''}(k_i) = p_{i+1}^{''}(k_i), \end{cases}$$

其中 p_i 為第i個區段內的多項式，$p_i{}'$ 及 $p_i{}''$ 為 p_i 的一次及二次微分。

- 引用數值分析的方法，讀者可找出滿足上述要求的多項式應具有以下的特性：

$$\begin{aligned} P(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \\ + I(x \geq k_1)\beta_4(x - k_1)^3 \\ + I(x \geq k_2)\beta_5(x - k_2)^3 \\ + \cdots\cdots \end{aligned}$$

→這個多項式可由一般的迴歸分析，即最小平方法或是加權最小平方法求出各β參數的估計值。

- 以矩陣表達，可表示為 $\underset{\sim}{t} = P = A\underset{\sim}{\phi}$，其中：

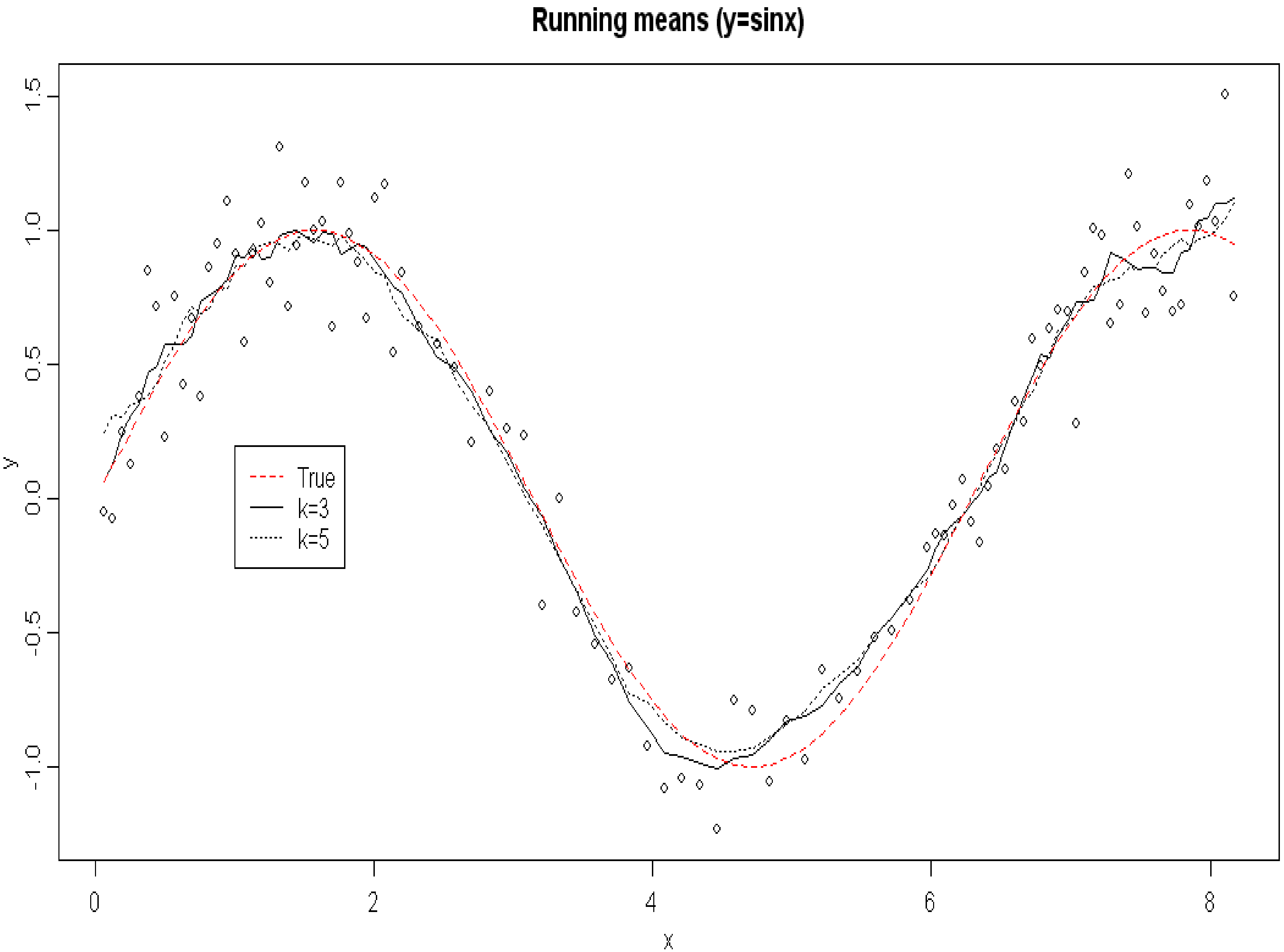
$$A = \left(\begin{array}{cccc} 1 & 1 & 1^2 & 1^3 \\ 1 & 2 & 2^2 & 2^3 \\ 1 & 3 & 3^2 & 3^3 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & n & n^2 & n^3 \end{array} \left(\begin{array}{c} \left. \begin{array}{c} 0 \\ 0 \\ \cdot \\ 0 \end{array} \right\} k_1 \\ 1^3 \\ \cdot \\ \cdot \\ \cdot \\ (n-k_1)^3 \end{array} \right) \cdots\cdots \right), \qquad \underset{\sim}{\phi} = \left(\begin{array}{c} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_{k+2} \end{array} \right)$$

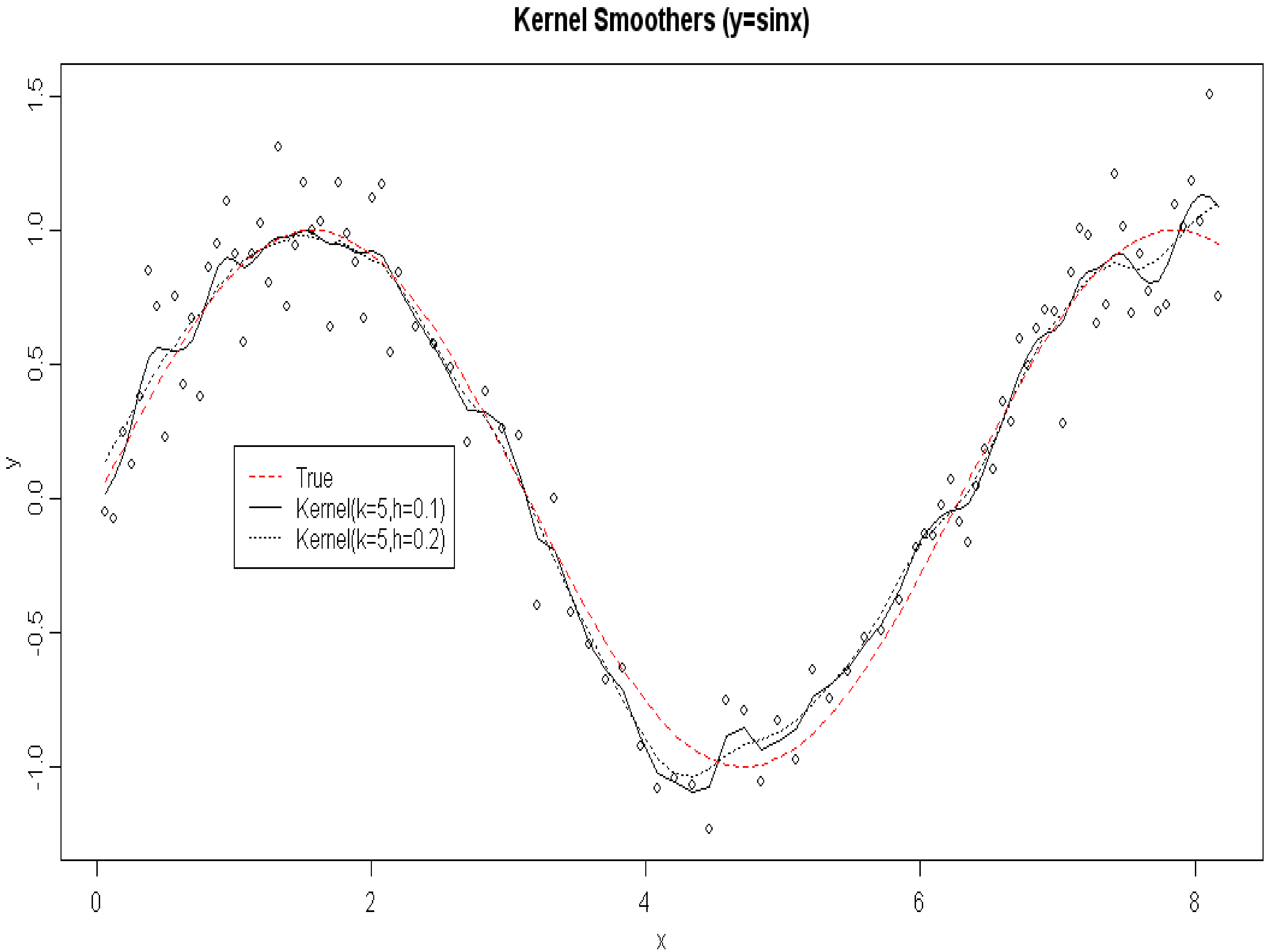
→修勻值可表為 $\underset{\sim}{v} = A(A'A)^{-1}A'\underset{\sim}{u}$。

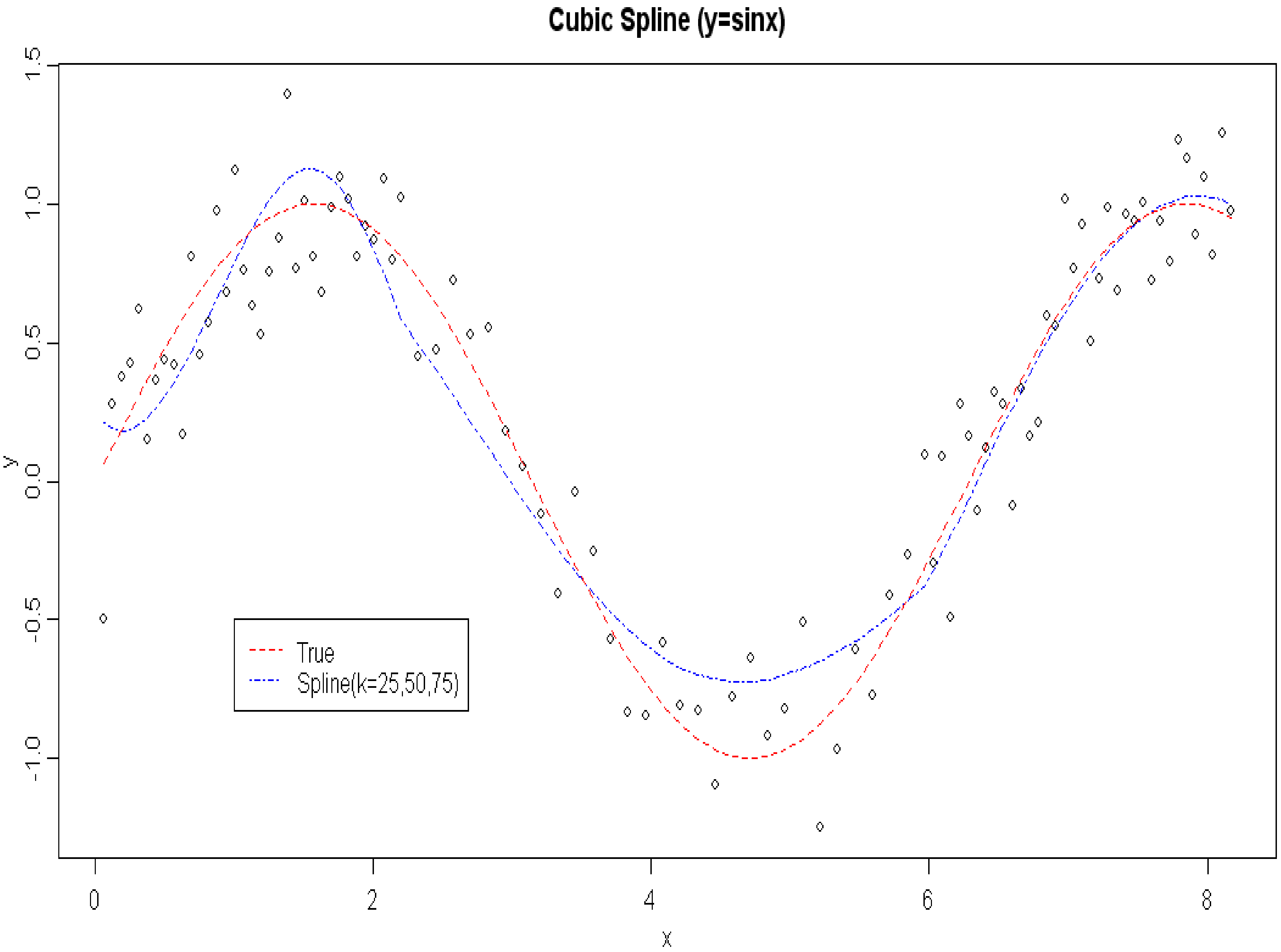
- Example 2. (continued)

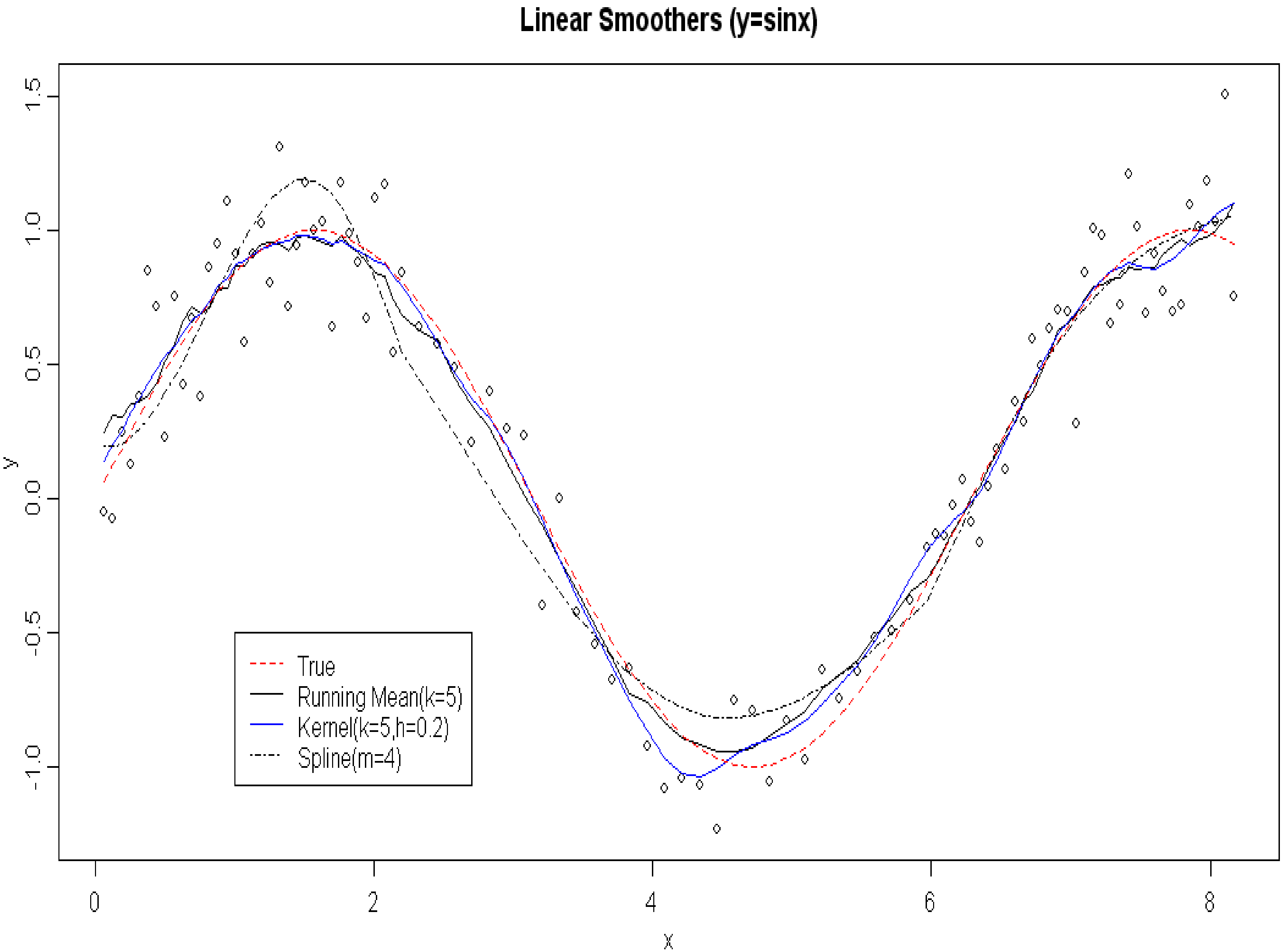
→ We shall use cubic splines with knots at $\{0, 2\pi/3, 4\pi/3, 2\pi\}$ and compare the results of smoothing for different methods.

Note: There are also other smoothing methods available, such as LOWESS (LOESS for an updated version) and running median (i.e., nonlinear smoothers), but we won’t cover these topics in this class.

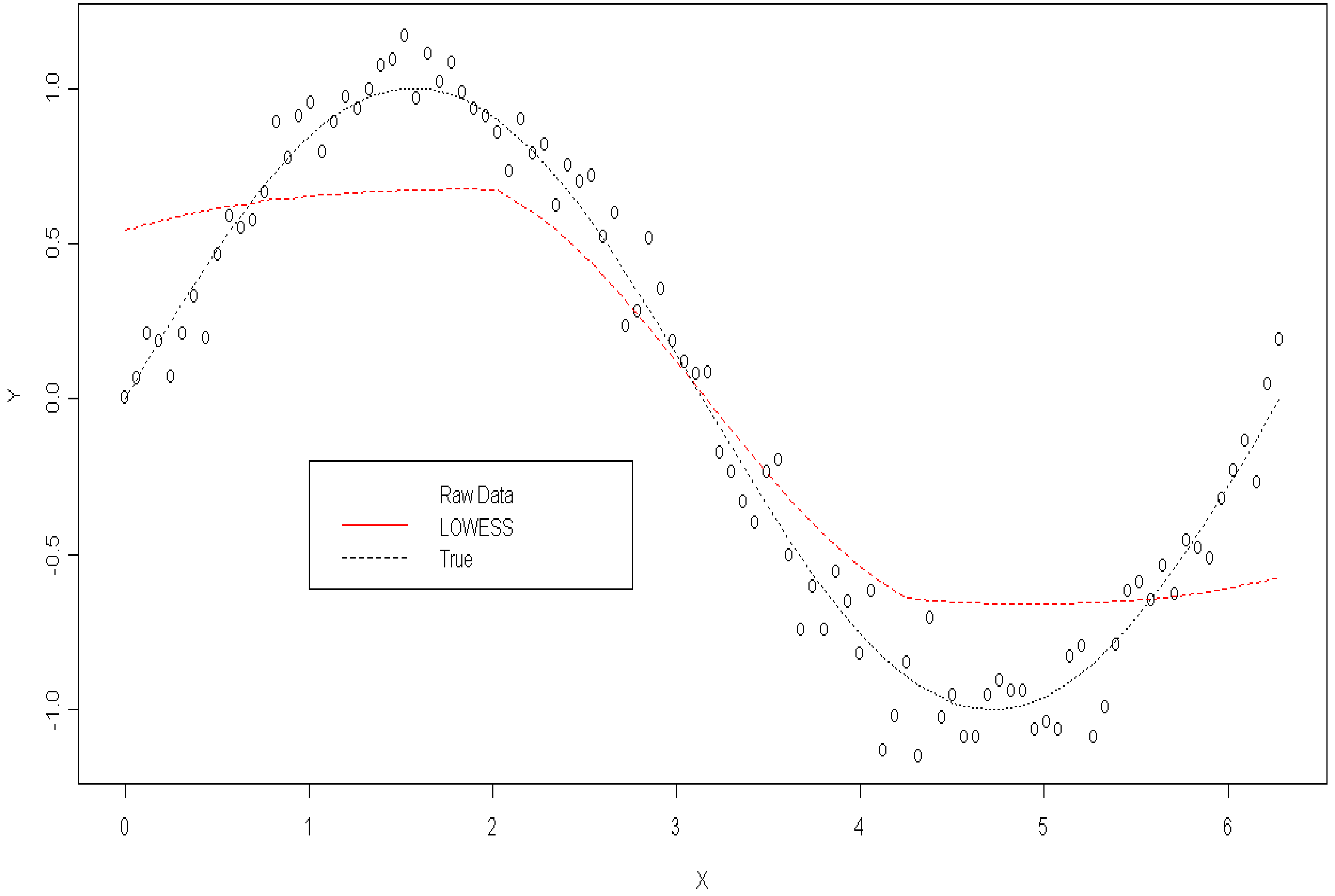

Running means (y=sinx)
True
k=3
k=5
y
x
1.5
1.0
0.5
0.0
-0.5
-1.0
0
2
4
6
8


Kernel Smoothers (y=sinx)
True
Kernel(k=5,h=0.1)
Kernel(k=5,h=0.2)
y
-1.0
-0.5
0.0
0.5
1.0
1.5
x
0
2
4
6
8

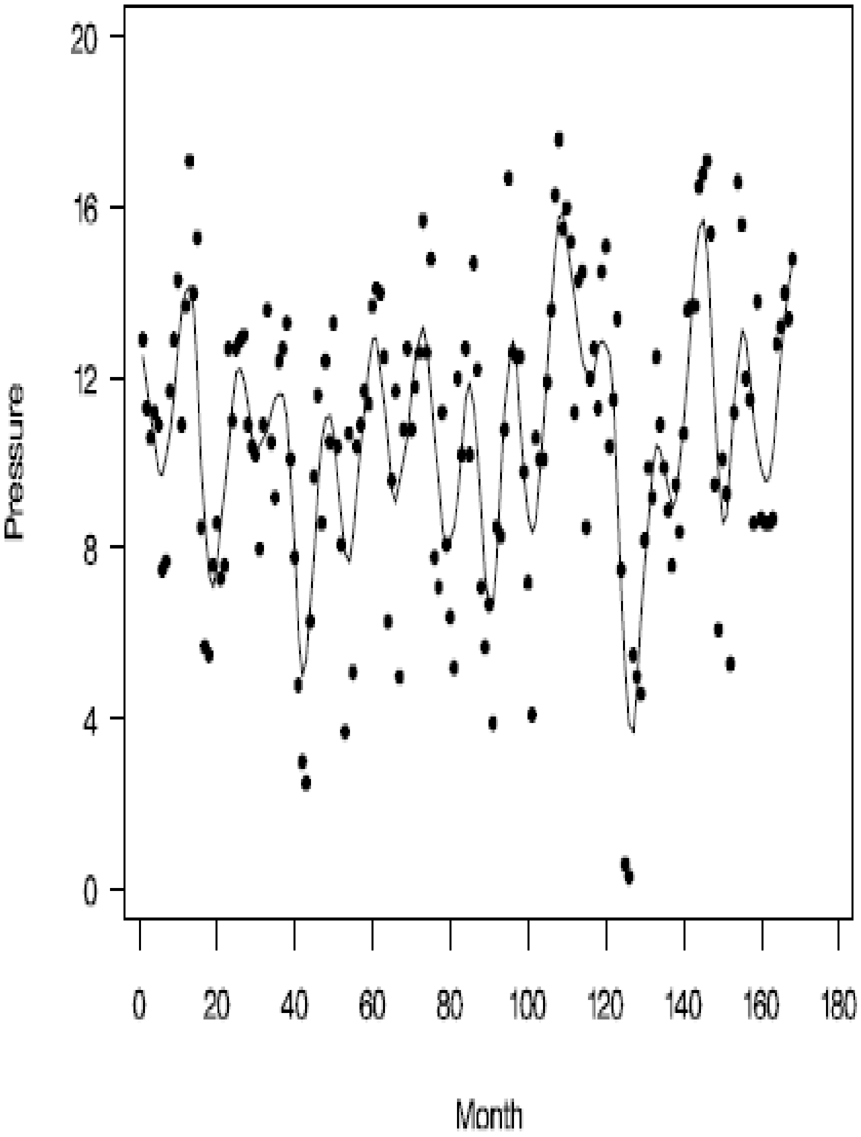

Cubic Spline (y=sinx)
1.5
1.0
0.5
0.0
-0.5
-1.0
y
0
2
4
6
8
x
True
Spline(k=25,50,75)

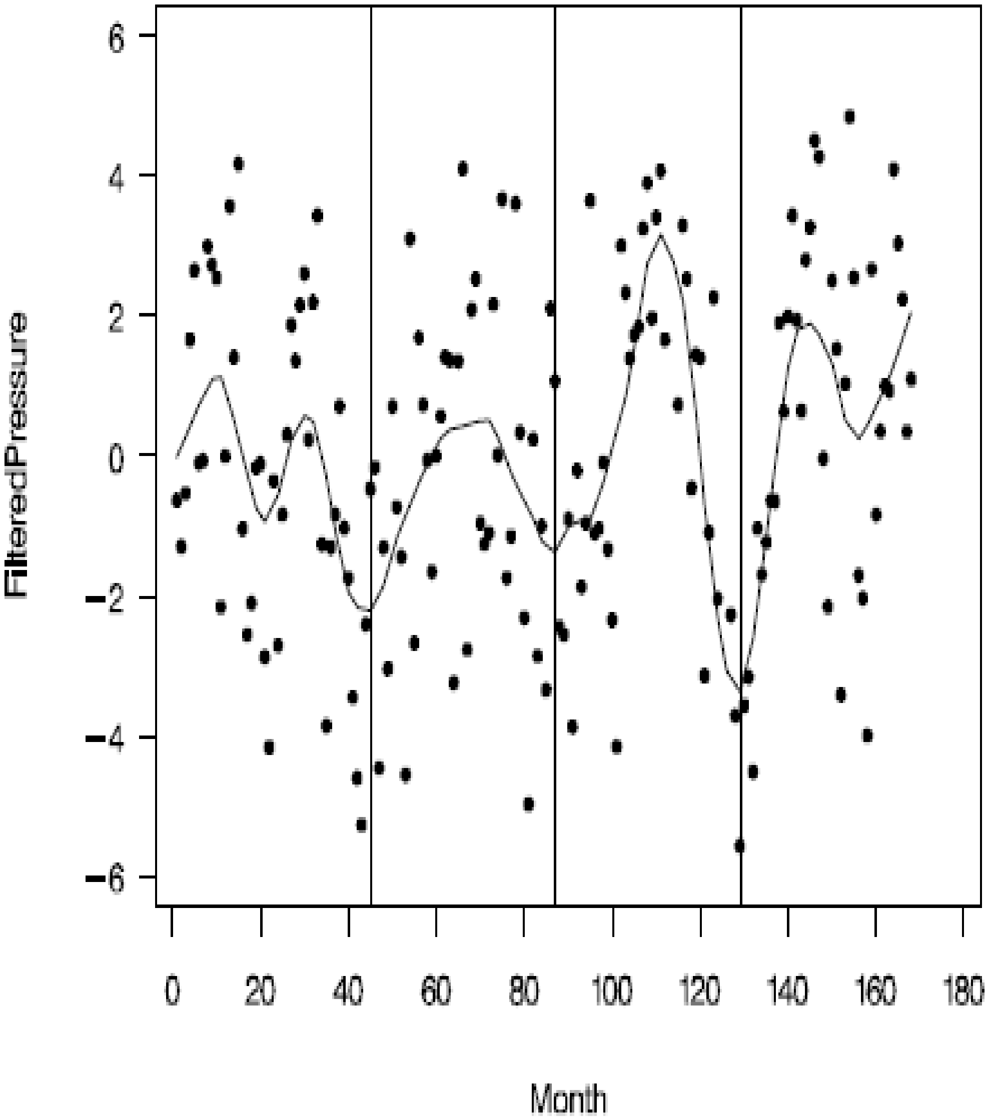

Linear Smoothers (y=sinx)
1.5
1.0
0.5
0.0
-0.5
-1.0
y
0
2
4
6
8
x
True
Running Mean(k=5)
Kernel(k=5,h=0.2)
Spline(m=4)

LOWESS

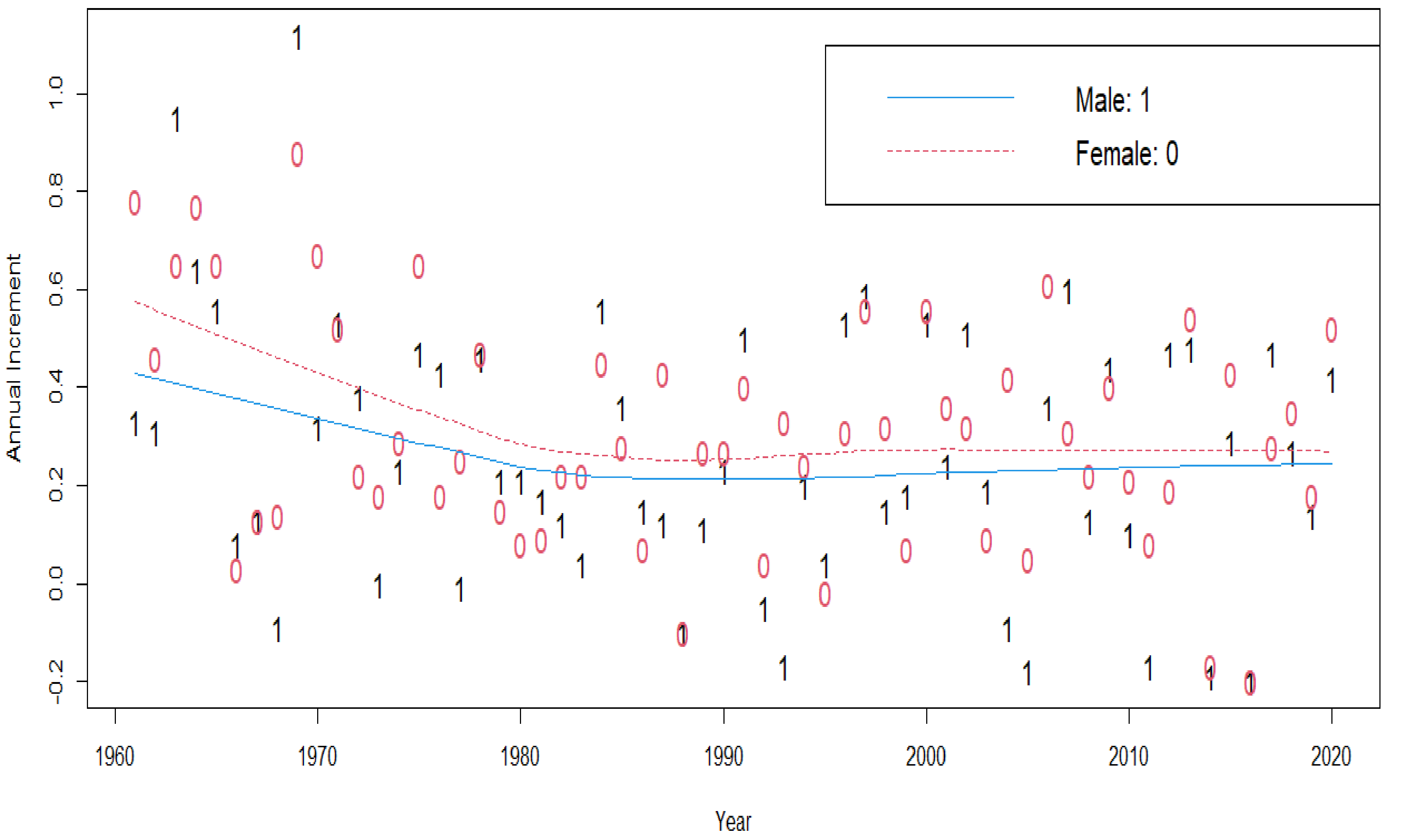


LOWESS (Locally-weighted Polynomial Regression) or *LOESS* (Local Polynomial Regression Fitting)





LOWESS分析範例(每年壽命增加幅度)



懲罰概似估計

- 懲罰概似估計(Penalized Likelihood Estimation)

→Smoothing Spline是使下列函數最小化

$$S_{\lambda}(M) = \frac{1}{n}\sum_{i=1}^{n}[y_i - M(x_i)]^2 + \lambda\int_a^b [M''(t)]^2 dt,$$

實證上目標函數設為

$$L(\theta \mid data) + \frac{\lambda}{2} J(\theta)$$

其中L(θ|data)為對數概似函數取負號，J(θ)二次粗略的懲罰函數(Quadratic Roughness Penalty)。

懲罰概似估計(續)

- 懲罰函數J(θ)一般選為

$$J(\theta)=\int\ddot{\theta}(x)^2\,dx \quad 或 \quad J(\theta)=\sum_{x}(\Delta^z\theta(x))^2$$

→若觀察值滿足 $Y_i=\theta(x_i)+\varepsilon_i,\ \varepsilon_i\sim N(0,\sigma_i^2)$，上述範例一可視為這種模型的特例，則PLE修勻即是將下列目標函數最小化：

$$\sum_{i=1}^{n}\left(\frac{Y_i-\theta(x_i)}{\sigma_i}\right)^2+\frac{\lambda}{2}\int\ddot{\theta}(x)^2\,dx$$

Note: Two terms of the right-hand side of the objective function usually represent constraints opposite to each other.

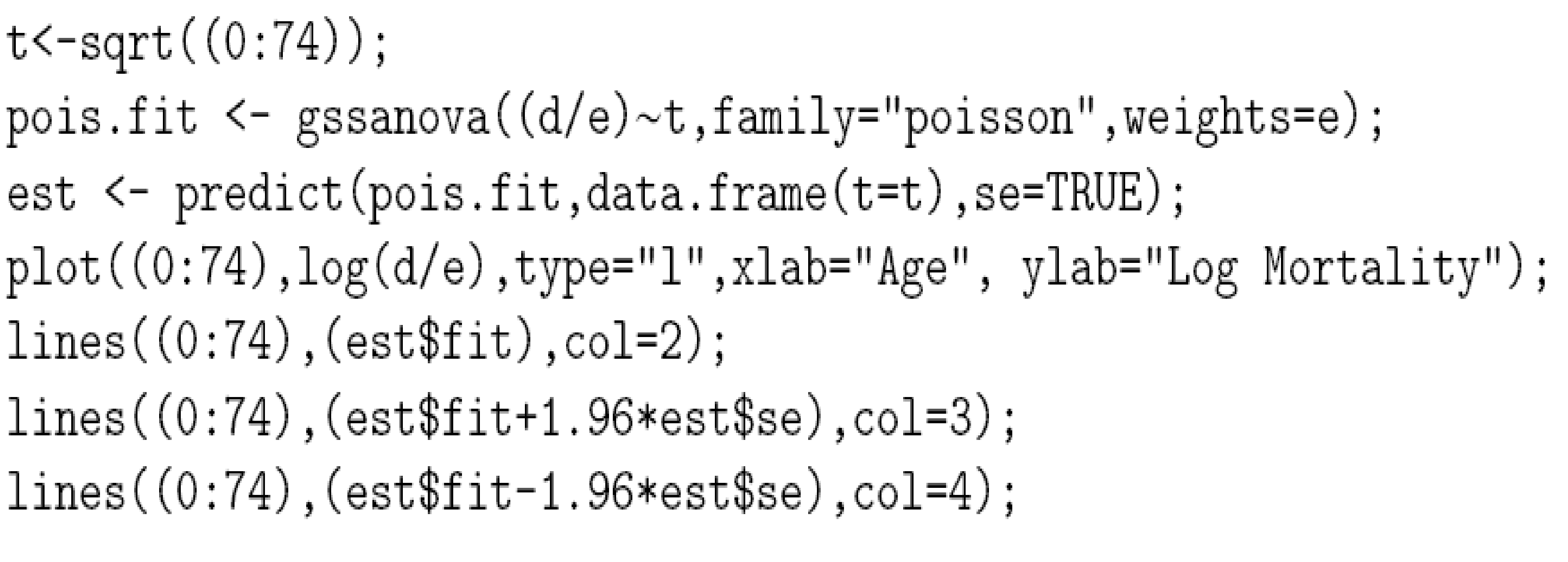
→ The first term measures how far the smoothers differ from the original observations.

→ The second term, also known as *roughness penalty*, measures the smoothness of the smoothers.

Note: Methods which minimize the objective function are called *penalized LS methods*.

Example 3、1988~2002年England 及Welsh 的女性死亡資料。

→使用 R 軟體的套裝程式「gss」模組，操作手冊可從www.r-project.org下載，指令非常簡單，程式如下，圖形在下一頁。



```
t<-sqrt((0:74));
pois.fit <- gssanova((d/e)~t,family="poisson",weights=e);
est <- predict(pois.fit,data.frame(t=t),se=TRUE);
plot((0:74),log(d/e),type="l",xlab="Age", ylab="Log Mortality");
lines((0:74),(est$fit),col=2);
lines((0:74),(est$fit+1.96*est$se),col=3);
lines((0:74),(est$fit-1.96*est$se),col=4);
```

English and Welsh Mortality Data

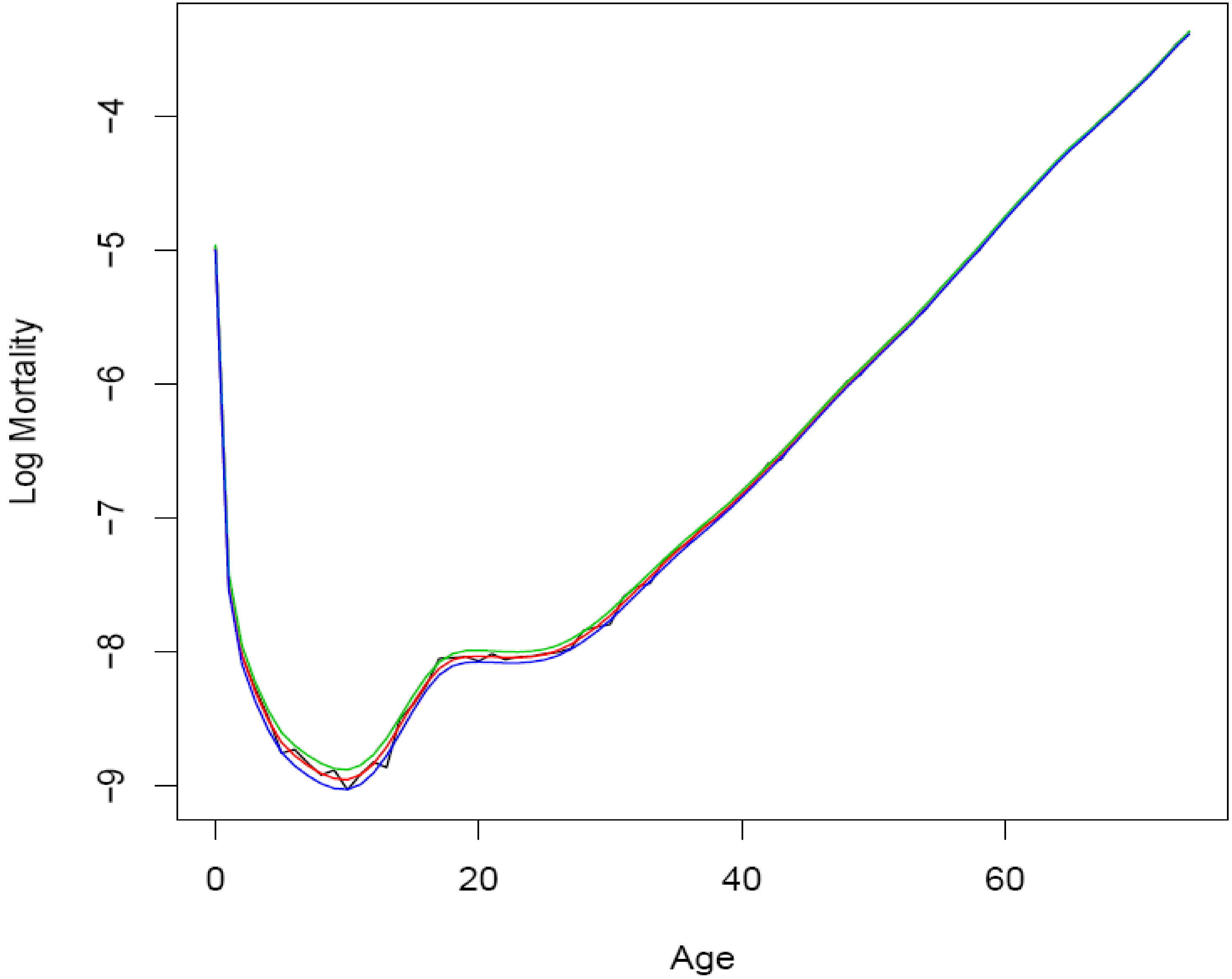


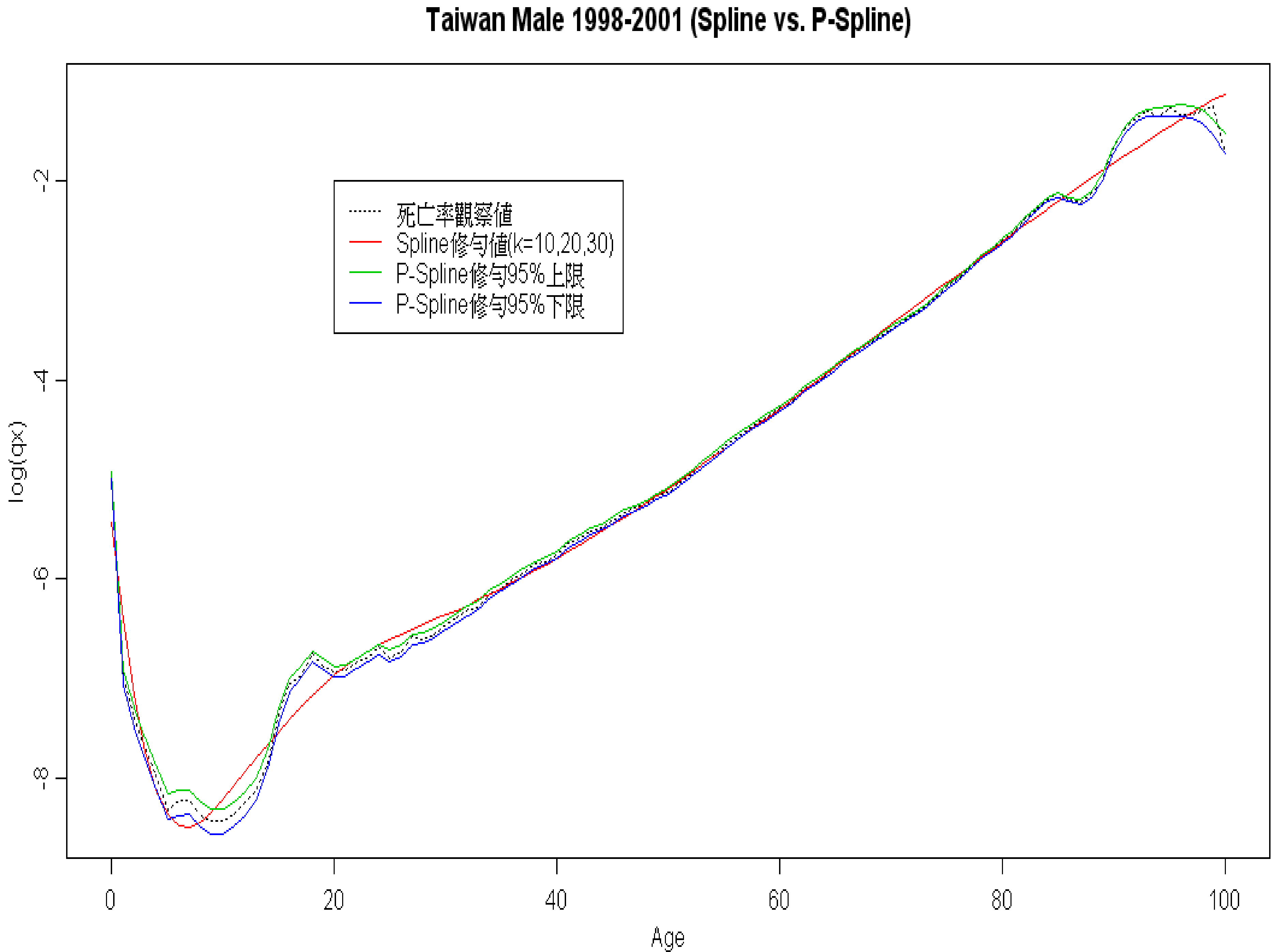
Figure 1 Raw Data (Black), Upper 95%, Lower 95% and Bayes Estimate.

- Example 4.Taiwan 1998-2001 Male Life Tables (Data can be downloaded from https://www.moi.gov.tw/cl.aspx?n=4404)

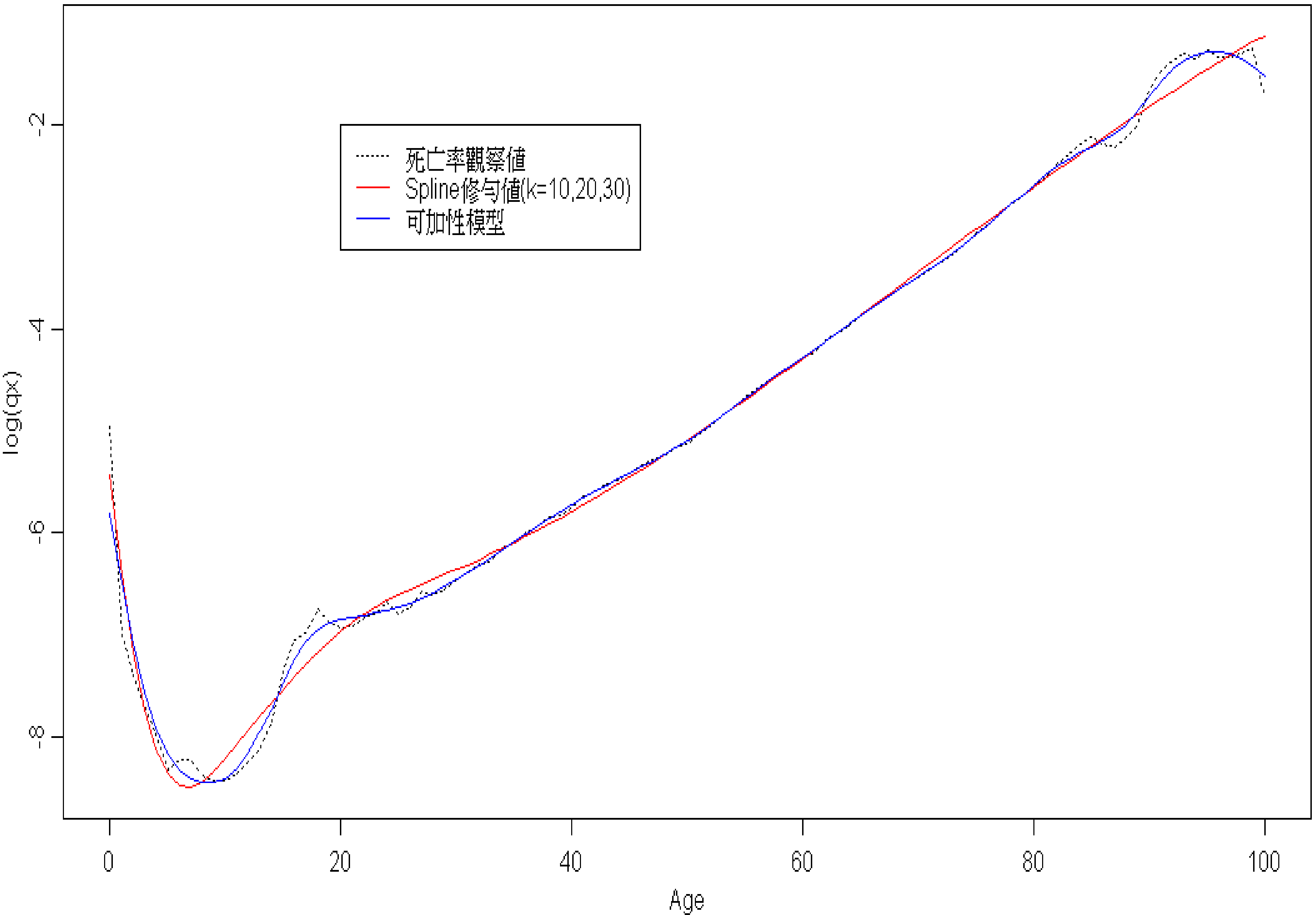
- First, we combine the numbers of deaths and population from 1998 to 2001, and then compute the age-specific mortality rates

→There are 3 knots in Spline smoothing

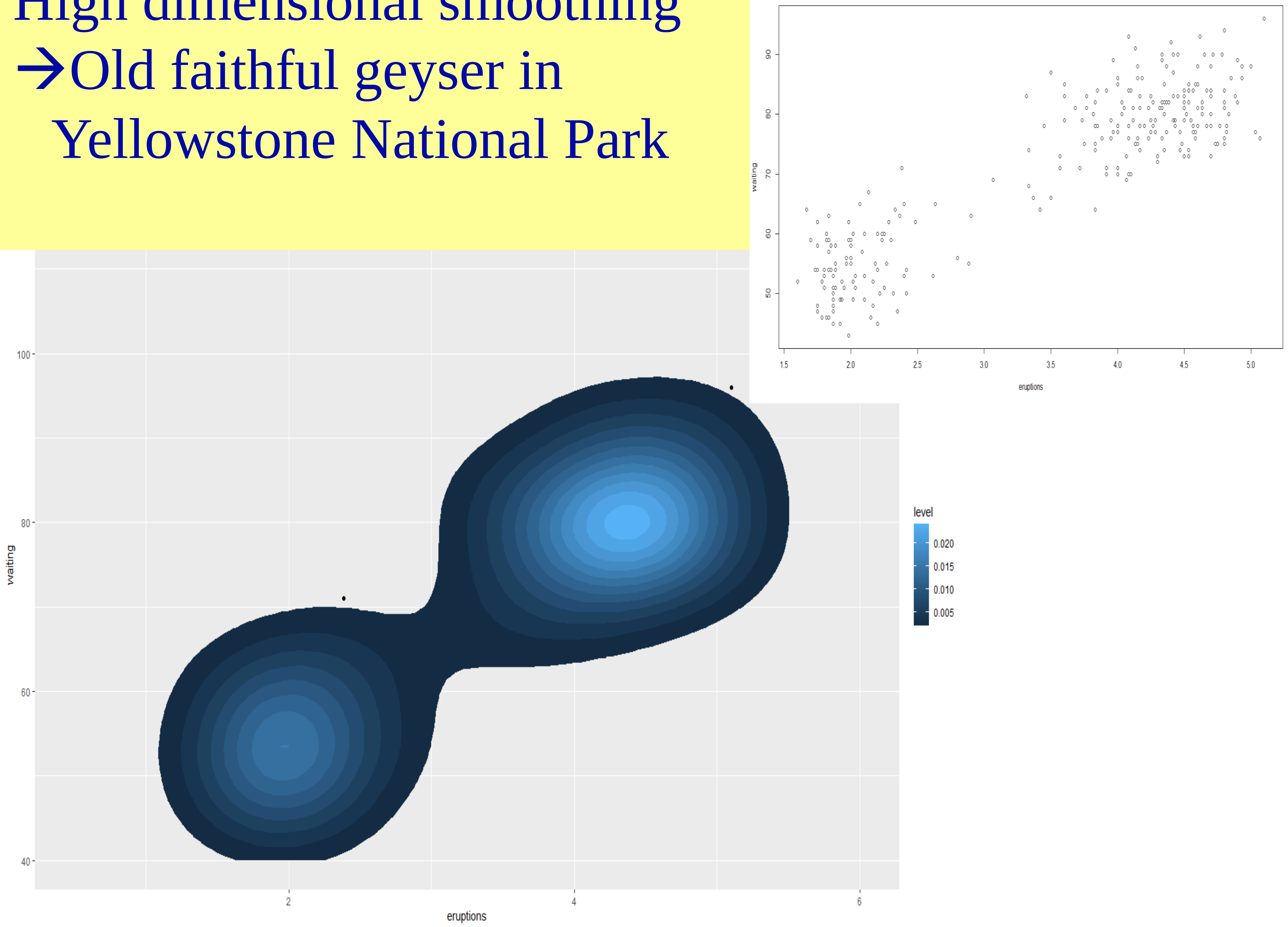
→We can also use "Generalized Additive Model"(「gam」 in R).


Taiwan Male 1998-2001 (Spline vs. P-Spline)
死亡率觀察值
Spline修勻值(k=10,20,30)
P-Spline修勻95%上限
P-Spline修勻95%下限
log(qx)
-2
-4
-6
-8
0
20
40
60
80
100
Age

Taiwan Male 1998-2001 (Spline vs. GAM)



High dimensional smoothing →Old faithful geyser in Yellowstone National Park



2d density plot with ggplot2

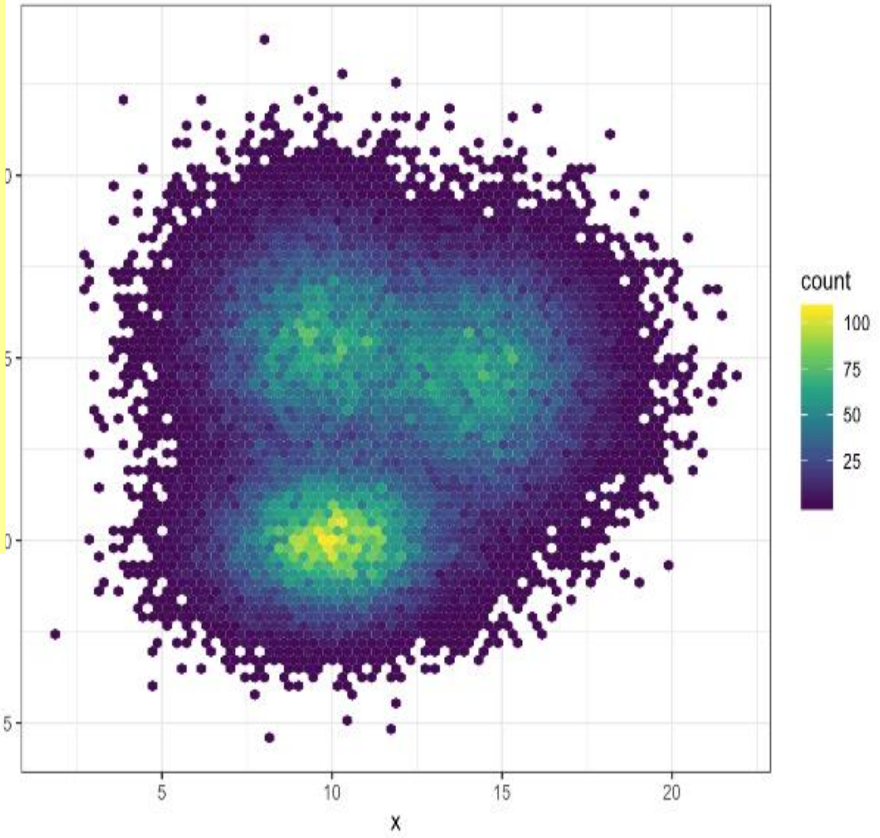
→Try the following commands

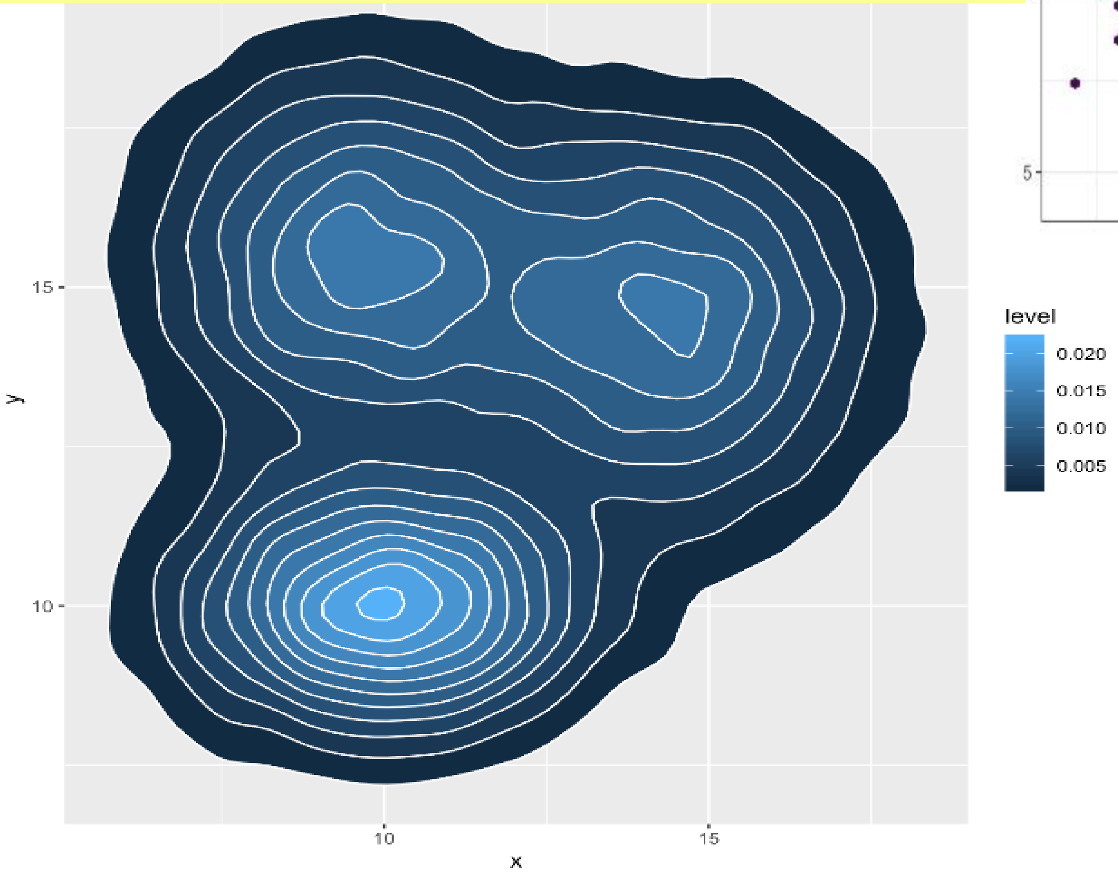
2d distribution with

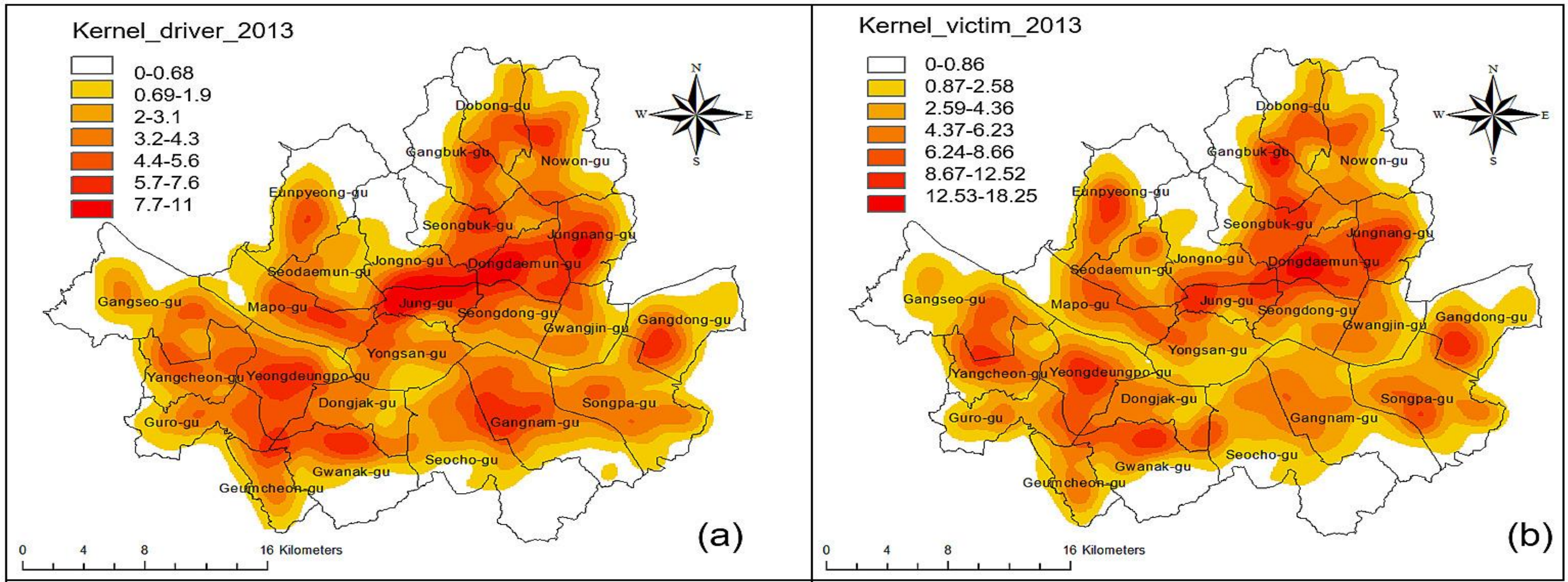
geom_density_2d or

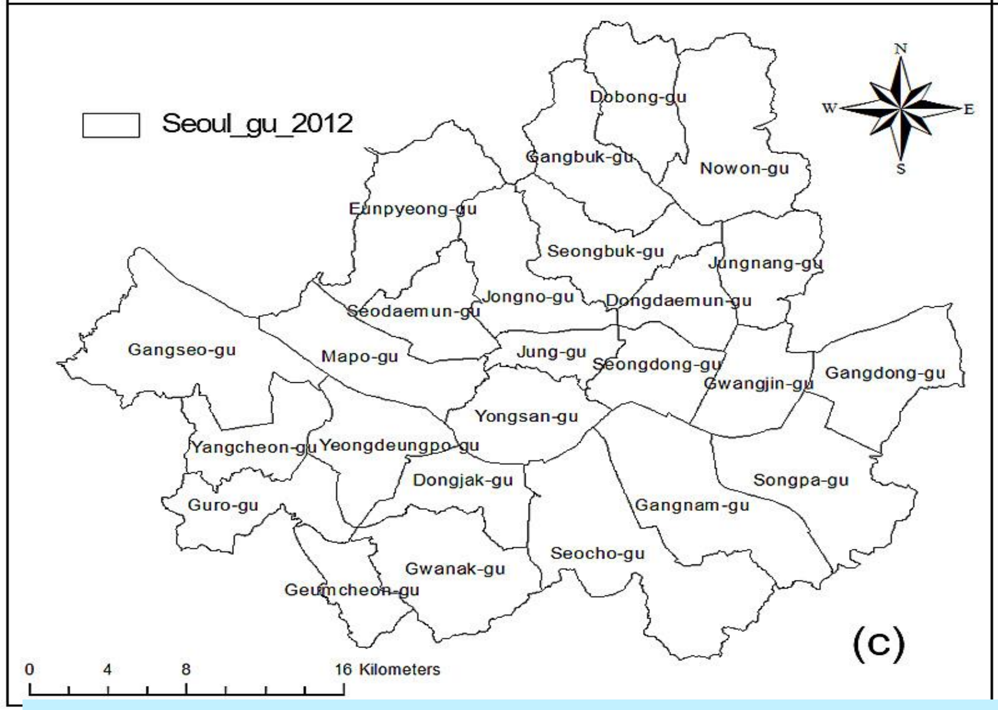
stat_density_2d;

2d Histogram with ***geom_bin2d()***





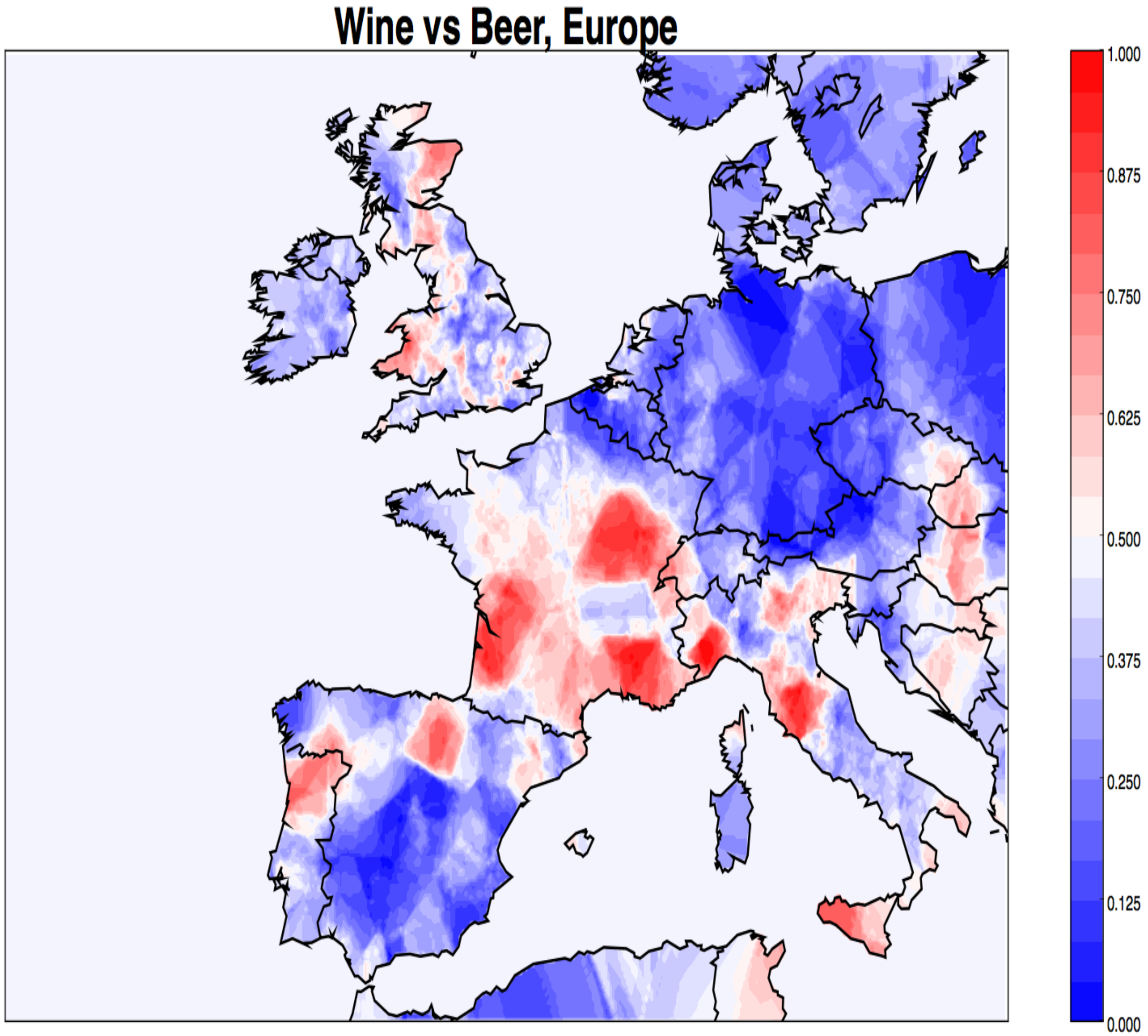




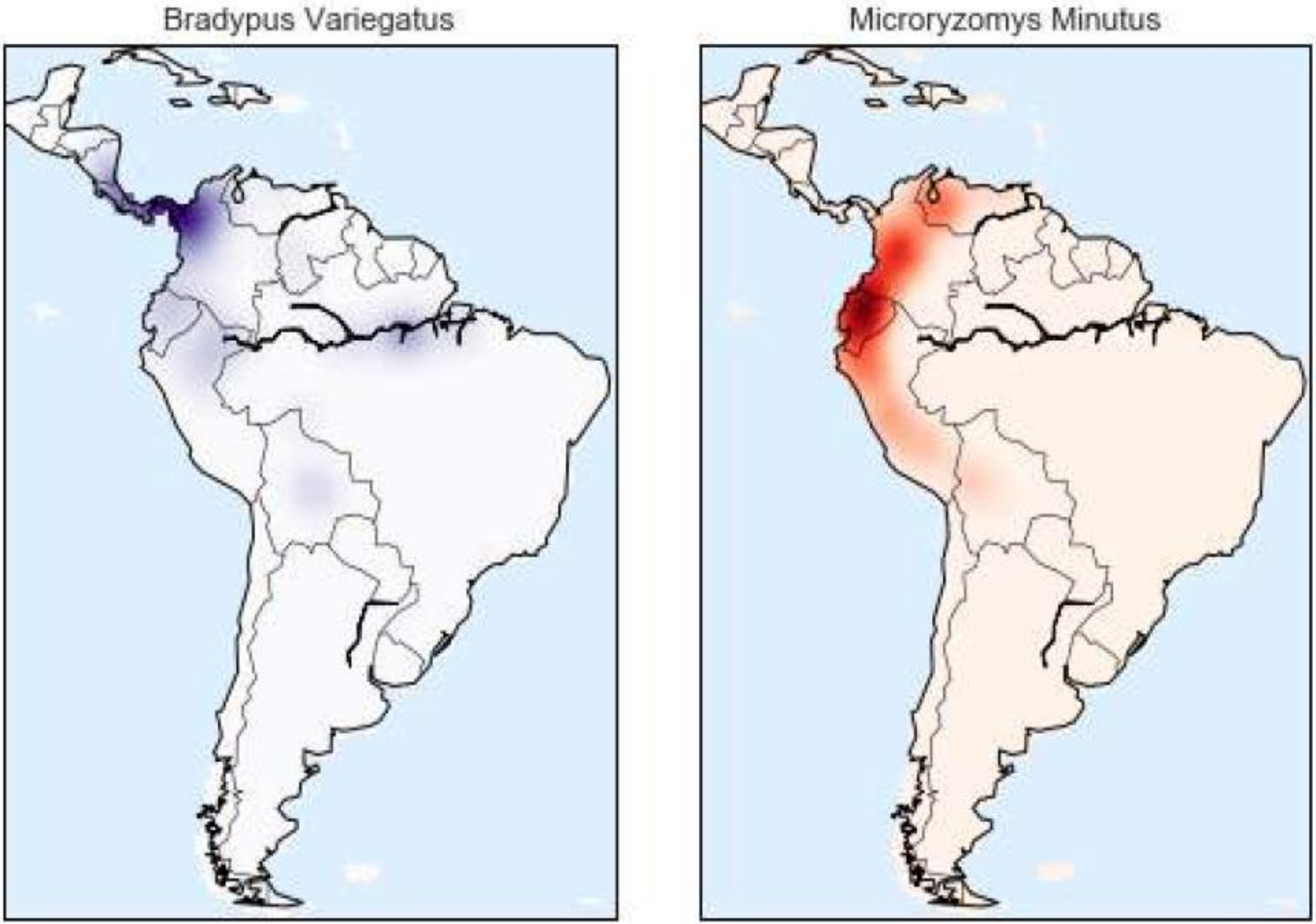
Concentrated regions of elderly people’s traffic accidents (首爾)

杨州市
楊州市
议政府市
議政府市
1. 东大门区
2. 衿川区
1. 東大門區
2. 衿川區
道峰区
道峰區
芦原区
蘆原區
高阳市
高陽市
南杨州市
南楊州市
江北区
江北區
恩平区
恩平區
城北区
城北區
中浪区
中浪區
金浦市
金浦市
西大门区
西大門區
钟路区
鍾路區
1
九里市
九里市
桂阳区
桂陽區
江西区
江西區
麻浦区
麻浦區
中区
中區
城东区
城東區
广津区
廣津區
江东区
江東區
龙山区
龍山區
永登浦区
永登浦區
阳川区
陽川區
河南市
河南市
松坡区
松坡區
铜雀区
銅雀區
富川市
富川市
九老区
九老區
瑞草区
瑞草區
江南区
江南區
冠岳区
冠岳區
2
光明市
光明市
果川市
果川市
城南市
城南市
安养市
安養市

https://ndownloader.figshare.com/files/11485925/preview/11485925/preview.jpg

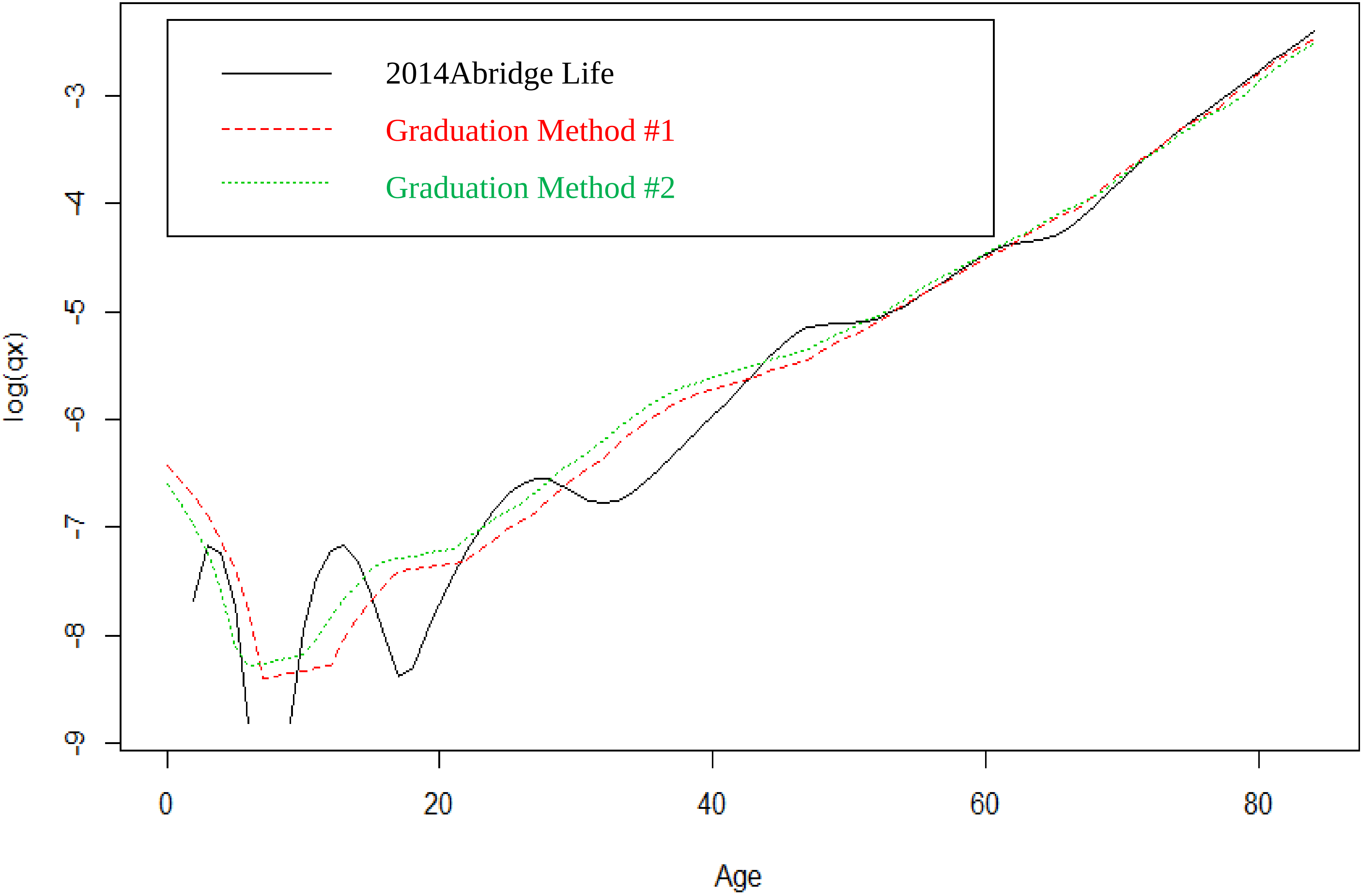


https://raw.githubusercontent.com/epierson9/densityMapping/master/wine_vs_beer.png

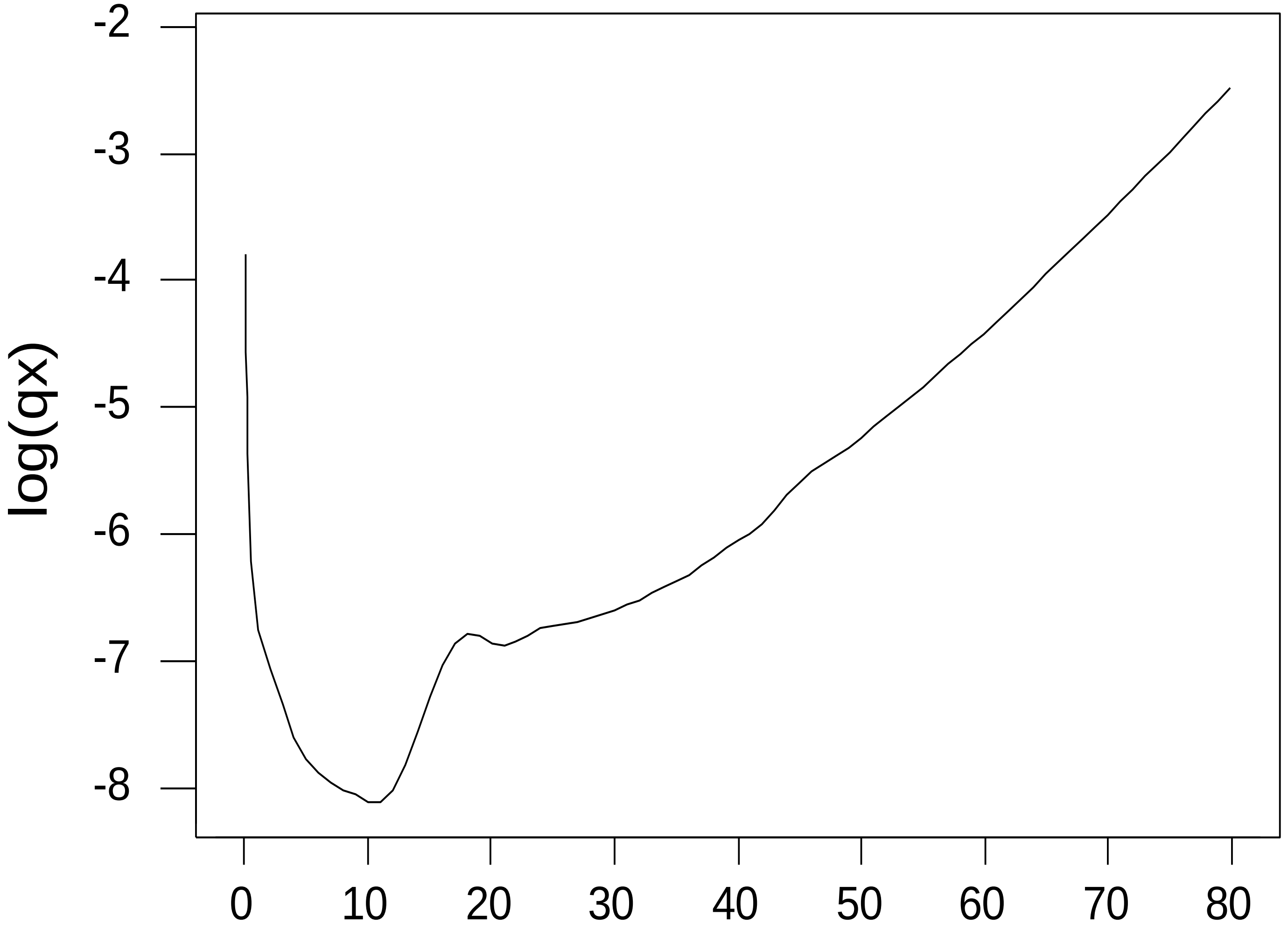


Species Distribution (Kernel Density Estimation)

Small Area Life Table (2014 Penghu Male)



An Example of Taiwan Life Table



Age