[Re] Local alignment statistics - EVD data exploration

Nathan Brouwer

11/18/2019

Data exploration

Sample analysis and exploration of n = m = 191 data

Load the data

```
random.scores.191 <- read.csv(file = "random_scores_191.csv")
```

Distribution of scores

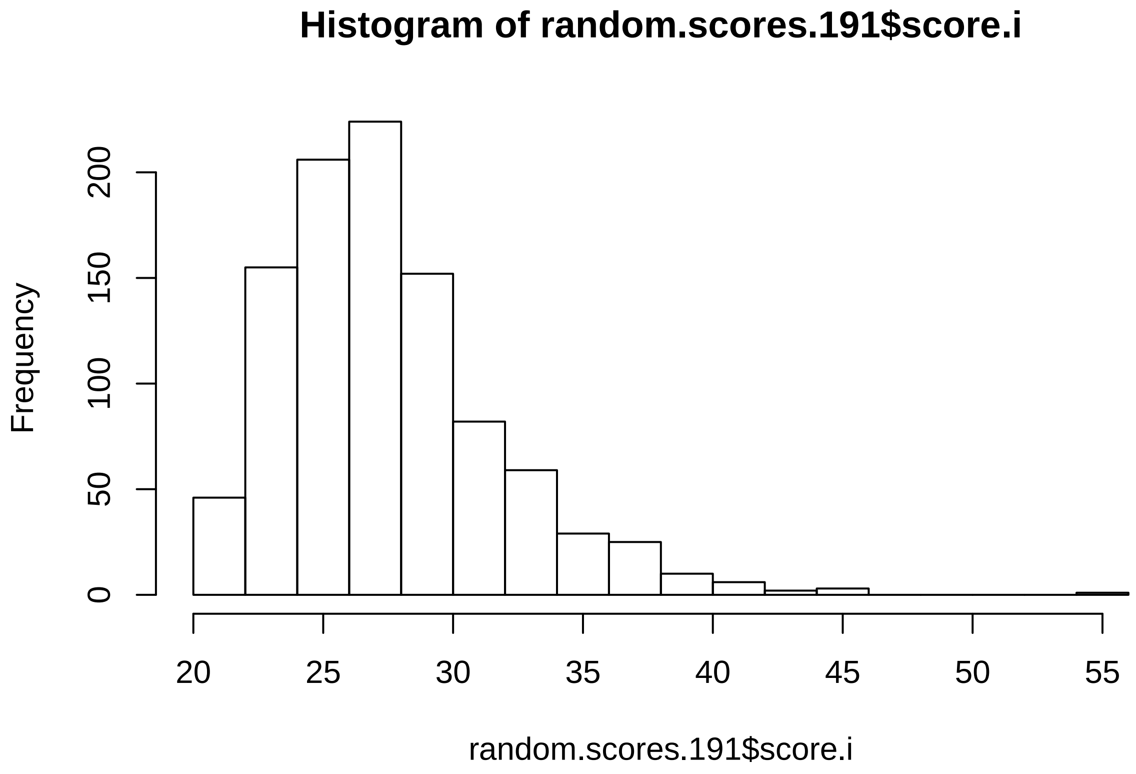
The scores produced by the simulation are similar in shape to an extreme value distribution (EVD)

```
head(random.scores.191$score.i)
```

```
## [1] 28 29 23 27 30 26
```

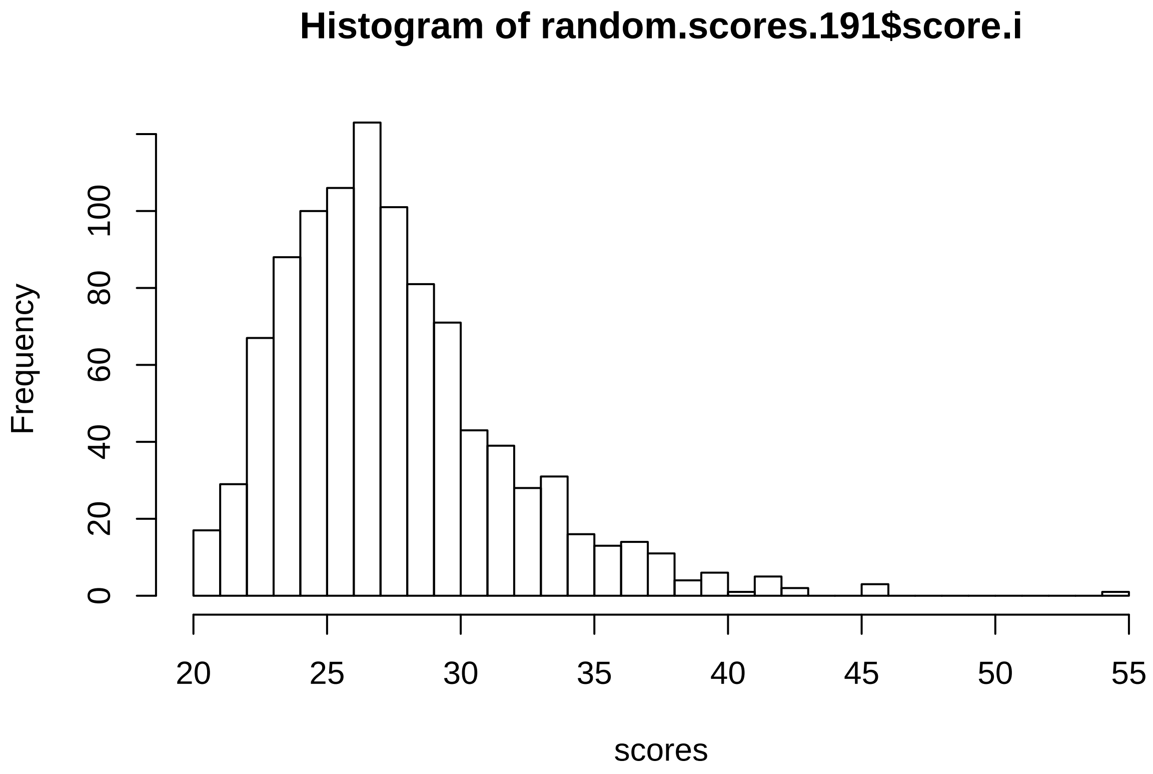
We can make the graph a bit smoother by setting breaks = 20, which determines the number of bars in the histogram

```
hist(random.scores.191$score.i,
     breaks = 20)
```



If we want we can make R give each score its own bar on the histogram. The code below does this, but is a bit dense and I won't explain it

```
x <- range(random.scores.191$score.i)
bins <- x[2]-x[1]+1
hist(random.scores.191$score.i,
     breaks = bins,
     xlab = "scores")
```



Determining the mode

The Gumbel extreme value distribution has two parameters: mu and lambda.

mu is defined as the highest point of the distribution. This is similar to the **mode** of the distribution. We can therefore approximate mu by calculate the mode. We can get the mode using the table() function, followed by some processing

Make a table of the scores

```
table.i <- table(random.scores.191$score.i)
table.i
```

```
##
##  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37
##   4  13  29  67  88 100 106 123 101  81  71  43  39  28  31  16  13  14
##  38  39  40  41  42  43  46  55
##  11   4   6   1   5   2   3   1
```

I can use which.max() to figure out which element of the table has the highest value.

```
i.max <- which.max(table.i)
table.i[i.max]
```

```
##  27
## 123
```

The scores are actually scored as the names of the table elements. I can get them using names()

```
mode.i <- names(table.i[i.max])
mode.i
```

```
## [1] "27"
```

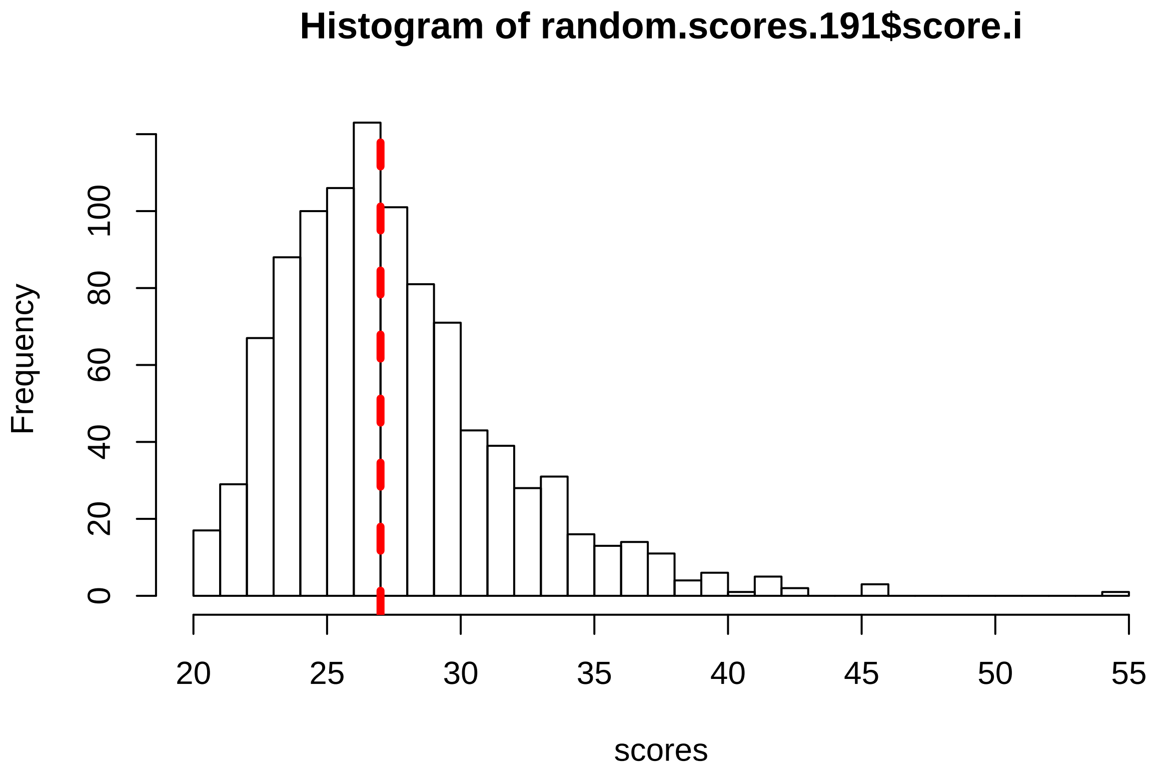
Names are character data so I use as.numeric() to turn it into a numeric value

```
mode.i <- as.numeric(mode.i)
mode.i
```

```
## [1] 27
```

I can now make a histogram with the mode shown. abline() with v = mode.i puts a line at the mode.

```
hist(random.scores.191$score.i,
     breaks = bins,
     xlab = "scores")
abline(v = mode.i, col = 2, lwd = 4, lty = 2)
```


Histogram of random.scores.191$score.i
Frequency
scores