The probability density function (or density) 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

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

Two important properties of density functions:

1. They are always nonnegative: $f(x) \geq 0$ for all $x$.

2. The total area under a density curve is always 1.

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

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

Example: Could assume the density is normal and get sample estimates of $\mu$ and $\sigma^2$. 
• The **nonparametric** approach is to make very **few** assumptions about the functional form of the density.

**Histograms**

• A simple density estimator is a **histogram**.

• 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 **total combined area** within all the bars is 1, we have a **histogram density estimate**.

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

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

![Histogram Diagram]

• In general, this histogram is:

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

where

\[ (b_j, b_{j+1}] \text{ is the interval for the } j\text{-th bin } \]

\[ n_j = \# \{ x_i : b_j < x_i \leq b_{j+1} \} = \text{count of observations falling in the } j\text{-th bin } \]

and \[ h = b_{j+1} - b_j = j\text{-th bin width } \]
• The total combined area within all bars is
\[ \sum_j \left( h \right) \left( \frac{n_j}{nh} \right) = \frac{1}{n} \sum_j n_j = \frac{1}{n} (n) = 1 \]

  width \quad \text{height}

• The R function \texttt{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
  \[ h = \frac{3.49 \hat{\sigma}}{n^{1/3}} \]

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

• Note: the sample standard deviation \( s \) is a consistent estimator of \( \sigma \), as is \( \text{IQR} / 1.34 \) 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.

• \textbf{Choosing} \( h \) too small produces many bins and a density estimate that is too rough.

• \textbf{Choosing} \( h \) too large produces few bins 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 $\rightarrow$ 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 **flatter** and a **small** width in regions where we expect the density to be **spiky**.

**Kernel Density Estimation**

- An obvious drawback to the histogram density estimate is that it is not **smooth**.

- A **kernel density estimate** (k.d.e.) produces a smooth estimate and works similarly to the kernel regression method.

- As $n \rightarrow \infty$, 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 \( F(\cdot) \) to obtain:
  \[ \hat{f}(x) = \frac{\# \chi_i \text{ in } (x-h, x+h]}{2nh} \]

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

  → 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 \( K(\cdot) \) include:
  \[ \text{Normal, Epanechnikov} \]

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

![Graph showing kernel density estimate with uniform kernel function.](image)
• 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**
  \[ \rightarrow \text{more sample values contribute to } \hat{f}(x) \]
  \[ \rightarrow \text{estimate is smoother overall} \]

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

• 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 varies over the
  region (being larger where the density is expected
to be flat and smaller where the density is
  expected to have bumps).