

Lecture 5

Estimating a Univariate Density

In this lecture we consider estimating the PDF of a continuous random variable based on a simple random sample from it.

Again suppose the data we have is a sample Y_1, Y_2, \dots, Y_m that is independently and identically distributed from the distribution F .

Estimating the Density Based on a Histogram

As the PDF is the derivative of the CDF, it will tend to be less smooth than the CDF and often more difficult to estimate.

Background: See Simonoff (1996), Chapter 2, whom we follow.

Suppose we can consider $F(y)$ to have a finite range from A to B .

The relationship between the PDF and CDF:

$$f(x) \equiv \frac{d}{dx}F(x) \equiv \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h}$$

suggests splitting up the range of the distribution into K equisized intervals each with width $h = (B - A)/K$.

If the number of intervals is large enough, then we can consider an estimator of the form:

$$\hat{f}(y) = \frac{F_m(b_{j+1}) - F_m(b_j)}{h}, \quad x \in (b_j, b_{j+1}],$$

where $b_j = A + (B - A)(j - 1)/K$, $j = 1, \dots, K + 1$ and $(b_j, b_{j+1}]$ defines the boundaries of the j th interval.

This is the familiar *histogram* estimator of $f(y)$.

The advantages of the histogram in this setting are:

ease of interpretability and
convenient construction with most statistical packages

How can we evaluate $\hat{f}(y)$ as an estimator of $f(y)$?

For a given value of y , it is natural to consider the squared error,

$$\text{SE}(y) = [\hat{f}(y) - f(y)]^2,$$

and its expected value (mean squared error),

$$\text{MSE}(y) = \mathbb{E}_f [\hat{f}(y) - f(y)]^2.$$

If we wish to measure *global accuracy* over the full interval $[A, B]$, we can consider the *integrated squared error*,

$$\text{ISE} = \int_A^B [\hat{f}(u) - f(u)]^2 du,$$

and its expected value, *mean integrated squared error* (MISE).

We can measure these for any sample size and K , but here will consider the *asymptotic behavior* when the sample size $m \rightarrow \infty$.

Idea: As we get more data we should increase the number of intervals to capture the detailed structure of $f(y)$
– but do so slower than the sample size increases to reduce the variability of $\hat{f}(y)$ within each interval.

Based on the binomial distribution for $mF_m(y)$ and the distribution of $m\hat{f}(y)$ is binomial with m trials and probability of success $F(b_{j+1}) - F(b_j)$ where $x \in (b_j, b_{j+1})$.

Simonoff (1996) shows that if the interval width $h \rightarrow 0$, and $mh \rightarrow \infty$, as $m \rightarrow \infty$ and $f(y)$ is smooth enough ($f'(y)$ is absolutely continuous and square integrable), then

$$\begin{aligned} \text{Bias} [\hat{f}(y)] &\equiv E_f [\hat{f}(y)] - f(y) \\ &= \frac{1}{2} f'(y) [h - 2(y - b_j)] + O(h^2), \quad y \in (b_j, b_{j+1}], \end{aligned}$$

while the variance is

$$\text{Var} [\hat{f}(y)] = \frac{f(y)}{mh} + O(m^{-1}).$$

Combining the squared bias and variance yields the mean squared error,

$$\begin{aligned}\text{MSE} [\hat{f}(y)] &= \text{Var} [\hat{f}(y)] + \text{Bias}^2 [\hat{f}(y)] \\ &= \frac{f(y)}{mh} + \frac{f'(y)^2}{4} [h - 2(r - b_j)]^2 \\ &\quad + O(m^{-1}) + O(h^3).\end{aligned}$$

Integrating over each interval, and then summing interval by interval, gives

$$\text{MISE} = \frac{1}{mh} + \frac{h^2 R(f')}{12} + O(m^{-1}) + O(h^3),$$

where

$$R(v) = \int_{-\infty}^{\infty} [v(x)]^2 dx.$$

We will write AMISE to represent the asymptotic MISE (that is, the leading two terms in the expansion of MISE).

The minimization of MISE requires explicitly balancing bias and variance through the choice of the number of bins K .

The minimizer of AMISE is

$$h_0 = \left[\frac{6}{R(f')} \right]^{1/3} m^{-1/3}.$$

In practice, we need to specify a particular estimate of f to operationalize this rule.

A reasonable candidate is a Gaussian distribution with mean and variance matching the sample. This leads to the rule:

$$h_0 = 3.491sm^{-1/3}.$$

where s is the sample standard deviation.

Many rules-of-thumb have been suggested that are similar to:

$$h_0 = 2IQRm^{-1/3},$$

where IQR is an estimate of the *interquartile range* of the distribution. If we use a distribution of normal shape but with the spread of the uniform we get the rule $h_0 = m^{-1/3}$.

See Simonoff (1996) for a discussion of these issues.

A function f defined over the interval $[A, B]$ is *absolutely continuous* if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that, for each finite collection of non overlapping intervals $\{x_i, y_i\}_{i=1}^n$ with

$$\sum_{i=1}^n |y_i - x_i| < \delta$$

we have that

$$\sum_{i=1}^n |f(y_i) - f(x_i)| < \epsilon$$

This condition is necessary and sufficient for there to exist a function $h(y)$ such that

$$f(y) = \int_{-\infty}^y h(x) dx$$

for each y .