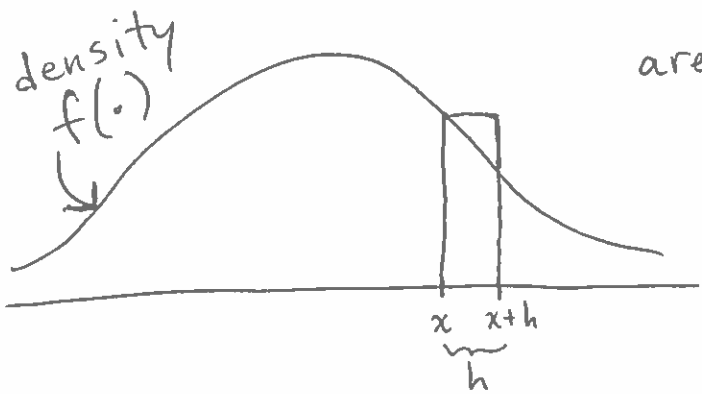
STAT 518 --- Nonparametric Density Estimation

• The <u>probability density function</u> (or <u>density</u>) of a continuous random variable *X* describes its probability distribution.

• We denote the density as $f(x)$

• Note that if F(*x*) is the c.d.f. of *X*, then

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



area of bar $= h f(x)$

$\approx F(x+h) - F(x)$

$P(x \le X < x+h)$

Two important properties of density functions

(1) They are always <u>nonnegative</u>: $f(x) \geq 0$ for all x

(2) The total area under a density curve is always <u>1</u>.

• In real data analysis, we do not know the true density, so we can estimate it using sample data $X_1, X_2, \ldots, X_n$.

<u>Parametric approach</u>: Assume a specific functional form (e.g., normal, gamma, etc.) for the density and use the sample data to estimate certain <u>unknown parameters</u>.

<u>Example</u>: Could assume the density is <u>normal</u> and get sample estimates of <u>μ</u> and <u>σ^2</u>.

• **The <u>nonparametric</u> approach is to make very <u>few</u> assumptions about the functional form of the density.**

Histograms

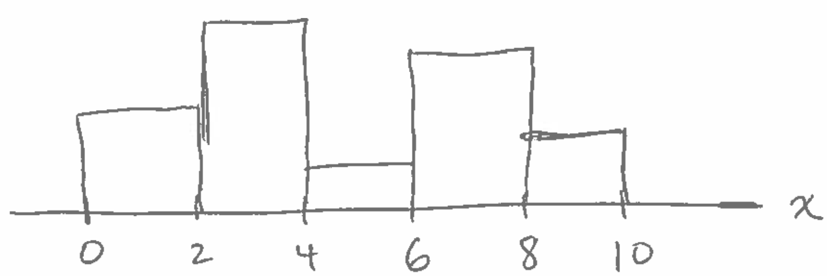
• **A simple density estimator is a <u>histogram</u>.**

• **In introductory statistics, we study the** frequency **histogram having bins with bars whose height is the count of sample observations falling in that bin.**

• **If we rescale the heights of each bar so that the <u>total combined area</u> within all the bars is 1, we have a <u>histogram density estimate</u>.**

• **Assume there are *K* bins, each of width *h*:**

Picture ($K = 5$, $h = 2$):



• **In general, this histogram is:**

$$\hat{f}(x) = \frac{n_j}{nh}, \qquad x \in (b_j, b_{j+1}]$$

where $(b_j, b_{j+1}]$ is the interval for the j-th bin

$n_j = \{\# x_i : b_j < x_i \le b_{j+1}\}$ = count of observations falling in the j-th bin

and $h = b_{j+1} - b_j$ = j-th bin width

• **The total combined area within all bars is**

$$\sum_j \underbrace{(h)}_{\text{width}} \underbrace{\left(\frac{n_j}{nh}\right)}_{\text{height}} = \frac{1}{n}\sum_j n_j = \frac{1}{n}(n) = 1$$

• **The R function `hist` produces such histograms.**

• **The choice of bin width *h* determines the number of bins, which can affect the appearance of the estimate.**

• **A simple rule of thumb for choosing *h* is derived from a normal density:**

Let $h = \dfrac{3.49\,\hat{\sigma}}{n^{1/3}}$

where $\hat{\sigma} = s \quad \text{or} \quad IQR/1.34$

• **Note: the sample standard deviation *s* is a consistent estimator of σ, as is *IQR* / 1.34 when the true density is normal.**

• **In reality, this provides a good initial choice of *h*, which may then be adjusted by trial and error.**

• **Choosing *h* too small produces <u>many bins</u> and a density estimate that is too** rough**.**

• **Choosing *h* too large produces <u>few bins</u> and a density estimate that is** oversimplified**.**

Example 1: Waiting time data (Old Faithful eruptions)

Default number of bins = 12

- Main characteristic of density estimate:

Bimodal → peaks around 50 minutes and 80 minutes

Example 2: New York City – windspeed measurements

– Default number of bins = 11

• We could also let the bin width vary across bins, choosing a __large__ width in regions where we expect the density to be <u>flatter</u> and a __small__ width in regions where we expect the density to be <u>spiky</u>.

Kernel Density Estimation

• An obvious drawback to the histogram density estimate is that it is not __smooth__.

• A <u>kernel density estimate</u> (k.d.e.) produces a smooth estimate and works similarly to the kernel regression method.

• As $n \to \infty$, the k.d.e. will approach the true density $f(x)$ more quickly than the histogram will.

Recall: $f(x) = \frac{d}{dx} F(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}$

- **Plug in the e.d.f. for *F*(·) to obtain:**

$$\hat{f}(x) = \frac{\#\, x_i \text{ in } (x-h,\, x+h]}{2nh}$$

- **This is exactly the same as**

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

with *K*(*u*) = $\begin{cases} \frac{1}{2} & \text{if } -1 < u \le 1 \\ 0 & \text{otherwise} \end{cases}$

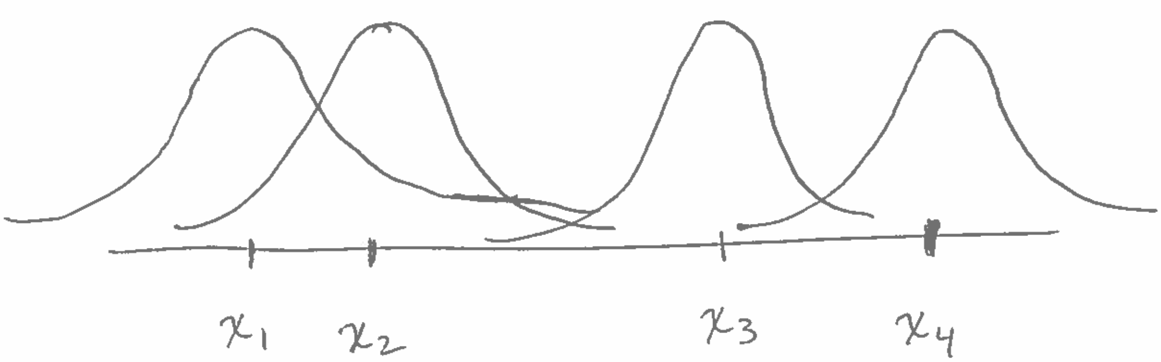
→ **a kernel estimate with a** uniform **kernel function.**

- **However, with the** uniform **kernel, the resulting density estimate is not smooth.**

- **Better choices of kernel function *K*(·) include:**

Normal, Epanechnikov

- **Let *K*(·) in the above k.d.e. formula be a standard normal kernel function.**
- **Then for, say, *h* = 1:**



• **We see at each point *x*, the k.d.e.** $\hat{f}(x)$ **is the average of normal densities, centered at** each x_i value

• **Sample values near *x* will contribute** substantially to $\hat{f}(x)$

• **Sample values far from *x* will** hardly contribute to $\hat{f}(x)$

Role of the Bandwidth *h*

• **If *h* increases, these normal densities become** flatter **and more** spread out
→ more sample values contribute to $\hat{f}(x)$
→ estimate is smoother overall

• **If *h* decreases, these normal densities become** taller **and** narrower
→ fewer sample values contribute to $\hat{f}(x)$
→ estimate is bumpier overall

• **Rule of thumb for choosing *h* (again based on the true density being normal):**

Let $$h \approx \frac{1.06\,\hat{\sigma}}{n^{1/5}}$$

where $$\hat{\sigma} = \min\left\{ s, \frac{IQR}{1.34} \right\}$$

• **In reality, this provides a good initial choice of *h*, which may then be adjusted by trial and error.**

• **The `density` function in R produces a kernel density estimate.**

Example 1: Old faithful waiting time data

→ density appears bimodal

highest peak around 80 minutes

2nd major peak around 50 minutes

Default bandwidth ≈ 4.7

Example 2: NYC wind speed data

default bandwidth ≈ 1.2

→ density appears very slightly skewed right

main peak around 10 mph

two "shoulders" around 15 mph

and 20 mph

• **As with kernel regression, kernel density estimators tend to be biased at the left and right edges:** boundary bias

• **The k.d.e. also has a tendency to be too flat (not rise or dip enough) in the peaks and valleys of the density.**

• **An option is to use a bandwidth that <u>varies</u> over the region (being** larger **where the density is expected to be flat and** smaller **where the density is expected to have bumps).**