- Histograms
- KDE
- ScatterPlot and multidimensional graphs

Histogram Example

Example: Plot the Frequency, Relative Frequency and Density Histograms for

$$0,\ 4,\ 2,\ 2,\ 0,\ 0.5,\ 1,\ 3$$

Now let us change the default bins for a Histogram. We can use the following - first define the vector of our class interval (Bins) endpoints: (note that you need to cover all Datapoints!)

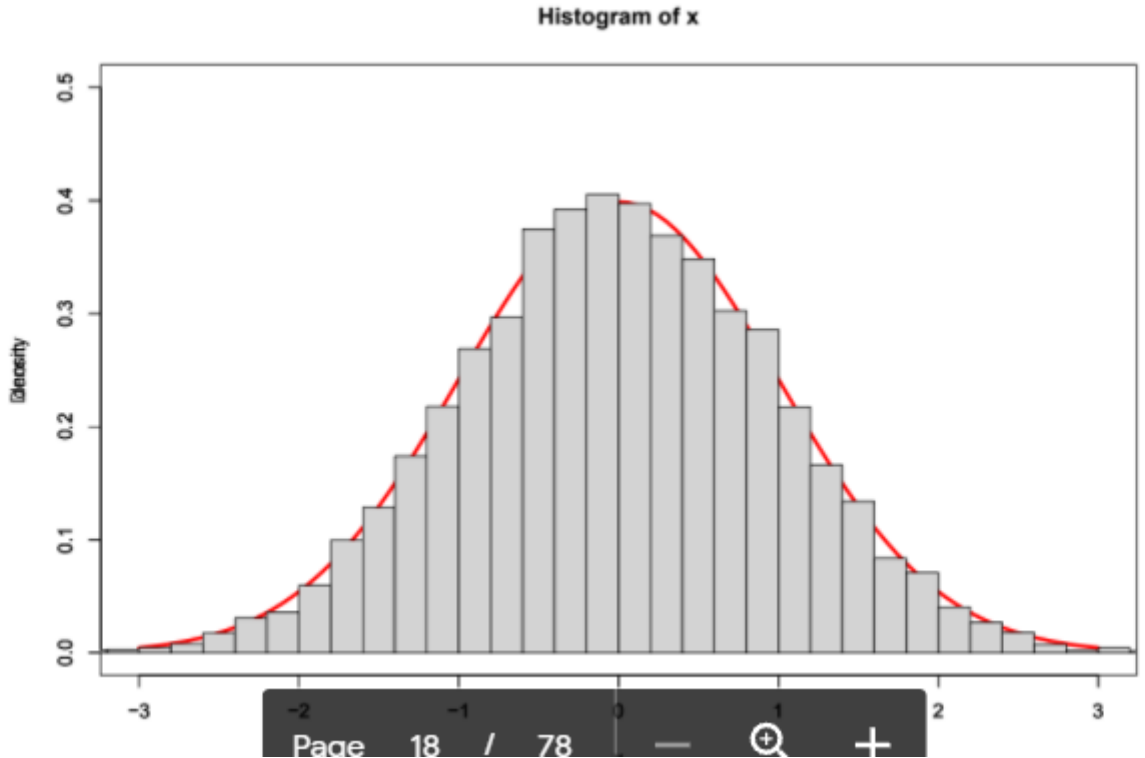
```
bins.endpoitns <- c(50, 65, 75, 90, 100)
hist(airquality$Temp, breaks = bins.endpoitns)
```

- By default, if we give custom bins with non-equal lengths, **R** is plotting the Density Histogram!

Estimation of the PDF through the Density Histogram

As it was stated above, the Density Histogram is an approximation (estimate) of the PDF of the Data unknown Distribution. To check this, let us take a synthetic Dataset from the Distribution we know:

```
plot(dnorm, lwd = 3, col= "red", xlim=c(-3,3), ylim=c(0,0.5))
x <- rnorm(10000)
par(new = TRUE)
hist(x, breaks = 40, freq = FALSE, xlim=c(-3,3), ylim=c(0,0.5))
```



Choosing Bin sizes correctly

It is important to choose the Bin sizes (lengths of the Bin, class, intervals) wisely. Otherwise you will skip some info or you will not get any valuable info.

Let us use another **R** standard dataset to show the effect of the choice of the bin size: *precip*. This Dataset shows the average amount of precipitation (rainfall) in inches for each of 70 United States (and Puerto Rico) cities.

```
head(precip)
```

```
##      Mobile      Juneau     Phoenix Little Rock Los Ange
##        67.0        54.7         7.0        48.5        1
```

Differences between the Barplot and Histogram

- Can you give some differences?

Here are some:

- *Barplot*'s rectangles widths are arbitrary, do not mean anything, rectangles are not adjacent; *Histogram*'s rectangles are adjacent, and the choice of the Bin widths is changing the graph
- *Barplot* is for a categorical or Discrete Data, *Histogram* is for both Discrete and Continuous
- We can exactly reconstruct the Dataset from the *Barplot*, but not the *Histogram*

What we can see from the Histogram

If we will not look at the Histogram as being an estimate for the unknown Distribution behind the Data, and if we will just try to get some info about our Dataset, Histogram is helping us to say if the Data:

- is symmetric about some point or is skewed to the left or right
- is spread out or concentrated at some point
- has some gaps
- has values far apart from others, has outliers (anomalies)
- is unimodal, bimodal or multimodal

KDE

Another estimate for the unknown Distribution PDF is the **Kernel Density Estimator**, KDE. It is, in some sense, the smoothed version of the Histogram: Histogram is a piecewise-constant function, with jumps, so it is not a smooth function.

KDE

To define the KDE, we first choose a smooth Kernel function $K(t)$, here, a function with

$$K(t) \geq 0, t \in \mathbb{R}, \qquad \text{and} \qquad \int_{-\infty}^{+\infty} K(t)dt = 1.$$

For example, we can take the Gaussian Kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}, \qquad t \in \mathbb{R},$$

or any other PDF.

Next, one defines the Kernel Density Estimator with Kernel K as

$$KDE_K(x) = KDE(x) = \frac{1}{nh} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

KDE

It is easy to see that $KDE(x)$ will give a function satisfying the properties of the PDF, i.e., will be nonnegative and will integrate to 1:

$$\int_{-\infty}^{+\infty} KDE(x)dx = \frac{1}{nh} \cdot \sum_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - x_i}{h}\right) dx =$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - x_i}{h}\right) d\frac{x - x_i}{h} \overset{u=\frac{x-x_i}{h}}{=}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \int_{-\infty}^{+\infty} K(u)du = \frac{1}{n} \cdot \sum_{i=1}^{n} 1 = 1.$$

KDE

Note: Like in the case of the Density histogram, where that histogram was depending on the bins choice, the KDE depends on the choice of $h > 0$. h is called the **bandwidth**, and its estimation is another story.

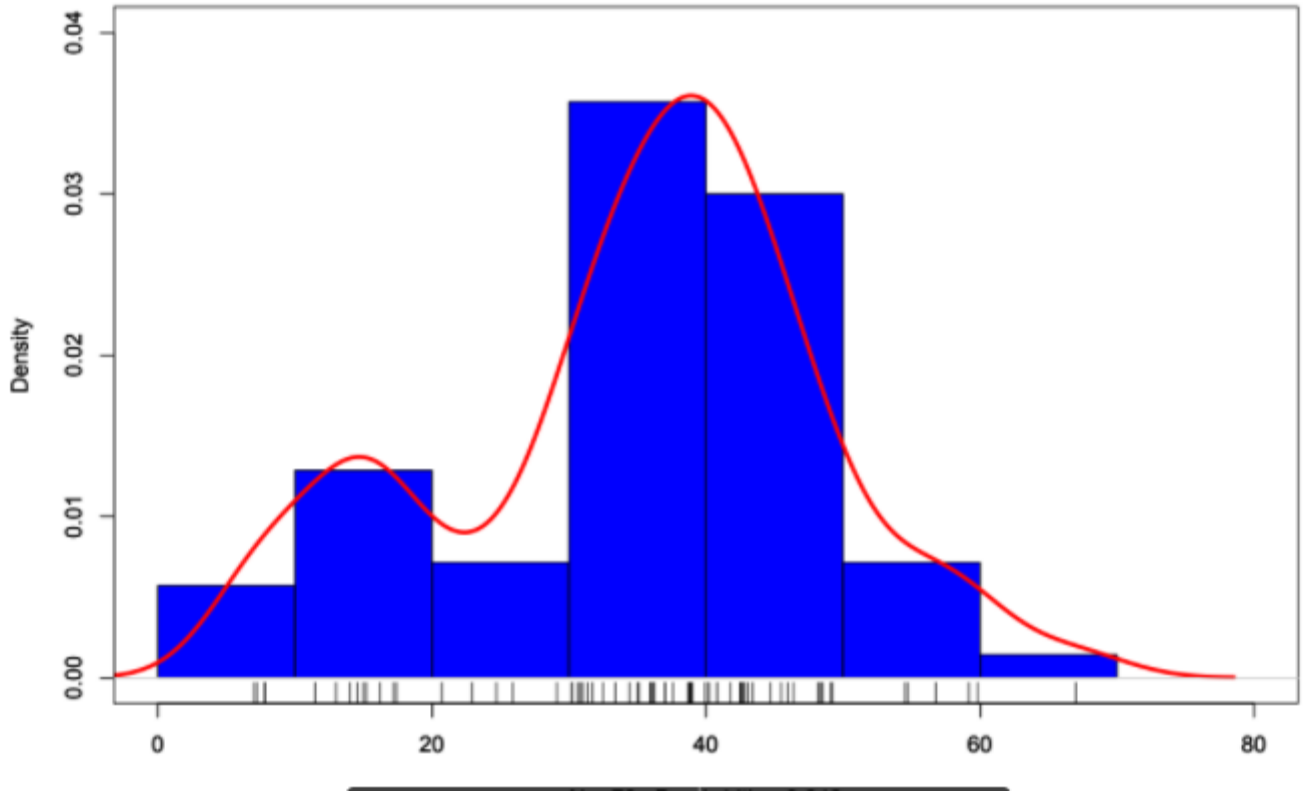
Note: One can prove that under some conditions, KDE is approximating well the unknown PDF behind the data. In fact,

Theorem: Assume we are constructing the $KDE = KDE(\cdot, h_n)$ for the IID r.v $X_1, X_2, ..., X_n$, coming from an unknown PDF f, and with the bandwidth h_n. If the PDF f is continuous at the point x, and if $h_n \to 0$ and $n \cdot h_n \to \infty$, then

$$KDE(x, h_n) \to f(x) \qquad in \quad \mathbb{P}.$$

KDE Example

```
x <- precip; d <- density(x)
hist(x, freq = FALSE, xlim = c(0, 80), ylim = c(0,0.04),
     col = "blue", main = "")
rug(x); par(new = TRUE)
plot(d, lwd = 3, col = "red", xlim = c(0,80), ylim = c(0,0.04),
     main = "")
```



KDE, how to choose the bandwidth

Here is the idea behind the theory of choosing the bin size for the Histogram, and bandwidth for KDEs.

Assume we have an *Estimator* for the unknown PDF $f(x)$ behind IID data $X_1, X_2, ..., X_n$: we denote that Estimator by $\hat{f}_n(x)$. We assume that $\hat{f}$ depends on some parameter h (smoothing parameter), and we want to find some *best value* for h. We define the **Risk** of Estimating f through $\hat{f}$, $Risk(\hat{f}, f)$. One of the standard ways to define the Risk is to choose the Mean Integrated Squared Error:

$$Risk(\hat{f}, f) = MISE(\hat{f}, f) = \mathbb{E} \int_{\mathbb{R}} (\hat{f}(x) - f(x))^2 dx.$$

Now, the idea is to choose h in such a way that the Risk of $\hat{f}$ will be the minimal.