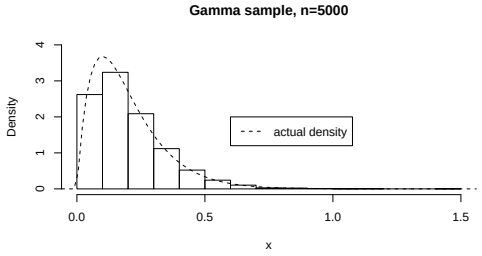
Histogram as Density Estimate

```
> x <- rgamma(5000,shape=2,rate=10)
> hist(x,freq=F,breaks=20,main="Gamma sample, n=5000")
> curve(dgamma(x,shape=2,rate=10),lty=2,add=T)
> legend(0.6,2,c("actual density"),lty=2)
```

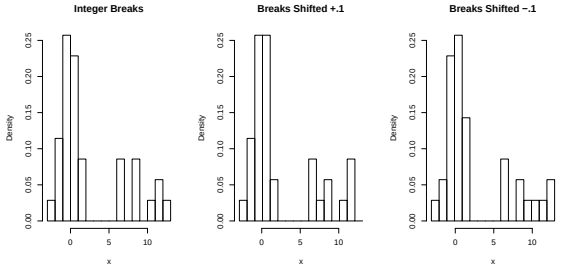


Pros:

- area under curve is 1;
- area between $x = a$ and $x = b$ is, roughly, proportion of observations in $[a, b)$ and so an estimate of $\Pr(a \leq X \leq b)$;
- looks like a pretty good estimate for *big* samples.

Histogram as Density Estimate

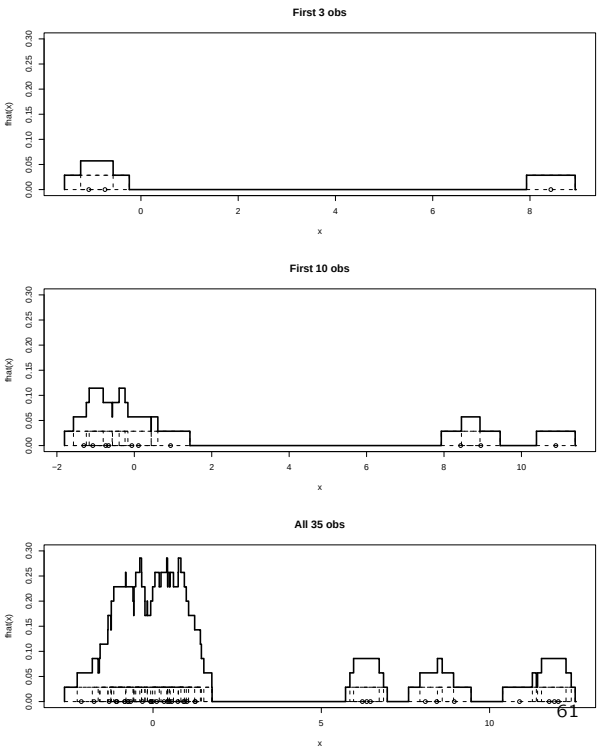
```
> x <- sample(c(rnorm(25,0,1),rnorm(10,10,2)))
> range(x)
[1] -2.129958 12.042381
> breaks <- seq(-3,13,by=1)
> opar <- par(mfrow=c(1,3))
> hist(x,freq=F,breaks=breaks, main="Integer Breaks")
> hist(x,freq=F,breaks=breaks+.1, main="Breaks Shifted +.1")
> hist(x,freq=F,breaks=breaks-.1, main="Breaks Shifted -.1")
> par(opar)
```



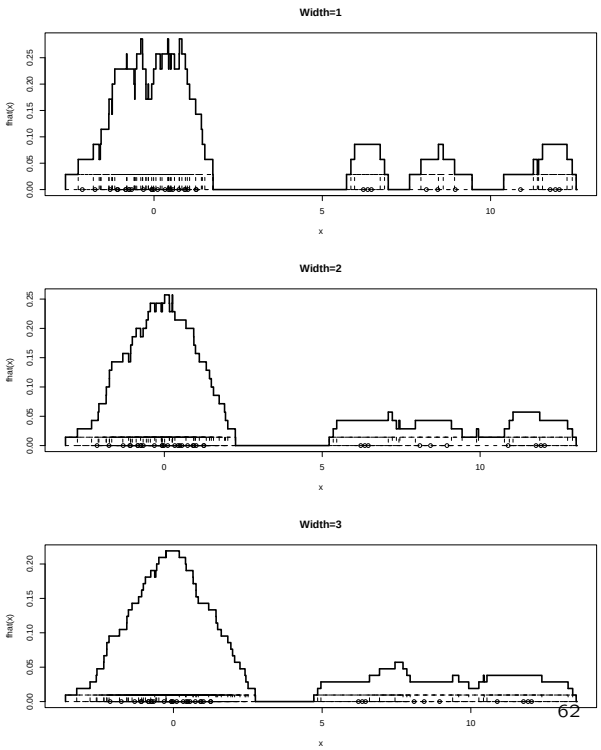
Cons:

- discontinuous and just plain ugly;
- very sensitive to choice of breaks;
- generally poor estimate for moderate sample sizes.

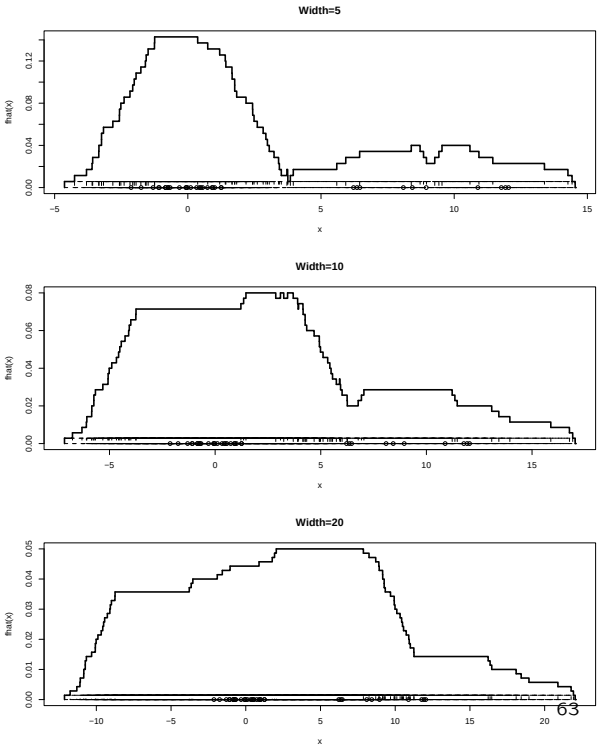
Adding up Little Rectangles



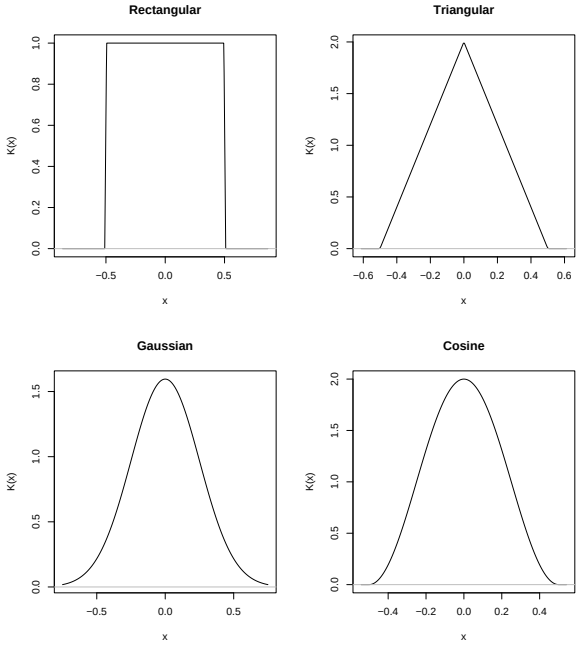
Using Different-Width Rectangles



Using Different-Width Rectangles



Some Different Kernels



The R `density` Function

`density(x, bw="nrd0", adjust=1, kernel="gaussian")`

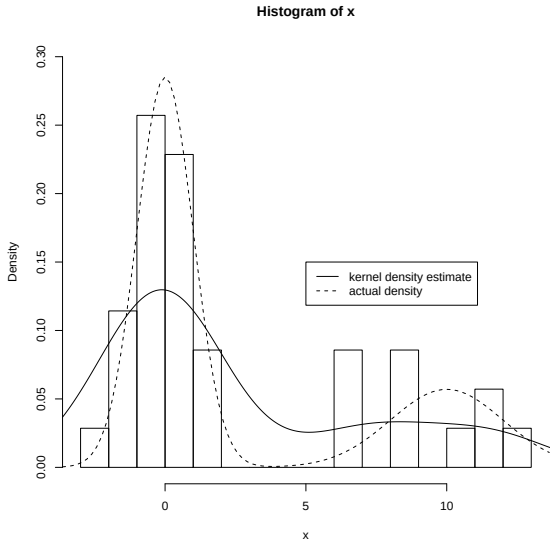
- `kernel` can be `rectangular`, `triangular`, `gaussian`, or `cosine` (or even `epanechnikov`, `biweight`, or `optcosine`);
- `bw` can be (a) a number, and the kernel will be scaled to have this standard deviation; or (b) the name of a special bandwidth-choosing algorithm: `nrd0`, `nrd`, `ucv`, `bcv`, `sj`, or `sj-dpi`;
- `adjust` is a constant to scale a bandwidth chosen by an algorithm.

Note the function returns a `density` object with the most useless default output possible:

```
> density(x)

Call:
     density(x = x)

Data: x (35 obs.); Bandwidth 'bw' = 1.984

       x                y
 Min.   :-8.082   Min.   :0.0001709
 1st Qu.:-1.563   1st Qu.:0.0094211
 Median : 4.956   Median :0.0292758
 Mean   : 4.956   Mean   :0.0383052
 3rd Qu.:11.475   3rd Qu.:0.0456803
 Max.   :17.994   Max.   :0.1297325
>
```

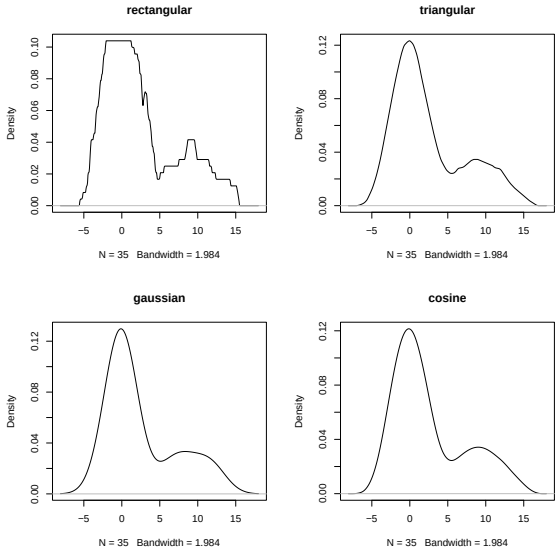

Superimposed on a Histogram

```
> hist(x, freq=F, nclass=20, ylim=c(0,.3))
> lines(density(x))
> curve(25/35*dnorm(x,0,1)+10/35*dnorm(x,10,2), lty=2, add=T)
> legend(5,.15,c("kernel density estimate","actual density"),
+ lty=c(1,2))
```



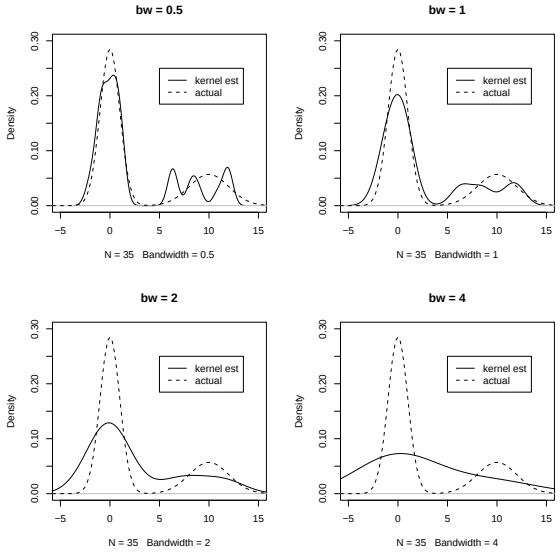
Effect of Kernel

All plots with bandwidth `bw` = 1.984 as chosen by `nrd0`.



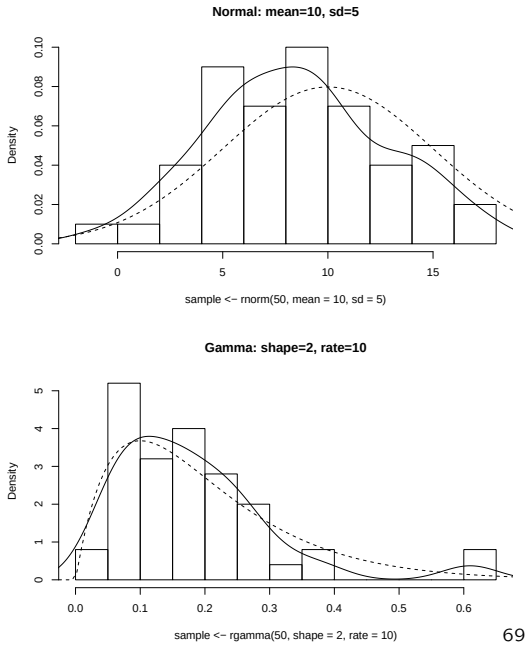
Effect of Bandwidth

All plots with default Gaussian kernel.



Default Bandwidth

The `nrd0`-chosen bandwidth works well with unimodal distributions and reasonable numbers of observations (here, $n = 50$). Solid is fitted, dashed is actual.



Sheather & Jones Bandwidth