IEOR 165 – Lecture 10
Distribution Estimation

1 Motivating Problem

Consider a situation where we have iid data x_i from some unknown distribution. One problem of interest is estimating the distribution that is generating the data. There are many useful examples of this abstract problem, including:

- We are analyzing a telephone call center, and we measure the amount of time required to provide service to each incoming phone call. Though an exponential distribution is a commonly used model, it can be useful to estimate the distribution of the call lengths to help validate the use of an exponential distribution model. In some cases, we may find by using the data that an exponential distribution is not a good model.
- We are analyzing midterm scores within a class, and would like to gain a better understanding of how well the class scored on the exam.
- We are taking pictures with a digital camera, and would like to determine if our picture is underexposed or overexposed.

Within the problem of estimating the distribution using iid x_i, there are two distinct problems:

1. Estimate the cdf $F_x(u)$.
2. Estimate the pdf $f_x(u)$.

We will consider each of these problems. We will drop the x subscript to simply the notation.

2 Empirical Distribution Function

One natural estimator for the cdf $F(u)$ is known as the empirical distribution function $\hat{F}(u)$, and it is defined as

$$\hat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq u).$$

In fact, the Glivenko-Cantelli theorem tell us that

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{u\in\mathbb{R}} \left|\hat{F}_n(u) - F(u)\right| = 0\right) = 1.$$

This means the empirical distribution function $\hat{F}_n(u)$ has good statistical problems, because it means that with probability 1 the discrepancy between $\hat{F}(u)$ and $F(u)$ goes to 0.

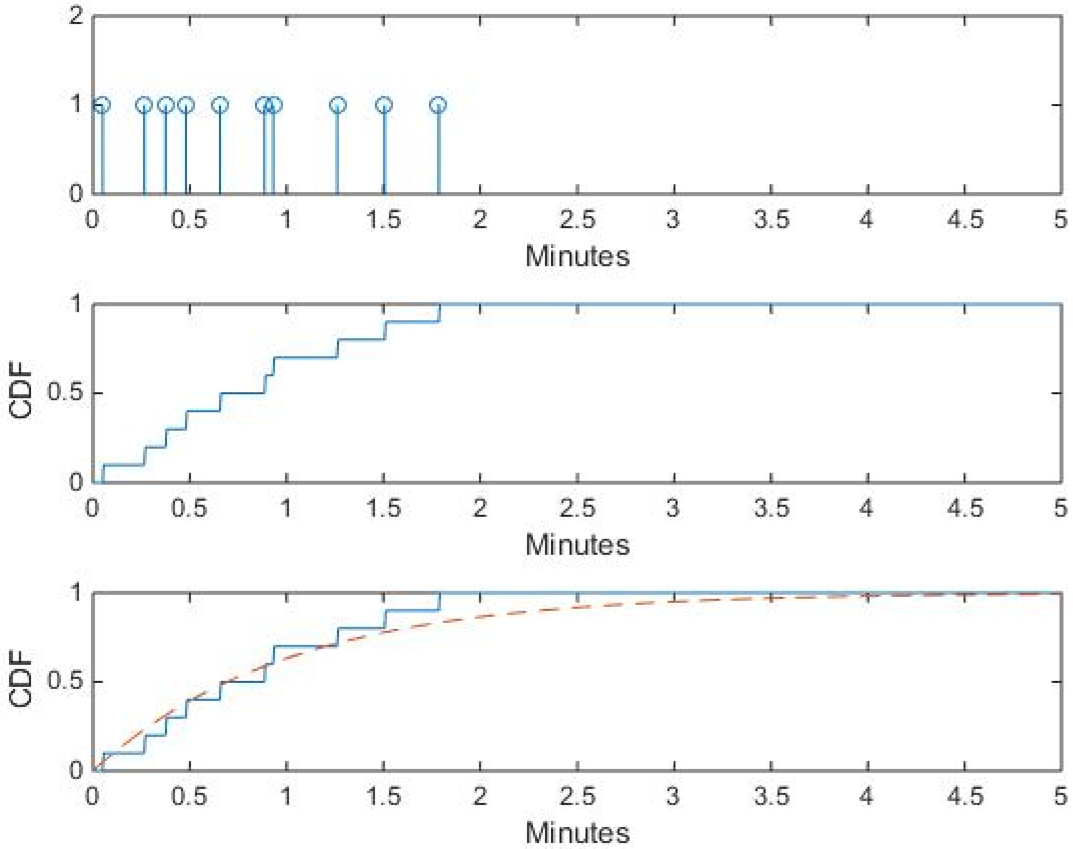
2.1 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

$$\{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}.$$

Q: Plot the empirical distribution function.

A: The first step is to plot the location of each data point (top plot). The second step is to draw the empirical distribution function (middle plot) by increasing the function value by $1/n = 1/10 = 0.1$ at the location of each data point. The bottom plot compares the empirical distribution to the true distribution function (an exponential distribution with rate parameter $\lambda = 1$) that was used to generate this data.



3 Challenge with Density Estimation

Though the empirical distribution function $\hat{F}_n(u)$ has good statistical problems, it is not well-suited for estimating the density (i.e., the pdf) of the distribution. Recall that when a density exists, it is equal to the derivative of the distribution function. And so the reason that $\hat{F}_n(u)$ is not well-suited for estimating the pdf is that it is not differentiable in the normal sense.

To better understand this, recall that the Dirac delta function is defined as a measure such that

$$\delta(A) = \begin{cases} 1, & \text{if } 0 \in A \\ 0, & \text{otherwise} \end{cases}$$

Informally, the Dirac delta function is a function defined such that

$$\delta(u) = \begin{cases} 0, & \text{if } u \neq 0 \\ +\infty, & \text{if } u = 0 \end{cases}$$

and

$$\int_{-\infty}^{+\infty} g(u)\delta(u)du = g(0).$$

Then, we can define an estimate of the density function by

$$\hat{f}_n^{\text{edf}}(u) = \frac{1}{n}\sum_{i=1}^{n} \delta(u - x_i)$$

since the following relationship holds

$$\int_{\mathbb{R}} \mathbf{1}(t \leq u)\hat{f}_n^{\text{edf}}(dt) = \hat{F}_n(u).$$

Observe that for a differentiable distribution $F(u)$ with corresponding density $f(u)$, we have that

$$F(u) = \int_{-\infty}^{u} f(t)dt = \int_{\mathbb{R}} \mathbf{1}(t \leq u)f(t)dt.$$

So we can interpret the $\hat{f}_n^{\text{edf}}(u)$ as a derivative of the non-differentiable $\hat{F}_n(u)$. However, there is a problem with using this $\hat{f}_n^{\text{edf}}(u)$ as an estimate of the density. It essentially places all the density mass at points where the data x_i has been measured; the density is zero at every other point on the real line. This is not a good estimate of the density because if we know that our unknown density is continuous, then we would rather have an estimate that interpolates between measured data points.

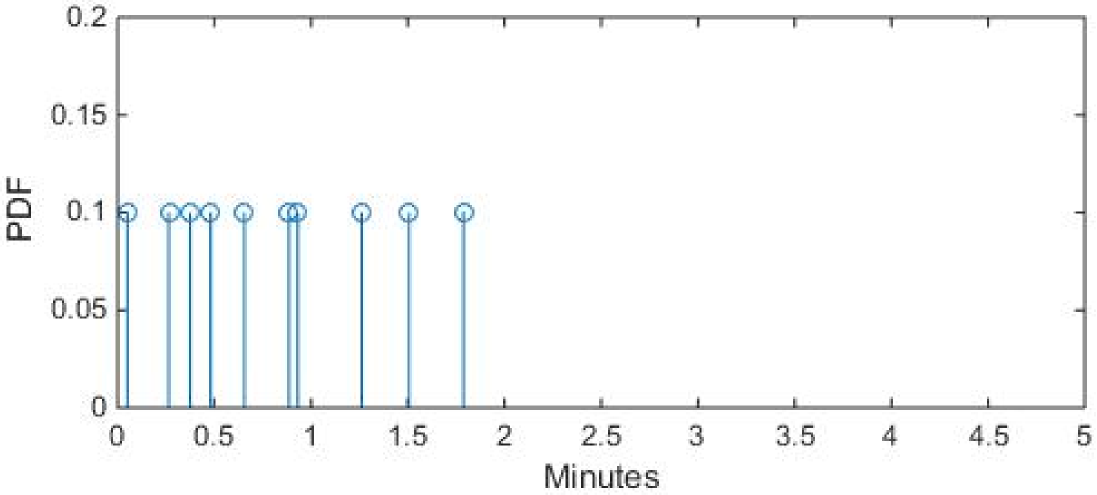
3.1 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

$$\{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}.$$

Q: Plot the estimated pdf using the derivative of the empirical distribution function.

A: We plot the location of each data point and give the Dirac delta an amplitude of $1/n = 1/10 = 0.1$.



4 Histograms

The simplest estimate of the density is instead a histogram. The idea of a histogram is to

1. begin by specifying a support for the distribution, meaning specifying the range over which the density is non-zero; suppose we specify that the support is $[u, v]$;
2. next specify a set of bins partitioning the support, meaning specifying a set of strictly increasing values $b_0 = u < b_1 < \ldots < b_N = v$ that start at u and end at v;
3. we count the number of data points falling into each bin; specifically, we define

$$C_j = \sum_{i=1}^{n} \mathbf{1}(b_{j-1} \le x_i < b_j), \qquad \text{for } j = 1, \ldots, N-1$$
$$C_N = \sum_{i=1}^{n} \mathbf{1}(b_{N-1} \le x_i \le b_N).$$

4. finally, we define our histogram estimate of the density by

$$\hat{f}_n^{\text{his}}(u) = \begin{cases} \frac{C_j}{n(b_j - b_{j-1})}, & \text{if } b_{j-1} \le u < b_j \\ \frac{C_N}{n(b_N - b_{N-1})}, & \text{if } u = b_N \\ 0, & \text{otherwise} \end{cases}$$

Note that instead of specifying the edges of the bins $b_0, \ldots, b_N$, we could have instead specified the desired number of bins N and set the bin edges to be $b_j = u + j \cdot (v - u)/N$.

4.1 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

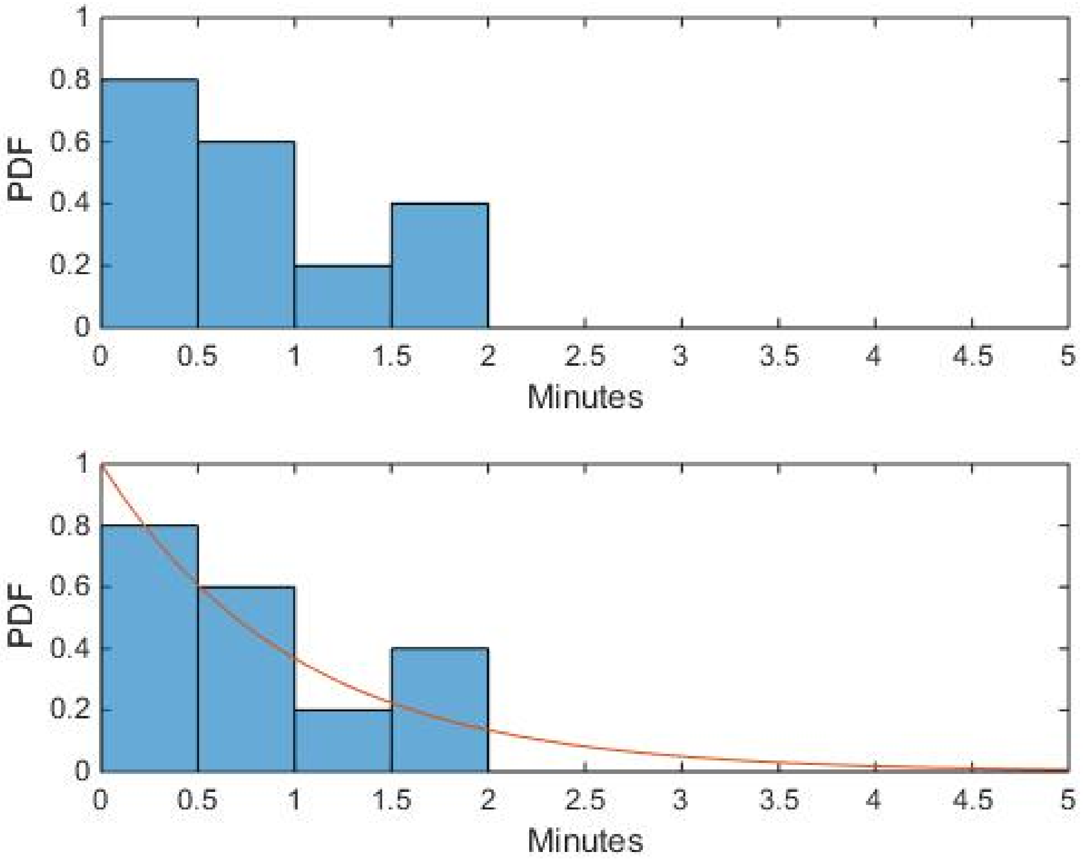
$$\{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}.$$

Q: Plot the histogram using bin edges $\{0, 0.5, 1, 1.5, 2, 10\}$.

A: We first count the number of data points falling into each bin. We have 4,3,1,2,0 points in bins 0–0.5, 0.5–1, 1–1.5, 1.5–2, 2–10, respectively. The next step is to use the formulate $\frac{C_j}{n(b_j - b_{j-1})}$ to normalize the data counts:

$$\begin{aligned}
&\tfrac{4}{10\cdot(0.5-0)} = 0.8\\
&\tfrac{3}{10\cdot(1-0.5)} = 0.6\\
&\tfrac{1}{10\cdot(1.5-1)} = 0.2\\
&\tfrac{2}{10\cdot(2-1.5)} = 0.4\\
&\tfrac{0}{10\cdot(10-2)} = 0
\end{aligned}$$

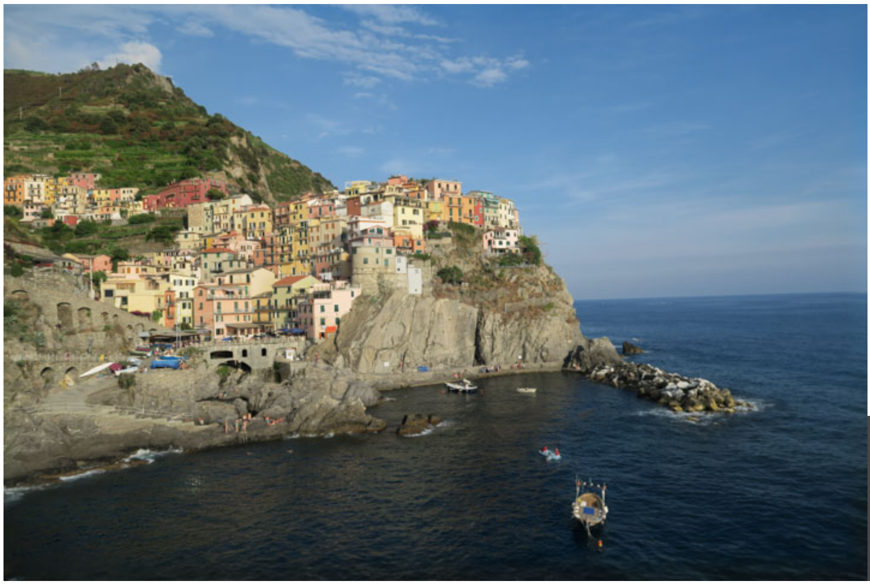
Finally, we plot the histogram:

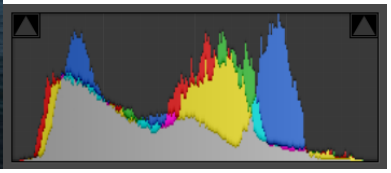


This top plot is the histogram, and the bottom plot compares the histogram to the actual pdf that was used to generate the data.

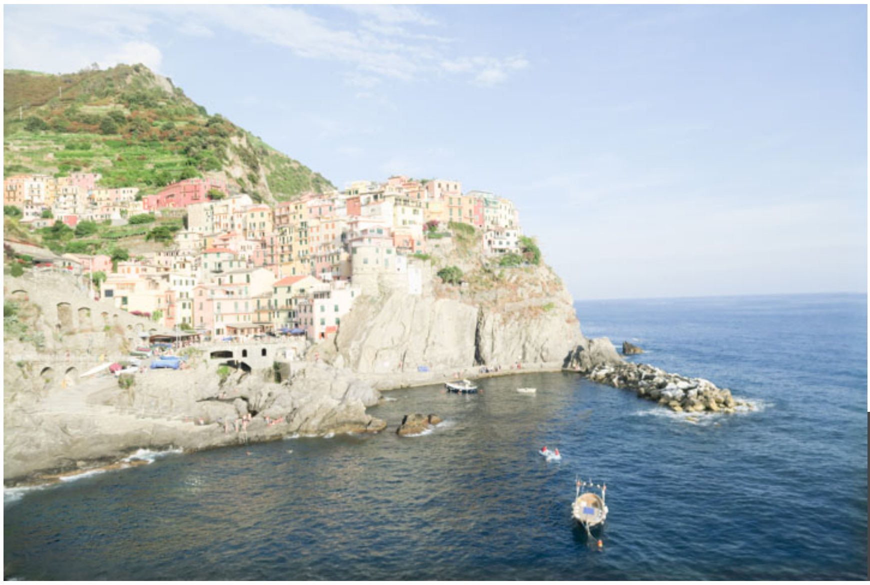
4.2 Example: Image Exposure

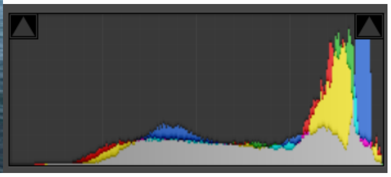
Histograms are useful in taking and editing digital photographs. For instance, an image with correct exposure and its histogram is:



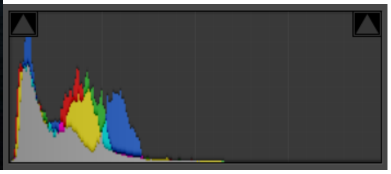


An overexposed image and its histogram is:





An underexposed image and its histogram is:



5 Kernel Density Estimation

5.1 Motivation

Histogram estimates of the density $f(u)$ are not continuous, and so it is interesting to consider other approaches that produce continuous estimates of the density. One idea is to combine the idea of a histogram with the density estimate (written in terms of Dirac deltas) that was generated by differentiating the empirical distribution function. To start, recall that the problem of the density estimate generated by the empirical distribution function

$$\hat{f}_n^{\text{edf}}(u) = \frac{1}{n}\sum_{i=1}^{n}\delta(u - x_i)$$

is that it is zero at all points except the x_i where data was seen. The next thing to note is that the histogram works by giving width to counting the amount of data in some region; the Dirac deltas were problematic because they give zero width to counting data in some region.

5.2 Kernel Function

Given these two ideas, a proposal for another approach to estimate the density is to give the Dirac deltas some width by replacing them by copies of a kernel function $K(\cdot) : \mathbb{R} \to \mathbb{R}$ that has

- finite support: $K(u) = 0$ for $|u| \geq 1$;
- even symmetry: $K(u) = K(-u)$;

- positive values: $K(u) > 0$ for $|u| < 1$;
- unit integral: $\int_{\mathbb{R}} K(u)du = 1$.

Furthermore, we would like to be able to specify the width of the copies of the kernel functions. We will use the variable h to denote *bandwidth*, and note that the function $\frac{1}{h}K(u/h)$ has

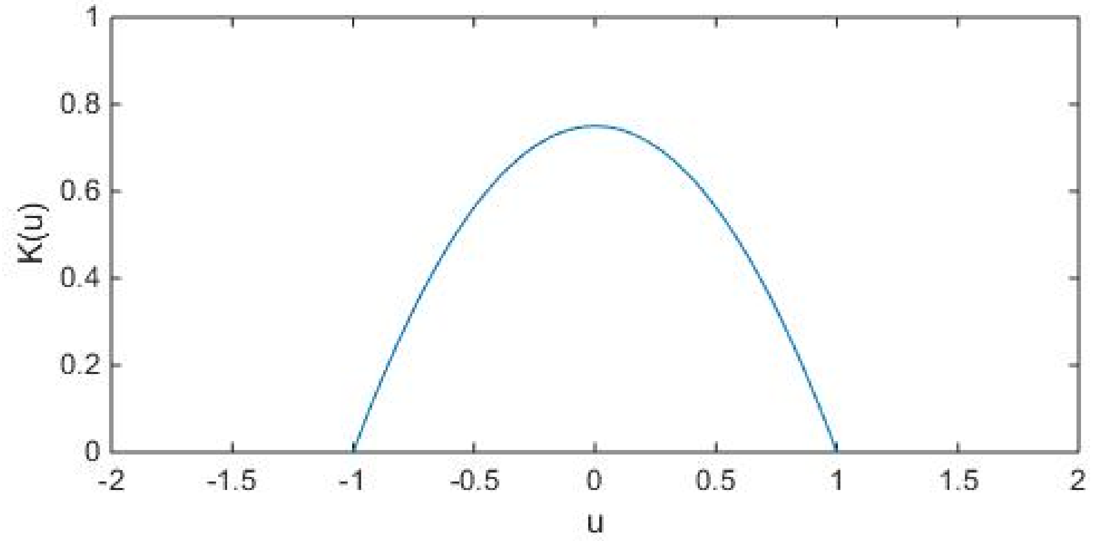
- support over $[-h, h]$;
- even symmetry $\frac{1}{h}K(u/h) = \frac{1}{h}K(-u/h)$;
- positive values $\frac{1}{h}K(u/h) > 0$ for $|u| < h$;
- unit integral: $\int_{\mathbb{R}} \frac{1}{h}K(u/h)du = 1$;
- maximum value of $\frac{1}{h}\max_u K(u)$.

An important conceptual point is that $\frac{1}{h}K(u/h)$ weakly converges to a Dirac delta when $h \to 0$.

It is interesting to consider some examples of kernel functions. The table below lists some common examples:

Kernel Function	Equation
Uniform	$K(u) = \frac{1}{2}\mathbf{1}(\|u\| \leq 1)$
Triangular	$K(u) = (1 - \|u\|)\mathbf{1}(\|u\| \leq 1)$
Epanechnikov	$K(u) = \frac{3}{4}(1 - \|u\|^2)\mathbf{1}(\|u\| \leq 1)$
Quartic/biweight	$K(u) = \frac{15}{16}(1 - \|u\|^2)^2\mathbf{1}(\|u\| \leq 1)$

The Epanechnikov kernel is plotted below, and the other kernel functions essentially look the same.



5.3 Estimate

The corresponding estimate of the density is then given by

$$\hat{f}_n^{\textsf{kde}}(u) = \hat{f}_n^{\textsf{edf}}(u) * \frac{1}{h}K(u/h),$$

where the symbol $*$ denotes a *convolution*. Recall that a convolution of two functions f, g is defined as

$$f(x) * g(x) = \int_{\mathbb{R}} f(\tau) \cdot g(\tau - x) d\tau.$$

In our context, we can simplify the convolution integral, which results in the following equivalent equation for the density estimate

$$\hat{f}_n^{\textsf{kde}}(u) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{u - x_i}{h}\right).$$

The kernel density estimate naturally generalizes to higher dimensions:

$$\hat{f}_n^{\textsf{kde}}(u) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\|u - x_i\|}{h}\right),$$

where d is the dimension of the random variables (i.e., $x_i \in \mathbb{R}^d$).

5.4 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

$$\{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}.$$

Q: Suppose we choose $h = 0.5$, and that we use the Epanechnikov kernel. Compute the estimated pdf using the kernel density approach at the point $u = 0.7$.

A: We first compute the quantities $K\left(\frac{u-x_i}{h}\right)$ for each data point. We have

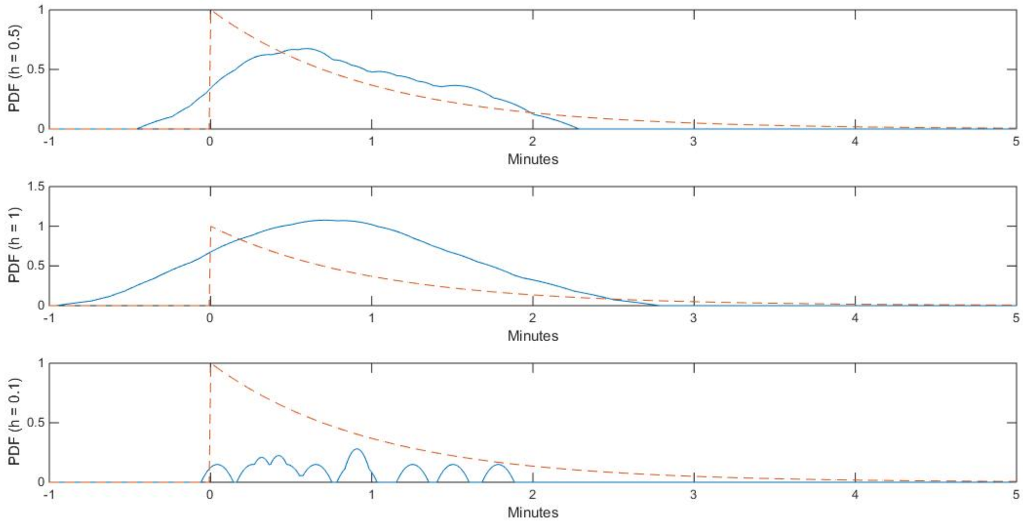
$$\begin{aligned}
&K((0.7 - 0.66)/0.5) = K(0.08) = 3/4 \cdot (1 - 0.08^2) = 0.7452\\
&K((0.7 - 0.05)/0.5) = K(1.3) = 0\\
&K((0.7 - 0.27)/0.5) = K(0.86) = 3/4 \cdot (1 - 0.86^2) = 0.1953\\
&K((0.7 - 1.26)/0.5) = K(-1.12) = 0\\
&K((0.7 - 1.51)/0.5) = K(-1.62) = 0\\
&K((0.7 - 0.38)/0.5) = K(0.64) = 3/4 \cdot (1 - 0.64^2) = 0.4428\\
&K((0.7 - 1.79)/0.5) = K(-2.18) = 0\\
&K((0.7 - 0.94)/0.5) = K(-0.48) = 3/4 \cdot (1 - 0.48^2) = 0.5772\\
&K((0.7 - 0.48)/0.5) = K(0.44) = 3/4 \cdot (1 - 0.44^2) = 0.6048\\
&K((0.7 - 0.89)/0.5) = K(-0.38) = 3/4 \cdot (1 - 0.38^2) = 0.6417
\end{aligned}$$

Finally, we compute

$$\hat{f}_n^{\textsf{kde}}(0.7) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{0.7 - x_i}{h}\right) =$$

$$\frac{0.7452 + 0.1953 + 0.4428 + 0.5772 + 0.6048 + 0.6417}{10 \cdot 0.5} = 0.64.$$

Plotting the kernel density estimate (KDE) by hand is difficult. A computer is typically used, and standard packages exist to do this. One important point is that the estimate is sensitive to the value of the *bandwidth*. Unfortunately, the bandwidth cannot be chosen using cross-validation. An example of the estimated density for three different values of h and compared to the true pdf is shown below:


PDF (h = 0.5)
Minutes
PDF (h = 1)
Minutes
PDF (h = 0.1)
Minutes