Goal: Smoother Density Estimation.

3 Kernel Density Estimation

Recall the definition of a density function

$$f(y) \equiv \frac{d}{dy}F(y) \equiv \lim_{h\to 0}\frac{F(y+h)-F(y-h)}{2h} = \lim_{h\to 0}\frac{F(y+h)-F(y)}{h},$$

where $F(x)$ is the cdf of the random variable Y.

take this and approximate w/ small fixed h, use Ecdf.

$$\hat{f}(x) = \frac{\hat{F}_n(y+h)-\hat{F}_n(y)}{h}$$ histogram.

What if instead, we replace $F(y+h) - F(y-h)$? with ecdf

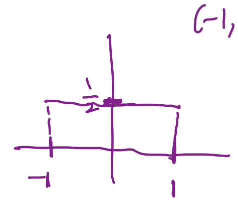
$$\Rightarrow \hat{f}(y) = \frac{\hat{F}_n(y+h)-\hat{F}_n(y-h)}{2h} = \frac{\#\{y_i \in (y-h, y+h]\}}{2nh}$$

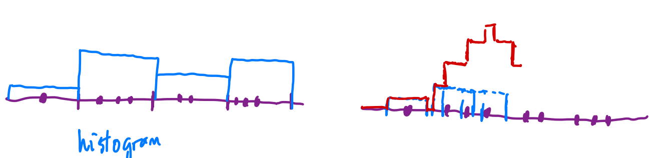
$$= \frac{\sum_{i=1}^n \mathbb{I}(y_i \in (y-h, y+h]\}}{2nh}$$

$$= \frac{1}{nh}\sum_{i=1}^n \frac{1}{2}\mathbb{I}(y-h < y_i \le y+h)$$

$$= \frac{1}{nh}\sum_{i=1}^n \frac{1}{2}\mathbb{I}\left(-1 < \frac{y-y_i}{h} \le 1\right)$$

$$= \frac{1}{nh}\sum_{i=1}^n K\left(\frac{y-y_i}{h}\right)$$ where K (Kernel function) is a Uniform density on $(-1,1]$





↑ still not continuous (because Uniform density not continuous!)

$\Rightarrow$ another Kernel may lead to smoother estimate.

A kernel function assigns weights to the contribution given by each y_i to $\hat{f}(y)$, depending on proximity to y.

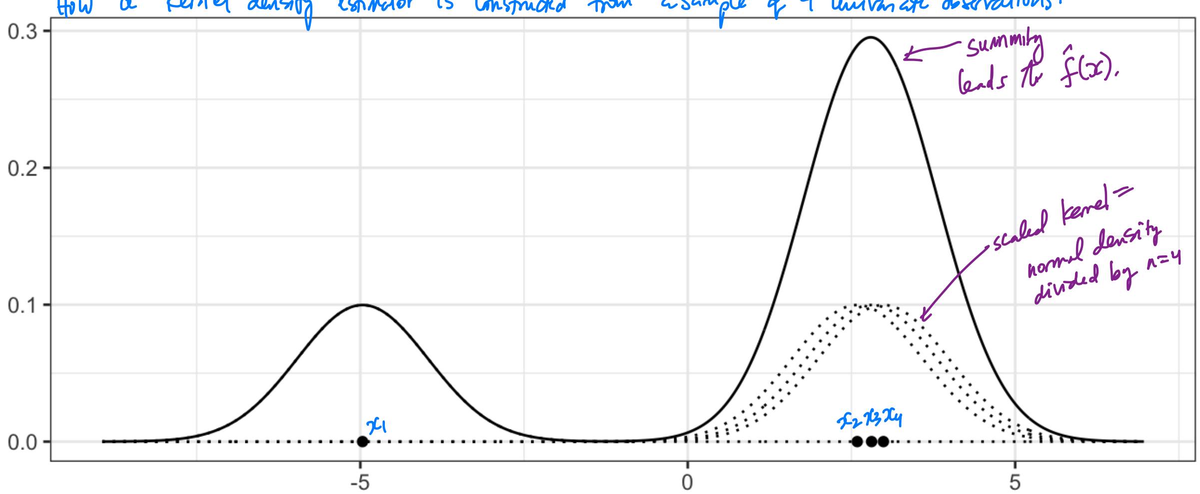
This will weight all points within h of x equally. A univariate *kernel density estimator* will allow a more flexible weighting scheme.

Typically, kernel functions are positive everywhere and symmetric about zero.

Examples of ideas for such functions? Normal density, Student t (others exist).

Additionally, constraining K so that $\int z^2 K(z)\,dz = 1$ allows h to play role of scale parameter (not required).

How a kernel density estimator is constructed from a sample of 4 univariate observations:



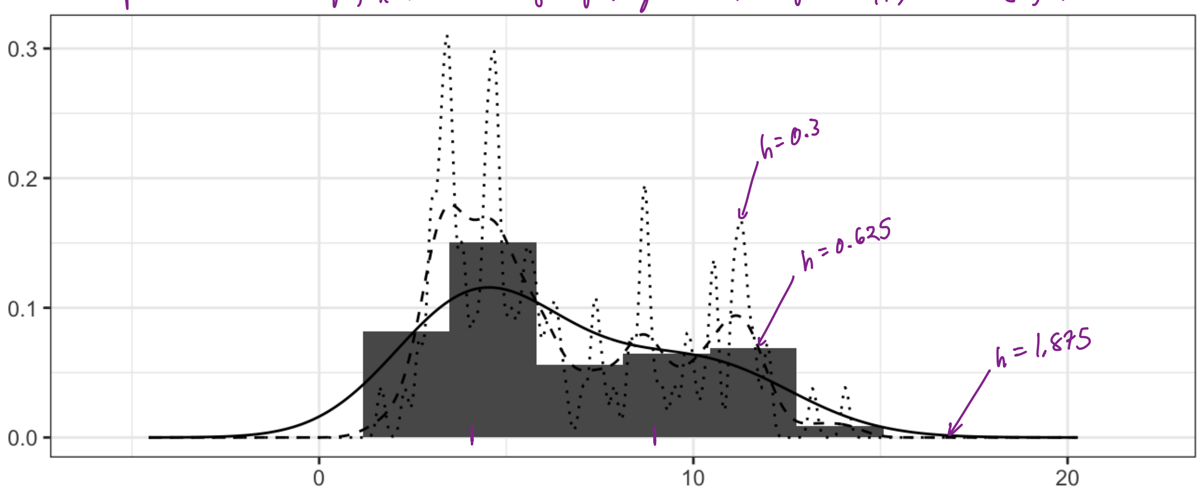
3.1 Choice of Bandwidth

The bandwidth parameter controls the smoothness of the density estimate. for a given kernel.

bandwidth determines tradeoff btw bias and variance.

The tradeoff that results from choosing the bandwidth + kernel can be quantified through a measure of accuracy of $\hat{f}$, such as MISE.

Example: Bimodal density, n=100 from equally weighted mixture of N(4,1) and N(9,4).



For large h, oversmoothing (lose 2nd mode).

For small h, undersmoothing (many false modes).

To understand bandwidth selection, let us analyze MISE. Suppose that K is a symmetric, continuous probability density function with mean 0 and variance $0 < \sigma_K^2 < \infty$. Let $R(g) = \int g^2(z)dz$. Recall that

$$\text{MISE} = \int \text{MSE}(\hat{f}(x))dx =$$

Now let $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

To minimize AMISE with respect to h,

The term $R(f'')$ measures the roughness of the true underlying density. In general, rougher densities are more difficult to estimate and require smaller bandwidth.

The term $[\sigma_K R(K)]^{4/5}$ is a function of the kernel function K.

3.1.1 Cross Validation

3.1.2 Plug-in Methods

If the reference density f is Gaussian and a Gaussian kernel K is used,

Empirical estimation of $R(f'')$ may be a better option.

3.2 Choice of Kernel

There are two choices we have to make to perform density estimation:

3.2.1 Epanechnikov Kernel

The *Epanechnikov kernel* results from choosing K to minimize $[\sigma_K R(K)]^{4/5}$, restricted to be a symmetric density with finite moments and variance equal to 1

3.2.2 Canonical Kernels

Unfortunately a particular value of h corresponds to a different amount of smoothing depending on which kernel is being used.

Let h_K and h_L denote the bandwidths that minimize AMISE when using symmetric kernel densities K and L. Then,

Suppose we rescale a kernel shape so that $h = 1$ corresponds to a bandwidth of $\delta(K)$,

3.3 Bootstrapping and Variability Plot