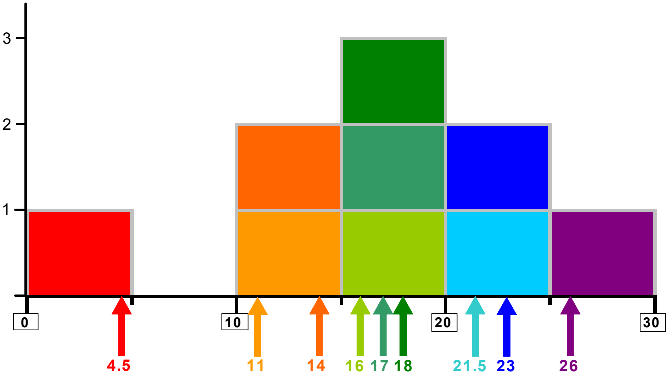
We create a **histogram** to graphically summarize the distribution of the data set

4.5,11,14,16,17,18,21.5,23,26.

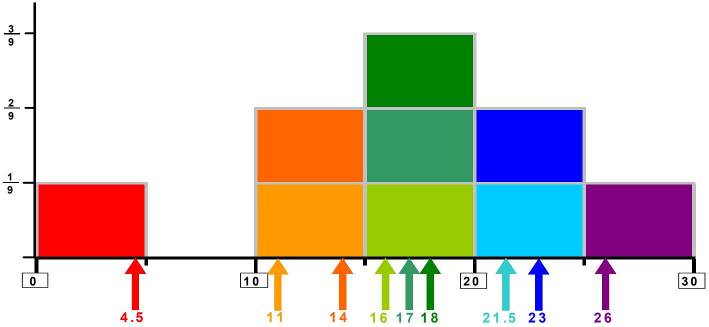
It shows the number of values that fall into each of the class intervals (or bins)

(0,5], (5,10], (10,15], (15,20], (20,25], (25,30].

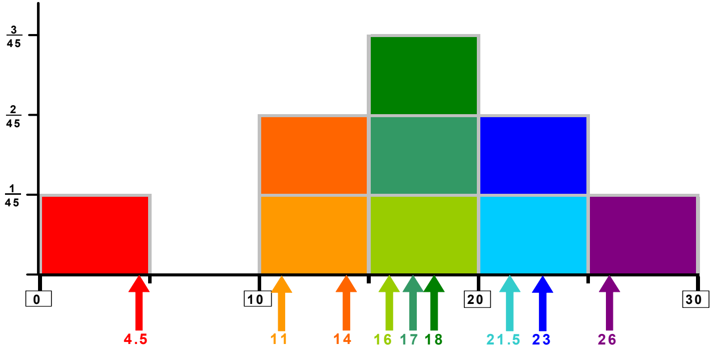


A histogram is a discontinuous function.
It inherits its jumps from its rectangular building blocks.

The **relative frequency histogram** and the **density histogram** are normalized variants of the histogram. In the case of the relative frequency histogram, the heights of all histogram bars sum to one.

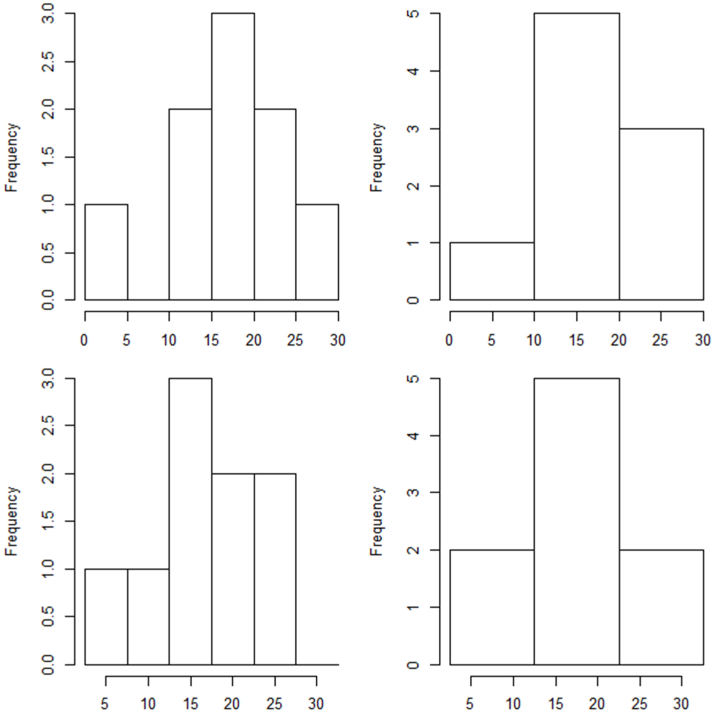


In the case of the density histogram, the histogram has a total area of one.

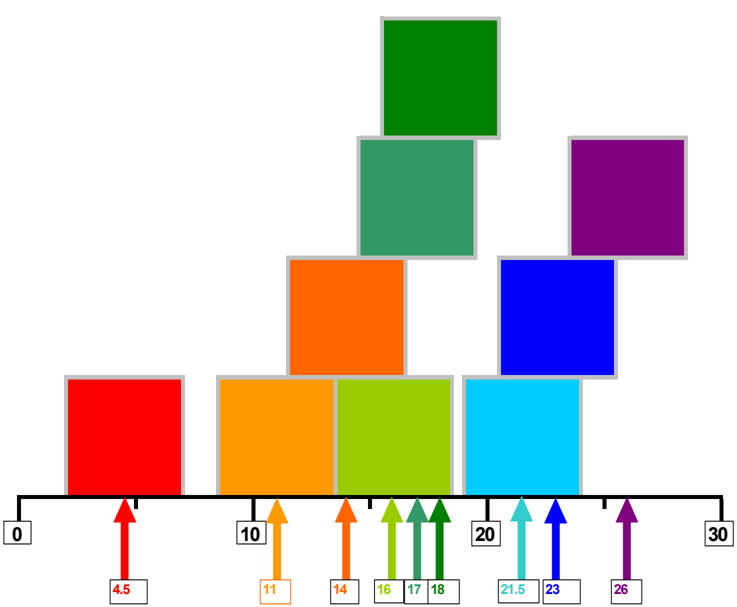


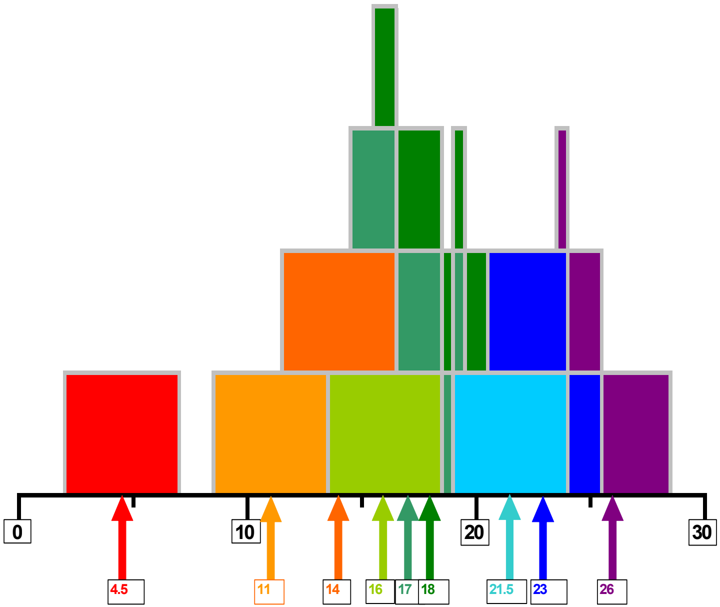
The appearance of a histogram depends strongly on the origin and the width of the class intervals.

```
x <- c(4.5,11,14,16,17,18,21.5,23,26) # data
par(mfrow=c(2,2),mar=c(2,4,1,1))
# subsequent figures in 2x2 array; narrow margins

hist(x,breaks=seq(0,30,5),right=TRUE,main="")
# right-closed (left-open) intervals: (0,5],...,(25,30]

hist(x,breaks=seq(0,30,10),right=TRUE,main="")

hist(x,breaks=seq(2.5,32.5,5),right=TRUE,main="")

hist(x,breaks=seq(2.5,32.5,10),right=TRUE,main="")
```

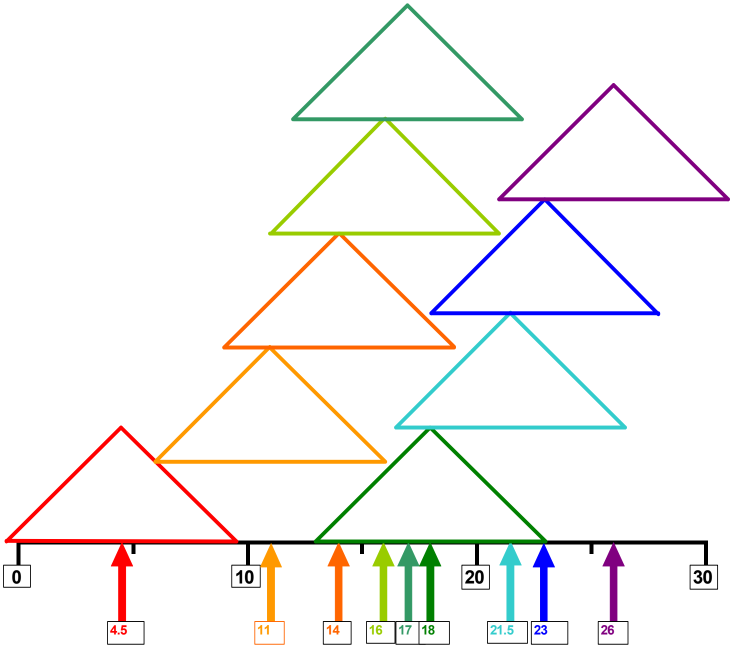


The dependency of a histogram on the origin of the bins can be removed if the rectangular boxes are centered around the individual data points.

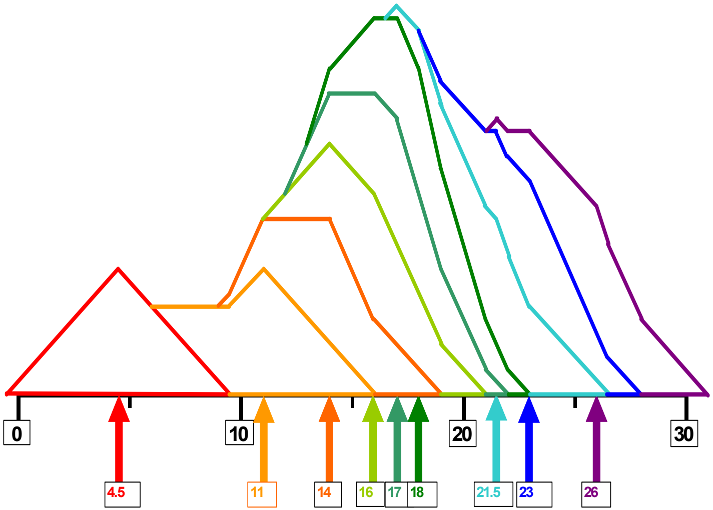




To avoid jumps we may use triangular weights.



The resulting function is continuous.
But it is not differentiable because there are kinks.



Basic rectangular and triangular functions are given by

$$w_R(s)=\begin{cases}\frac{1}{2},\ if\ |s|\le 1,\\ 0,\ otherwise\end{cases}$$

and

$$w_T(s)=\begin{cases}1-|s|,\ if\ |s|\le 1,\\ 0,\ otherwise,\end{cases}$$

respectively. Both functions are non-negative and integrate to one, hence they can be regarded as density functions.

We have:

$$\mu_R = \int_{-\infty}^{\infty} s\, w_R(s) ds = \int_{-\infty}^{-1} 0\, ds + \int_{-1}^{1} \tfrac{s}{2} ds + \int_{1}^{\infty} 0\, ds = \frac{1^2}{4} - \frac{(-1)^2}{4} = 0\,,$$

$$\mu_T = \int_{-1}^{0} s(1+s)ds + \int_{0}^{1} s(1-s)ds = -\frac{(-1)^2}{2} - \frac{(-1)^3}{3} + \frac{1^2}{2} - \frac{1^3}{3} = 0\,,$$

$$\sigma_R^2 = \int_{-\infty}^{\infty} s^2 w_R(s)ds - \mu_R^2 = \int_{-1}^{1} \frac{s^2}{2} ds = \frac{1^3}{6} - \frac{(-1)^3}{6} = \frac{1}{3}\,,$$

$$\sigma_T^2 = \int_{-1}^{0} s^2(1+s)ds + \int_{0}^{1} s^2(1-s)ds = -\frac{(-1)^3}{3} - \frac{(-1)^4}{4} + \frac{1^3}{3} - \frac{1^4}{4} = \frac{1}{6}$$

The rescaled density functions

$$w_r(s)=\begin{cases}\frac{\sqrt{3}}{2},\ if\ |s|\le 1,\\ 0,\ otherwise\end{cases}$$

and

$$w_t(s)=\begin{cases}\sqrt{6}(1-|s|),\ if\ |s|\le 1,\\ 0,\ otherwise\end{cases}$$

imply zero means and unit variances.

Given a standardized density $w(s)$, the density

$$v(u)= w(\tfrac{u-x}{h})\tfrac{1}{h}$$

obtained via the transformation

$$s \rightarrow hs+x$$

is centered around x and implies a standard deviation of h.

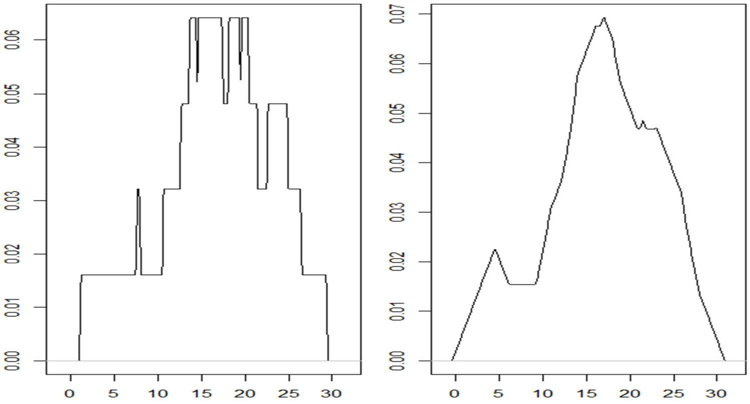
If $x_1,\ldots,x_n$ is a random sample from an unknown density f,

$$\hat{f}\,(\mathrm{s})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w(\tfrac{s-x_i}{h})$$

is called a **kernel density estimator** of f, the function w is called the **kernel**, and the parameter h is called the **bandwidth**.

We use the R function density for density estimation.

```
par(mfrow=c(1,2),mar=c(2,2,1,1))
plot(density(x,kernel="rectangular",bw=2),main="")
plot(density(x,kernel="triangular",bw=2),main="")
```

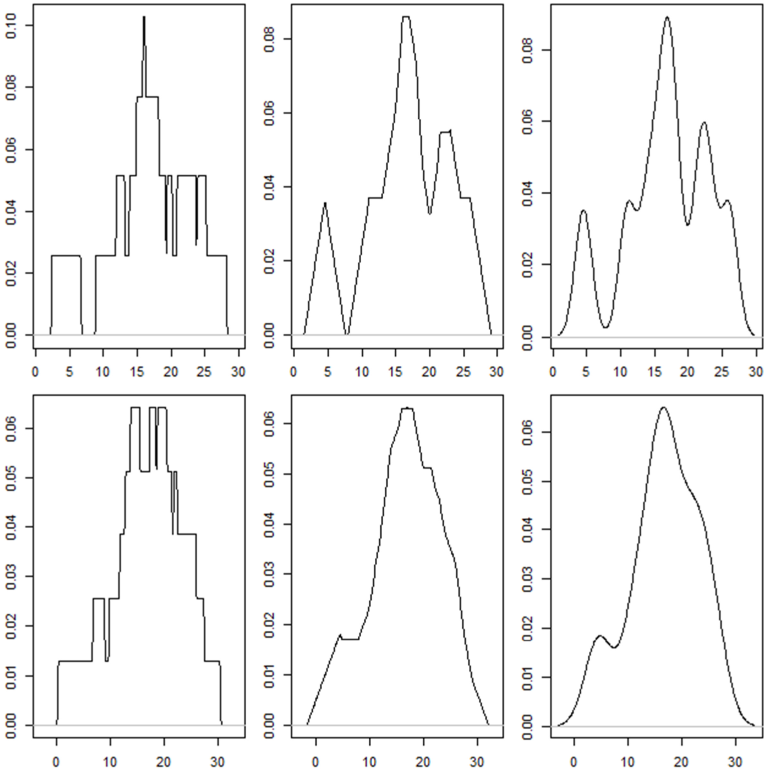


A kernel density estimator inherits its smoothness properties from the kernel. The R function density allows also the choice of smooth kernels such as the Gaussian kernel

$$w(s)=\frac{1}{\sqrt{2\pi}}e^{-\frac{s^2}{2}}.$$

The bandwidth h controls the resolution. The larger h, the lower the resolution. The choice of the bandwidth h is by far more important than that of the kernel w.

```
par(mfrow=c(2,3),mar=c(2,2,1,1))
for (b in c(1.25,2.5)) {
plot(density(x,kernel="rectangular",bw=b),main="")
plot(density(x,kernel="triangular",bw=b),main="")
plot(density(x,kernel="gauss",bw=b),main="") }
```



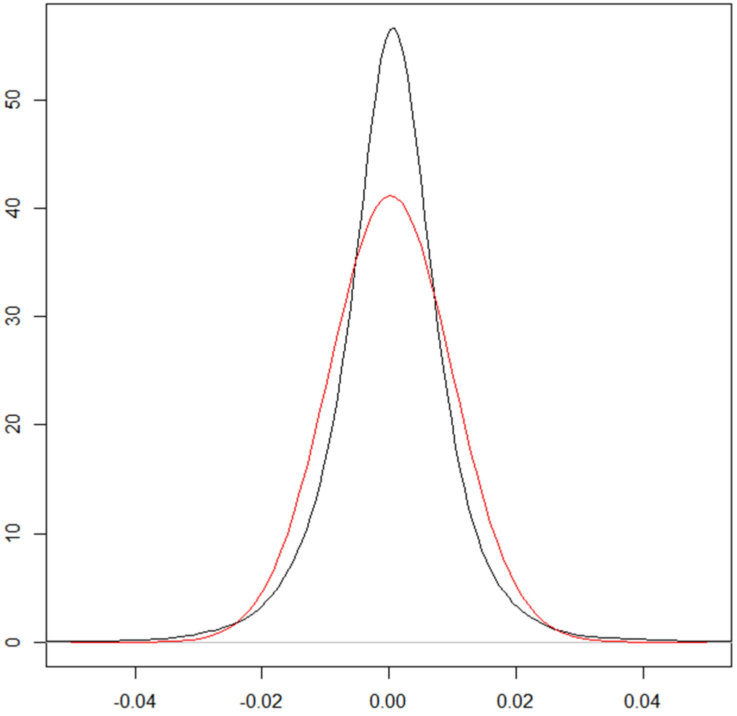
Estimating the density of index returns

We download the daily historical data of the S&P 500 Index from Yahoo!Finance as a csv file ^GSPC.csv into our working directory C:\SP500, import the data into R, and calculate the log returns.

```
setwd("C:/SP500")  # R uses / as path separator
Y <- read.csv("^GSPC.csv",header=T,na.strings="null")
# in the downloaded file, missing values are represented
# by the string "null" rather than by the symbol NA
Y <- na.omit(Y)  # rows with missing values are omitted
N <- nrow(Y); D <- as.Date(Y[,1])  # dates in column 1
cl <- log(Y[,6])  # adjusted close prices in column 6
r <- cl[2:N]-cl[1:(N-1)]; n <- N-1# n (log) returns
```

We use the Gaussian kernel to estimate the density of the returns and compare the estimated density of the returns with a normal density with the same mean and variance.

```
par(mfrow=c(1,1),mar=c(2,2,1,1))
x <- seq(-0.05,0.05,0.001); R <- range(x)
plot(density(r,kernel="g",bw=0.0025),xlim=R,main="")
lines(x,dnorm(x,mean=mean(r),sd=sd(r)),col="red")
```



Determining the tail behavior

If the inverse function of a distribution function F: $\mathbb{R} \to [0,1]$ exists and $q \in (0,1)$, then the value π_q=F$^{-1}(q)$ is called the ***q*-quantile**.

Clearly, π_q satisfies $F(\pi_q)=F(F^{-1}(q))=q$.

To check whether a sample $x_1,\dots,x_n$ comes from a specified theoretical distribution function F we might plot the "sample quantiles"

$$\hat{\pi}_{\frac{1}{n+1}}=x_{(1)},\ \hat{\pi}_{\frac{2}{n+1}}=x_{(2)},\ \dots,\ \hat{\pi}_{\frac{n}{n+1}}=x_{(n)}$$

against the corresponding theoretical quantiles

$$\pi_{\frac{1}{n+1}}=F^{-1}(\tfrac{1}{n+1}),\ \pi_{\frac{2}{n+1}}=F^{-1}(\tfrac{2}{n+1}),\ \dots,\ \pi_{\frac{n}{n+1}}=F^{-1}(\tfrac{n}{n+1})$$

of F, where

$$x_{(1)}\le x_{(2)}\le\dots\le x_{(n)}$$

is the sample arranged in ascending order.

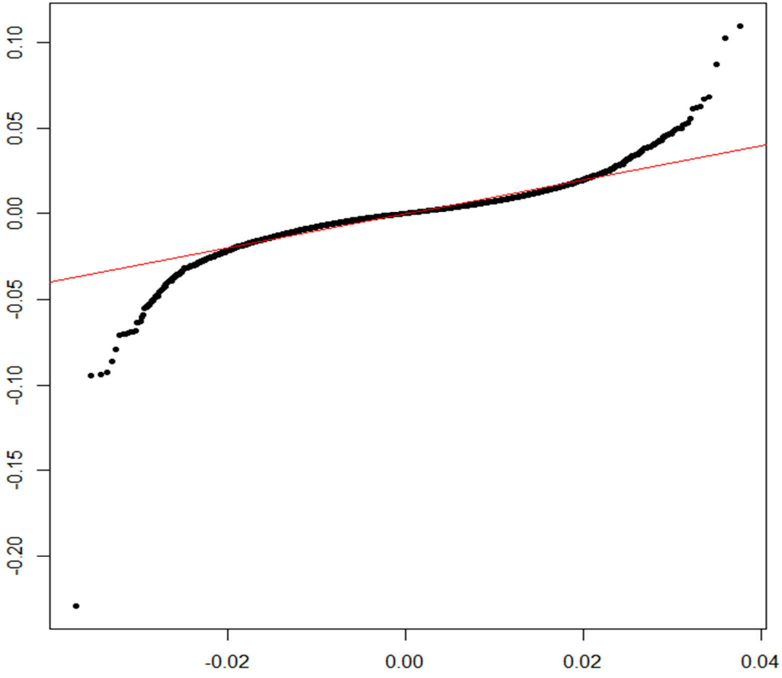
If the sample quantiles are approximately of the same size as the theoretical quantiles, the points

$$(F^{-1}(\tfrac{i}{n+1}),x_{(i)}),\ i=1,\dots,n,$$

should roughly lie on a straight line with intercept zero and slope one.

We create a **quantile-quantile (Q-Q) plot** to "test" the returns of the S&P 500 Index against a normal distribution with the same mean and variance.

```
par(mfrow=c(1,1),mar=c(2,2,1,1),pch=20)
q <- (1:n)/(n+1)
plot(qnorm(q,mean=mean(r),sd=sd(r)),sort(r))
abline(a=0,b=1,col="red")
```


0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.02
0.00
0.02
0.04