The answer is each interval must contain enough data points to allow estimation & maximization & they suggest a approximate quantile method.

In practice - use software & add & delete knots is to try to improve the estimation. Other strategies exist as well.

Density Estimation

Often in statistics, the function that we wish to estimate is a probability density. i.e. we are given a sample $X_1, \ldots, X_n$ of iid observations from unknown density f on D and and we construct $\hat{f}$ with

- $\hat{f}(x) \geq 0 \quad \forall \; x \in D$
- $\int_D \hat{f}(x)\,dx = 1$

hoping to find $\hat{f}$ with

- small error (ex M.S.E).
- $E[\hat{f}_n(x)] \to f(x) \quad \forall \; x \in D$ as $n \to \infty$.

If it is believed that f is a parametric density $f_{X|\theta}$ there are a variety of techniques to est. f.

- MLE
- MOM
- Fitting by matching quantiles.
- mixtures.
- logspline

So we assume NOT.

Nonparametric Density Estimation (Chpt. 10).

- We will discuss methods to estimate f when very little is known about its form. We will focus on 3 common methods.
  1. Orthoganal Series Estimators
  2. Histogram Estimators
  3. Kernel Estimatos.

I. Orthog. series est.

- we have already covered these.

$$\hat{f}(x) = \frac{1}{n} \sum_{k=0}^{J} \sum_{i=1}^{n} g_k(y_i)\, g_k(y)$$

where g_i is an orthog. series.

Comments:

not in book

1. # of terms has a major effect & more is not necess. better
2. $\hat{f}$ may not be smooth & it may have infinite variance.
3. Convergence rate (to f) is ind of dim $\therefore$ may be a good cand. for multivariate problems.
4. Most commonly Fourier & Hermite series used.

II Histogram Estimators.

not in book

A histogram is a piecewise constant density estimator. (Need not be univariate). What is $\hat{f}$? Consider how we construct the histogram

- Assume the support D is finite
- Construct a fixed partition of D using m nonoverlapping bins B_k, ie

$$B_j \cap B_k = \emptyset \;\; \& \;\; D = \bigcup_{k=1}^{m} B_k$$

- Let V_K be the volume of B_K. In one-dim this is simply the length, h_K, of the bin B_K, which is a subinterval of D. & often all equal.

- Let n_K be the # of obs in B_K

$$n_K = \sum_{i=1}^{n} I\,(X_i \in B_K)$$

- The proportions of obs in B_K is

$$\hat{P}_K = \frac{n_K}{n}$$

- The probability content of the bin is

$$P_K = \int_{B_K} f(u)\,du \qquad \text{(of course } f \text{ is often unknown).}$$

- The histogram estimator of f is

$$\hat{f}_n(x) = \begin{cases} \hat{P}_1/V_1 & x \in B_1 \\ \hat{P}_2/V_2 & x \in B_2 \\ \vdots & \\ \hat{P}_m/V_m & x \in B_m \end{cases}$$

or $$\hat{f}_n(x) = \sum_{K=1}^{m} \frac{\hat{P}_K}{V_K} I\,(x \in B_K) = \sum_{K=1}^{m} \frac{n_K}{nV_K} I\,(x \in B_K)$$

Comments: ① $\hat{f}(x) \geq 0 \quad \forall x \in D$

② $\int_D \hat{f}(x)\,dx = \sum_{K=1}^{m} \frac{n_K}{nV_K} \cdot V_K = 1$

③ $E[\hat{f}_n(x)] = P_K/V_K$ for $x \in B_K$

④ $Var(\hat{f}(x)) = \frac{P_K(1-P_K)}{n\,V_K^2}$ for $x \in B_K$

easy to see since n_K is a binomial r.v.

Var ↓ as bin size ↑. Variance differs from bin to bin.

⑤ Under certain cond. (Lipschitz continuity). you can bound Bias, var, & MSE (& others).

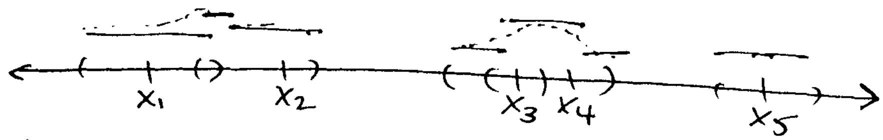
(6) In the multivariate case, can easily see extension of bins as boxes, but other shapes (triangles, squares, hexagons, & more work). and so choice of size & # of bins becomes the problem.

III Kernel Estimators. (Chapter 10).

Begin w/ univariate

Motivation: Let's think about how a prob. density function assigns prob. to intervals.

- If we observe $X_i = x_i$ we assume that f assigns density not only at x_i, but in a region around x_i (f is smooth).
- $\therefore$ to estimate f from $X_1, \dots, X_n \sim_{iid} f$ it makes sense to accumulate contrib to the regions.



Formally,
to estimate the density at point x, we consider the region $dx = 2h$ (h is some fixed #) centered at x. Then the prop of obs. that fall in the interval

$$\gamma = [x-h, x+h]$$

gives an indication of the density at x. and we have

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} I\{|x - x_i| < h\} \qquad (*)$$

as our estimator.

It can be shown (see discussion on pg 278) in order for $\hat{f}$ to be reasonable & pointwise consistent,

$$nh \to \infty \text{ and } h \to 0 \text{ as } n \to \infty.$$

This estimator is an example of a Kernel estimator with uniform kernel.

(*) weights all points w/in h of x equally.
Suppose we allow a more flexible weighting scheme, ie.

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

where K is a kernel function & h is constant the bandwidth.

Comments: (1) The kernel function weights the contributions given by each obs. X_i to the density est. $\hat{f}(x)$ based on proximity of X_i to x.

(2) K is usually pos. everywhere & sym. about 0.

(3) Often a density, such as normal, student's t. but there are others that are popular. (See pg. 292. Table 10.1).

(4) Kernel in (*) is uniform.

Example pg 280 Fig 10.1.
- 4 points, Kernel normal density.

When constructing a Kernel density estimator you need to choose 2 things
(1) the Kernel K
(2) the bandwidth h.

Turns out (2) is of much greater importance than (1).

Sec 10.2.1 <u>Choice of Bandwidth.</u>

The bandwidth h strongly influences $\hat{f}(x)$.
- h too small - density assigned to locally, & $\hat{f}$ is very wiggly w/ many false modes.

- causing high variance

- h too big → density spread out too much & features are lost; causing higher bias.

Example Fig 10.2 pg 282

--- $h = 1.875$ too big oversmoothing
— $h = .3$ too small too much variability
▬ $h = .625$ best.

Q: How to choose h?

- Often - in practice - if density est. is used for exploratory data analysis - visual [redacted] (trial & error) is "defensible". Since there is no correct choice you have some room 10-20%.

- Some software has automatic bandwidth selection based on various approaches.

- A little more formally: Let's consider how the bandwidth affects the MISE (= IMSE)

Recall: $MISE(\hat{f}) = IV(\hat{f}) + ISB(\hat{f})$ often main error measures want to minimize this.
and clearly is affected by bandwidth.

MISE = AMISE + error

If lt $h \to 0$, $nh \to \infty$ and $n \to \infty$

It can be shown*. (See derivation pg 283-284) That the Asymptotic mean integ. squared error. AMISE is minimized when

$$h = \left(\frac{R(K)}{n \sigma_K^4 R(f'')} \right)^{1/5}$$

*if K is a symm. cont prob density w/ mean 0 & var $0 < \sigma_K^2 < \infty$

and MISE → 0

Where $R(g) = \int g^2(z)\,dz$ is a measure of the roughness of g.

This is the exact balancing of the orders of the bias & variance terms.

How helpful is this?
- We don't know f.
- Does tell us that optimal $h = O(n^{-1/5})$ & MISE $= O(n^{-4/5})$.
- Employ one of several b.w. selection strategies.

Method 1. Cross-validation.

- Think of $\hat{f}$ as a function of h.
- Want to estimate f while optimizing some quality $Q(h)$. (min error, MSE).
- As we know, using $X_1, \ldots, X_n$ to find $\hat{f}$ and then again to calculate $\hat{Q}(h)$ can cause overfitting.
- So instead ~~to eval~~ $\hat{Q}$ at x_i we use

$$f_{-i}(X_i) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)$$

(X_i omitted from fitting).

Basic idea is the same.

Options for bandwidth selection

① $\hat{Q}(h)$ is the pseudo-likelihood

$$PL(h) = \prod_{i=1}^{n} f_{-i}(X_i)$$

Then choose h to minimize PL.
- Not the best, often to wiggly & sens. to outliers. or not consistent.

Unbiased cross validation UCV

(2) Minimize $ISE(h) = \int \hat{f}^2(x)\,dx - 2E[\hat{f}(x)] + \int f(x)^2\,dx$

$$= R(\hat{f}) - 2E\{\hat{f}(x)\} + R(f).$$

$$\approx R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i) + \cancel{R(f)} \quad \text{constant}$$

and so choose h to minimize

$$UCV(h) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i).$$

Comments:

- $R(\hat{f})$ can be found analytically for some kernels (ex Normal Eq 10.23 pg 286). so use one you know.
- $\hat{h}$ is asyptotically as good as best possible, but convergence can be slow. and has a strong dependence on the observed data.

Ex/ pg 287 Fig 10.4.

Method 2 : Plug-in methods

Basic idea: Apply a pilot b.w. to est one or more important features of f, then estimate better h at a 2nd stage based on initially estimated features.

Recall opt. $h = \left(\dfrac{R(K)}{n\sigma_K^4 R(f'')}\right)^{1/5}$ (Eq BW)

So 1st estimate $R(f'')$ the obtain h.

Replace $R(f'')$ w/ $R(\phi)$

Option 1. Silverman's Rule

- Assume f is normal w/ variance = sample var.
- The solving for h we get

$$h = \left(\frac{4}{3n}\right)^{1/5} \hat{\sigma}.$$

Then use this to obtain $\hat{f}$.

Comments 1 If multimodel, this will oversmooth.

2 sometimes replace $\hat{\sigma}$ w/ Interquartile Range (another measure of the spread.

3 Often used as a method to obtain a pilot h.

Replace $R(f'')$ w $R(\hat{f}'')$

Option 2 Sheather-Jones Method.

- Choose pilot bandwidth h_0 & suff diff kernel L
- Estimate $\hat{f}''$ empirically

$$\hat{f}''(x) = \frac{d^2}{dx^2}\left(\frac{1}{nh_0}\sum_{i=1}^{n} L\left(\frac{x-x_i}{h_0}\right)\right)$$

$$= \frac{1}{nh_0^3}\sum_{i=1}^{n} L''\left(\frac{x-x_i}{h_0}\right)$$

- compute opt. $\hat{h}$ based on $\hat{f}''$.
- estimate $\hat{f}$ using $\hat{h}$.

Comments.

1 The best b.w. for est. $\hat{f}''$ is not the same as that for f. (var$\{f''\}$ playes a greater roll) $\therefore h_0 > h$.

2 This method generally does very well. and is often used.

Example Fig 10.5 pg 291

Method 3: Maximal Smoothing Principle (Terrell)

Basic idea: Replace $R(f'')$ w/ the most conservative (smallest) possible value.

- Terrell looked at a collection of h that would minimize eq BW for various f.
- wanted to maximize eq BW w/r.t. f. (worse case).
- found worse case will be a polynomial.

So use

$$h = 3\left(\frac{R(K)}{35n}\right)^{1/5}\tilde{\sigma}$$

Comments: - creates a b.w. biased against undersmoothing in order to avoid false modes.
- oversmooths often
- Easy Table 10.1 gives $R(K)$ for several K.

Ex pg 291.

Sec 10.2.2 pg 292.

Choice of Kernel.

Turns out that the choice of Kernel has much less influence than Bandwidth.

Epanechnikov showed that if we minimize AMISE w.r.t K, we obtain a min w/

$$K^*(z) = \begin{cases} \frac{3}{4}(1-z^2) & \text{if } |z|<1 \\ 0 & \text{o.w.} \end{cases}$$

Ex/pg 294 So $K^*(z)$ is optimal, but does not do markedly better. so it really doesn't matter. See Table 10.1.

* if using a kernel other than the normal, you need to multiply $K(z)$ by $1_{\{|x|<1\}}$.

(Other ideas using rescaling or reshaping.)

10.2.2.2 Rescalings.

Suppose we have found a b.w. h that works well for Kernel K and we wish to change to kernel L. Unfortunately, h will correspond to a different amount of smoothing with L than w/ K. Can we easily switch?

Suppose h_K & h_L are optimal (AMISE) for sym, mean zero, pos var. kernels K & L resp. Then

$$\frac{h_K}{h_L} = \frac{\delta(K)}{\delta(L)}$$

where

$$\delta(K) = \left(R(K)/\sigma_K^4\right)^{1/5}$$

$\therefore$ to go from h_K for K to L we use $h_L = h_K\,\delta(L)/\delta(K)$. Some common values are listed on page 292 Table 10.1

Sec10.4 Multivariate Methods (A brief word)

Now we wish to esimate* f based on iid samples $X_i = (X_{i1}, \dots, X_{ip})^T$ ~~from~~ of a p-dim r.v.

Method 0: histogram. from earlier.
Method 1: Kernel estimation - product kernel.

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\frac{1}{h_j}K\left(\frac{x_j - X_{ij}}{h_j}\right)$$

where $K(z)$ is a univariate kernel and h_j are fixed b.w. for each coord. $j = 1, \dots, p$.

Can be shown that for the normal kernel, the opt bandwidth is

$$h_i = \left(\frac{4}{n(p+2)}\right)^{1/(p+4)} \hat{\sigma}_i$$

where $\hat{\sigma}_i$ is an est. of the st. dev along the ith coord. → Can rescale to get other b.w. for nonnormal kernels.

Sec 10.4.3.1 Method 2: Nearest neighbor.

- previously we fixed a region of influence & # of obs varied.
- here allow region of influence to vary, but contain a fixed number of observations.

The kth-nearest neighbor density estimator

$$\hat{f}(x) = \frac{k}{n V_p d_k(x)^p}$$

where

- $d_k(x)$ = is the Euclidean distance from x to the kth nearest observed point.
- V_p = is the volume of the unit sphere in p dimensions = $V_p = \pi^{p/2} / \Gamma(p/2+1)$
- p-dim of data.

Note: ① only unknown is $d_k(x)$.

② k plays role similar to b.w.

large k - smooth small k - wiggly.

There are other adaptive methods.

10.4.1. Problems: Multivariate density est is a diff task than univariate: It is hard to visualize any density estimate in more than 2 or 3 dims. Not useful as an exploratory tool.

Also,

Curse of dimensionality: # of points req for a good est. goes up radically as p increases.

Ex Est. a p-variate normal w/ optimal relative mean squared error $\doteq .0289$.

p	n
1	30
2	180
3	806
5	17,400
10	112,000,000
15	2,190,000,000,000