Part II : Descriptive Statistics
Histograms

In these notes we will introduce the histogram, to describe the shape of a distribution.

Histogram

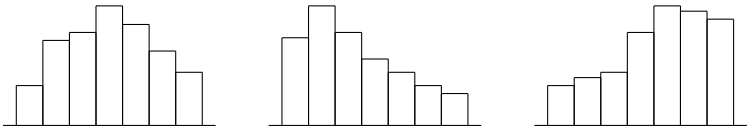
A histogram is a visual tool that we use to describe the shape of distribution of a random variable. We can construct a histogram of either the **frequency** or **density** using statistical computer software, (such as the `R` programming language).

Construction of the histogram :

1. Divide the horizontal axis into sub-intervals (preferably of equal length). Each sub-interval is a subset of the values that the random variable can take. (Usually, it is a good idea to divide the axis into between 5 and 20 such bins. Alternatively, using $\sqrt{n}$ bins also tends to work well, where n is the sample size).
2. The vertical axis will be the density or frequency (that is, the frequency divided by the length of the interval).
3. For each bin, draw a rectangle with height equal to the value of the density or frequency at that interval.
4. In practice, one would usually use computer software to construct histograms.

If the sample size n is not too small, a histogram allows us to describe the shape of the distribution of the random variable. For example, we can use a histogram to see whether the probability density function is symmetric.

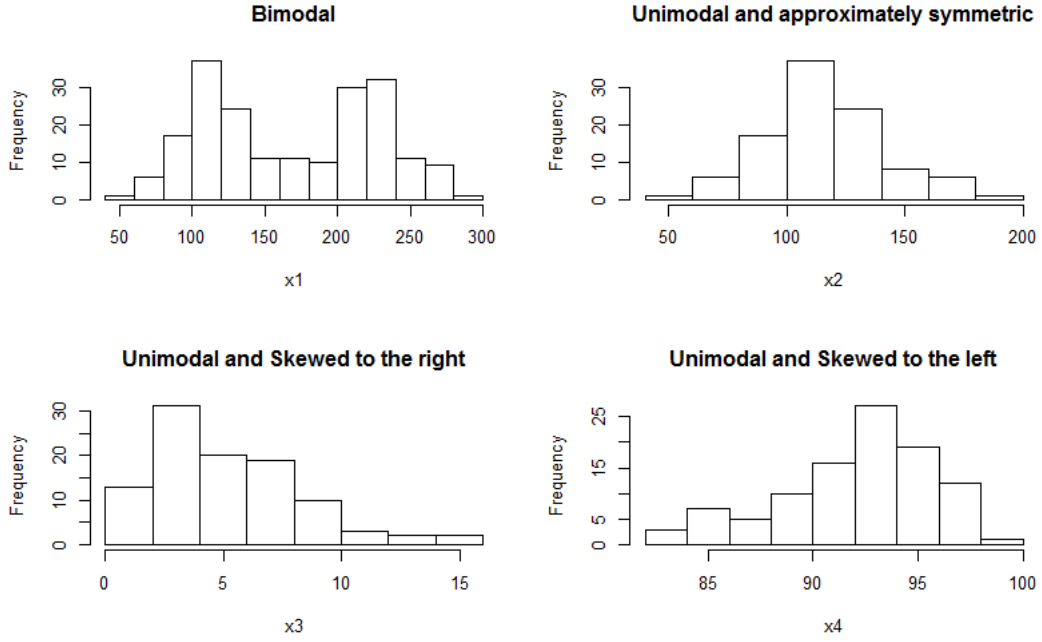
Here are some examples of histograms. The one on the left is (approximately) symmetric. The other two are not symmetric : the second histogram has a positive skewness (since the values that are far from the mean skew to the right) while the third histogram has what is called a negative skewness (since the values that are far from the mean skew to the left).



approximately symmetric | positive skewness | negative skewness

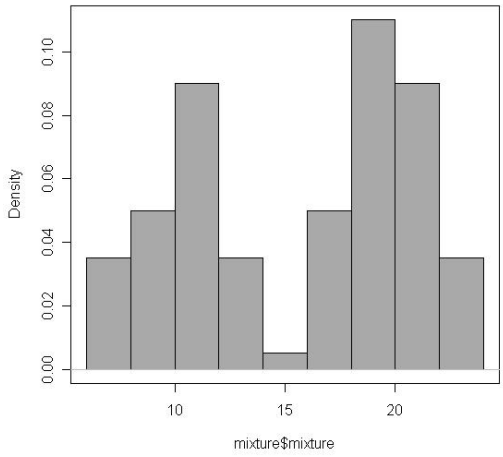
We also may want to determine whether the distribution is *unimodal* (i.e., it has one "bump") or *bimodal* (i.e., it has two "bumps"). If the distribution is bimodal this can sometimes indicate that the samples come from a heterogeneous population (i.e., they come from two different groups).

Here are examples of a random variable with distribution that are (from left to right) : bimodal, unimodal and approximately symmetric, a unimodal with positive skewnesss, or unimodal with negative skewness.

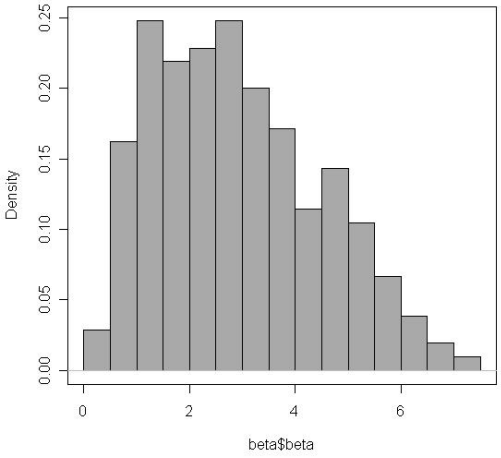


Example 40 : Consider the following three histograms. Describe the shape of the distribution of each of the random samples. Is it unimodal or bimodal ? If it is unimodal, does it have positive skewness, negative skewness or is it approximately symmetric ?

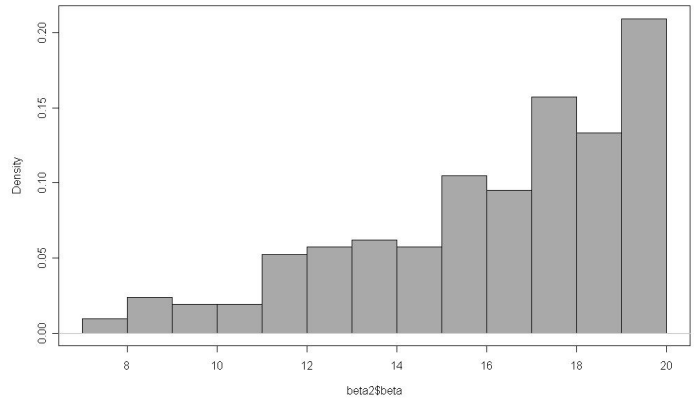
(a)



(b)



(c)



solution : (a) is bimodal, (b) is unimodal with positive skewness, and (c) is unimodal with negative skewness.