Statistics II: Histograms in R

Richard Gill*

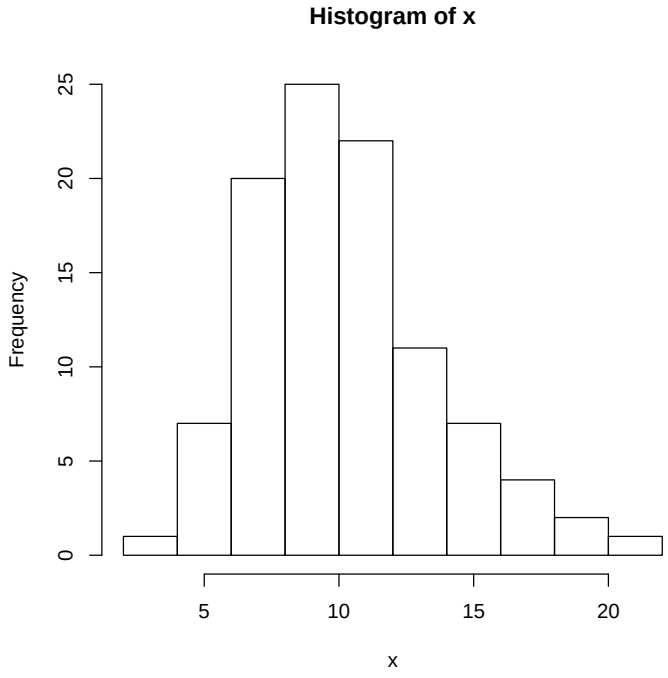
March 5, 2009

These scripts and notes illustrate the histogram as a density estimator. First we make some data. And before that, I set the random seed, so that the results will be perfectly reproducible.

```
> set.seed(11091951)
> x <- rgamma(100, 10)
```

Now we show the plain vanilla R histogram.

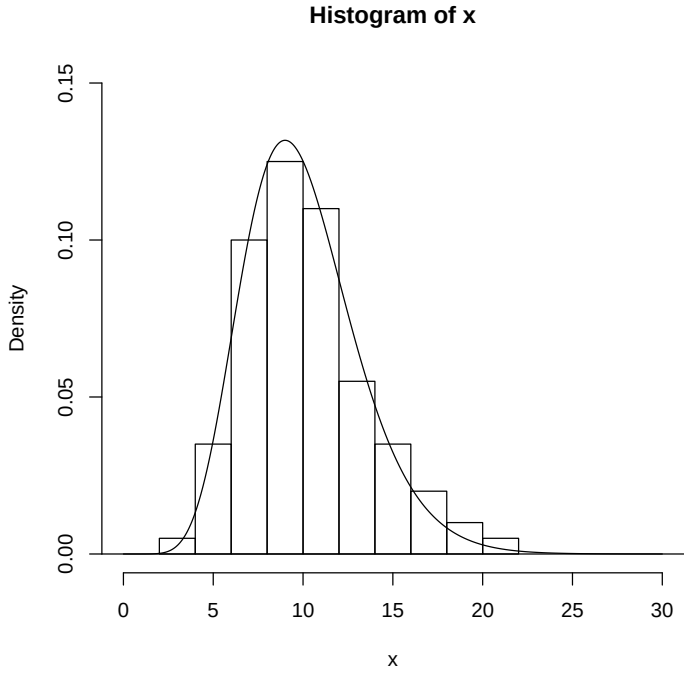
```
> hist(x)
```



*http://www.math.leidenuniv.nl/∼gill/teaching/statistics

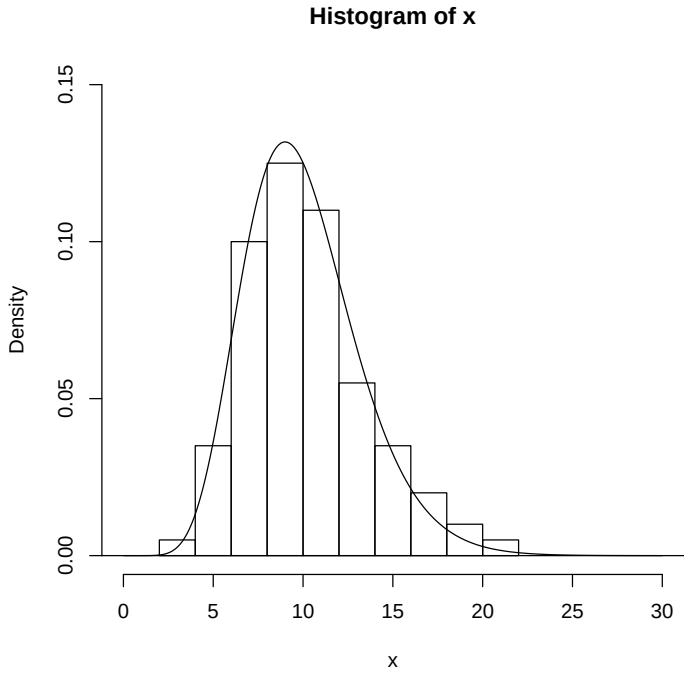
The area under the graph was equal to the number of observations. Using the option prob=TRUE we make the area under the graph equal to one. On this graph I will superimpose a plot of the true density of the data. I'll also fix the axes to get a nicer plot. The bin-width is determined by Sturges' method, the default, it doesn't hurt to make that explicit.

```
> hist(x, prob = T, xlim = c(0, 30), ylim = c(0, 0.15), breaks = "Sturges")
> xd <- seq(from = 0, to = 30, length = 1000)
> yd <- dgamma(xd, shape = 10)
> lines(xd, yd)
> abline(h = 0)
```



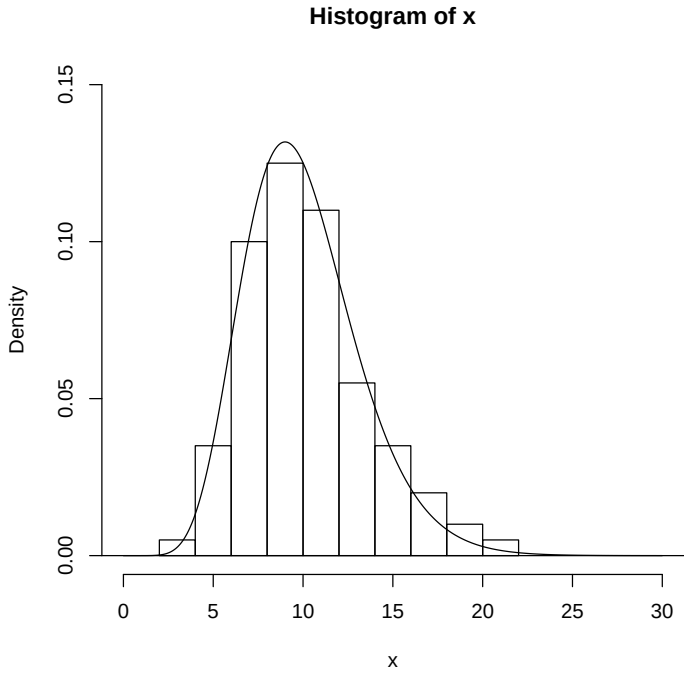
Now let's look at the results of the other bin-width algorithms. First, Scott:

```
> hist(x, prob = T, xlim = c(0, 30), ylim = c(0, 0.15), breaks = "Scott")
> lines(xd, yd)
> abline(h = 0)
```



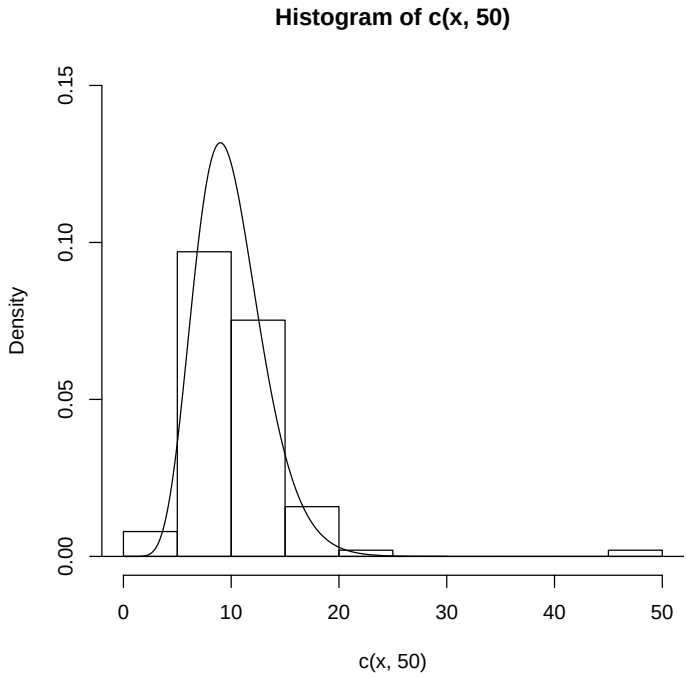
And Freedman-Diaconis:

```
> hist(x, prob = T, xlim = c(0, 30), ylim = c(0, 0.15), breaks = "FD")
> lines(xd, yd)
> abline(h = 0)
```



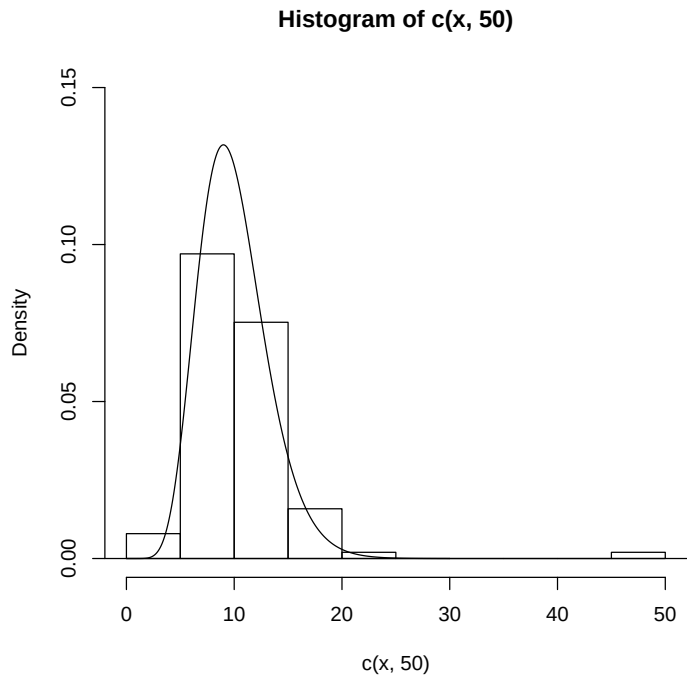
Not much to see, right? But I'll draw these all again, but now with an outlier added to the dataset. First, plain vanilla (Sturges):

```
> hist(c(x, 50), prob = T, xlim = c(0, 50), ylim = c(0, 0.15),
+     breaks = "Sturges")
> lines(xd, yd)
> abline(h = 0)
```



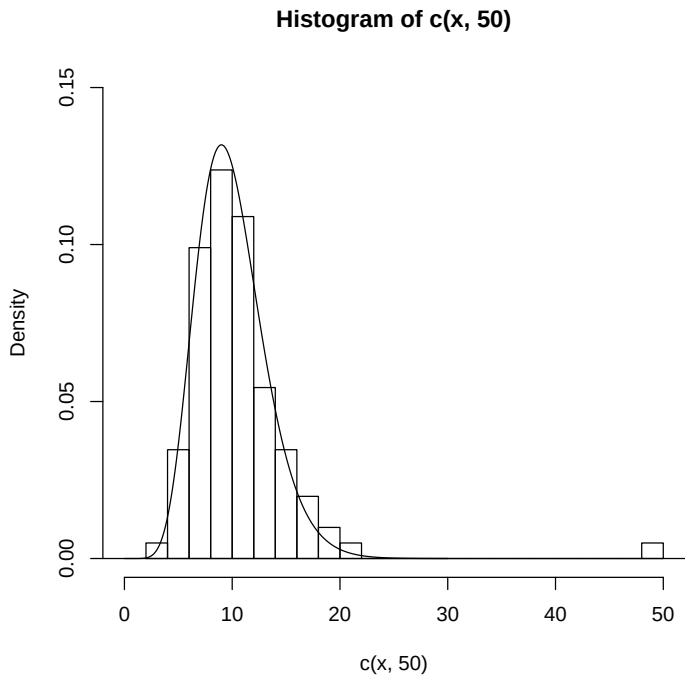
Then Scott:

```
> hist(c(x, 50), prob = T, xlim = c(0, 50), ylim = c(0, 0.15),
+     breaks = "Scott")
> lines(xd, yd)
> abline(h = 0)
```



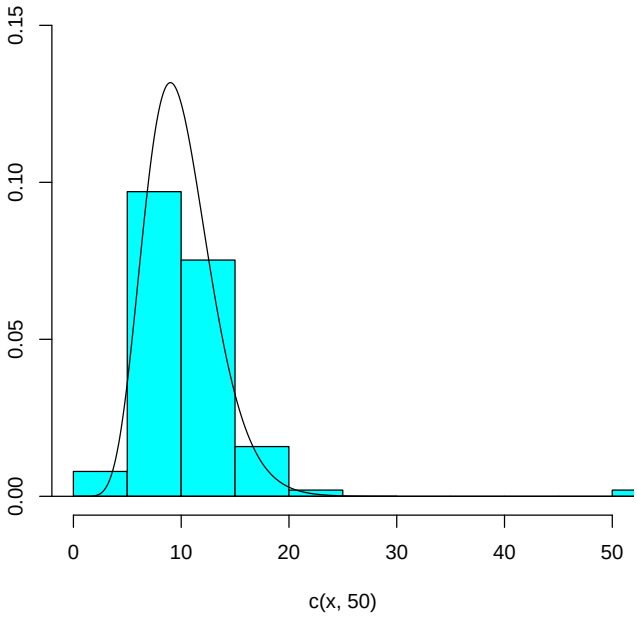
And Freedman-Diaconis

```
> hist(c(x, 50), prob = T, xlim = c(0, 50), ylim = c(0, 0.15),
+     breaks = "Freedman-Diaconis")
> lines(xd, yd)
> abline(h = 0)
```



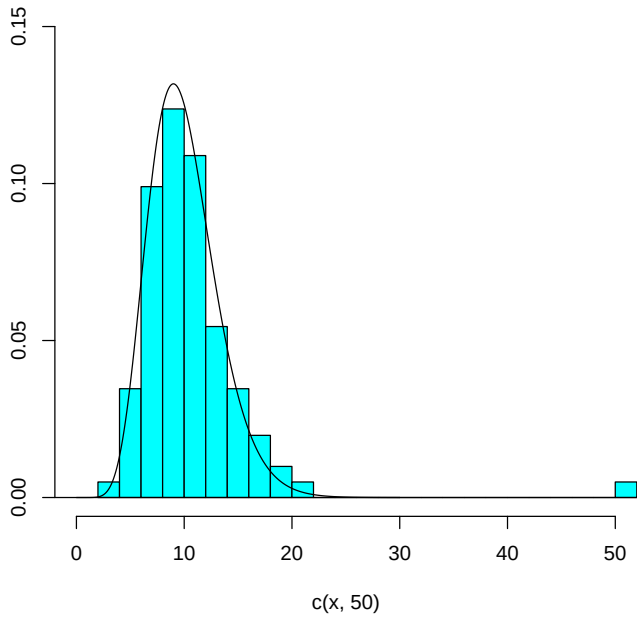
Venables and Ripley think they have a much better histogram. Let's take a look. Scott:

```
> library(MASS)
> truehist(c(x, 50), prob = T, xlim = c(0, 50), ylim = c(0, 0.15),
+     nbins = "Scott")
> lines(xd, yd)
> abline(h = 0)
```



And Freedman-Diaconis

```
> truehist(c(x, 50), prob = T, xlim = c(0, 50), ylim = c(0, 0.15),
+     nbins = "Freedman-Diaconis")
> lines(xd, yd)
> abline(h = 0)
```



Exercise. Figure out how to draw a frequency polygon (join the mid-points of histogram bars by straight lines...). Theoretically you can use a larger bin-width, at least, if the true density is twice differentiable.