INTRODUCTION

Data analysts often use the shape of a histogram to propose the type of process that generates a set of random values. This example, based on a classroom exercise, shows four of the most common processes and shapes.

THE UNIFORM PROCESS

Students were asked to roll a single die and report the spot.

```r
u <- c( rep(1, 10),
       rep(2, 11),
       rep(3, 8),
       rep(4, 12),
       rep(5, 11),
       rep(6, 9))

t <- as.data.frame(table(u))
barplot(t$Freq, names.arg=t$u)
```

Each of the 6 values are equally likely, so a random sample gives similar frequencies for each value.
THE ADDITION PROCESS

Students were asked to roll a die 4 times, and report the sum of the rolls. Here, instead of using student data, the die rolls are simulated using a random number generator.

\[
\begin{align*}
\text{NSTUDENTS} &\leftarrow 100 \\
r1 &\leftarrow \text{sample(seq}(1,6), \text{NSTUDENTS, replace=TRUE}) \\
r2 &\leftarrow \text{sample(seq}(1,6), \text{NSTUDENTS, replace=TRUE}) \\
r3 &\leftarrow \text{sample(seq}(1,6), \text{NSTUDENTS, replace=TRUE}) \\
r4 &\leftarrow \text{sample(seq}(1,6), \text{NSTUDENTS, replace=TRUE}) \\
s &\leftarrow r1+r2+r3+r4 \\
t &\leftarrow \text{as.data.frame(table(s))} \\
\text{barplot}(t$Freq, names.arg=t$s)
\end{align*}
\]

With 4 values, the high and low values tend to “balance out,” so most of the sums are in the middle of the range of possible outcomes, giving a mound-shaped distribution.
A WAITING PROCESS

Students rolled a die until a 6 came up, and reported the number of rolls, required. Again, this process is simulated here, using a different random number generator.

```r
w <- 1+rgeom(NSTUDENTS, 1/6) # try rgeom(10, 1/6) to see why "1+" is used
t <- as.data.frame(table(w))
barplot(t$Freq, names.arg=t$w)
```

A short wait is more likely than a long wait, so smaller values are more frequent than larger ones, giving a right-skewed distribution.
**A RECORD PROCESS**

Students were asked to roll a die 4 times, and report the largest of the rolls. Rolls are simulated as in the Addition Process.

```r
r1 <- sample(seq(1,6), NSTUDENTS, replace=TRUE)
r2 <- sample(seq(1,6), NSTUDENTS, replace=TRUE)
r3 <- sample(seq(1,6), NSTUDENTS, replace=TRUE)
r4 <- sample(seq(1,6), NSTUDENTS, replace=TRUE)
r <- pmax(r1, r2, r3, r4)

t <- as.data.frame(table(r))
barplot(t$Freq, names.arg=t$r)
```

Getting all small values is less likely than getting at least one large value, so large outcomes are more frequent than small ones, giving a left-skewed distribution. What if this process were changed to report the smallest roll, instead of the largest one?