Python code

```
import pandas as pd

## Histogram of Jammael Charles' rushing attempts in 2010
charles = pd.read_csv('Dataset_2_2.csv')

## Creating a new figure
plt.figure(figsize = (10, 8))

plt.hist(charles, histtype = 'bar', bins = 10)
plt.xlabel('Yards')
plt.ylabel('Frequency')
plt.grid()
plt.show()
```

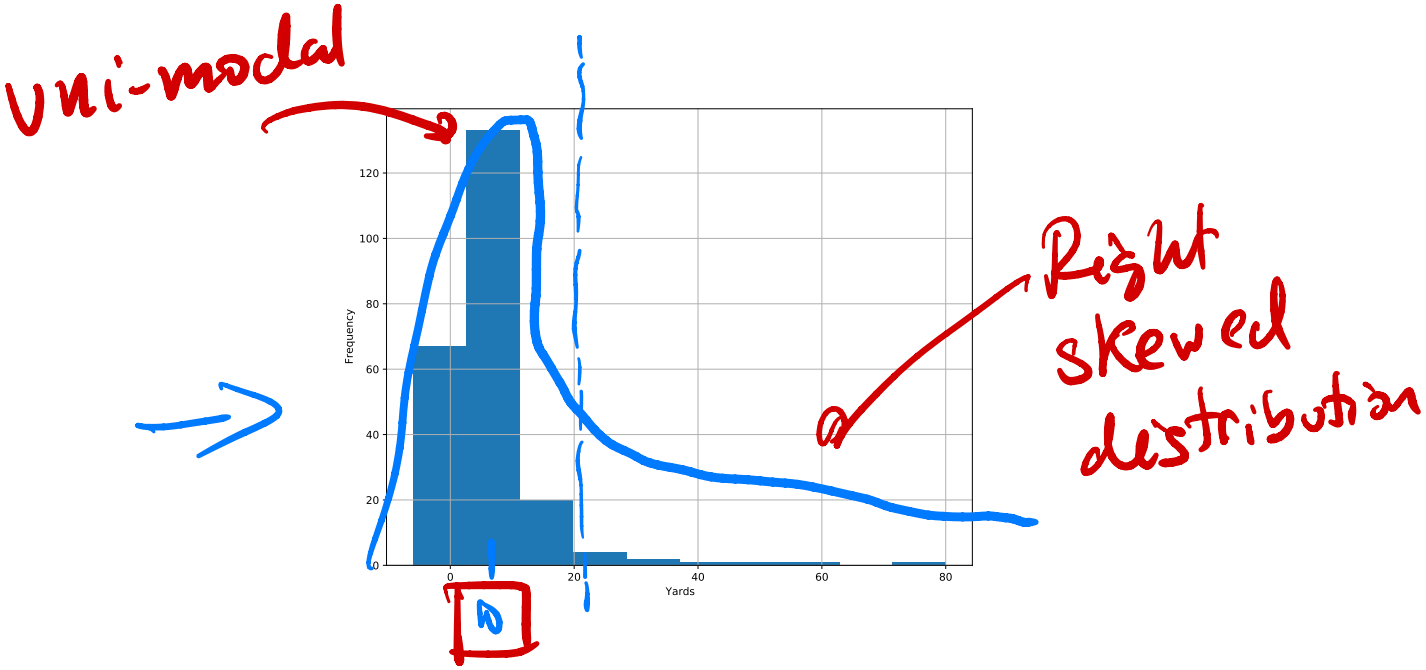

Figure 2.2: Charles' 2010 rushing yards by attempts

In Figure 2.2, we see that the histogram peaks around 5 or 6 yards; however, there are several values much greater than 5 yards. Such a distribution is said to be *right-skewed*, and this characteristic can be important in analyzing the data. Another important property of the normal distribution is that it has only one peak or *mode*; such a distribution is said to be *unimodal*. The histogram shown in Figure 2.3 is *bimodal*. A bimodal distribution has two peaks, separated by a valley. Note that the two peaks do not have to be the same height. More generally, a distribution might have several modes.

The data that was used to generate Figure 2.3 is data on shooting percentages of NBA players for the 2010-2011 season. Only "qualifying players," those with at least 300 field goals, are included. The following box shows the R code that was used to generate Figure 2.3.

R code

```
## Histogram of shooting percentage
shooting = read.csv(file = 'Dataset_2_3.csv')

hist(shooting$SPCT, col = 'gray', xlab = 'Shooting Percentage', ylab = 'Frequency')
box()
```

The following box shows the Python code that was used to generate Figure 2.3.

Python code

```
import pandas as pd

## Histogram of shooting percentage
shooting = pd.read_csv('Dataset_2_3.csv')

## Creating a new figure
plt.figure(figsize = (10, 8))

plt.hist(shooting['SPCT'], histtype = 'bar', bins = 12)
plt.xlabel('Shooting Percentage')
plt.ylabel('Frequency')
plt.grid()
plt.show()
```

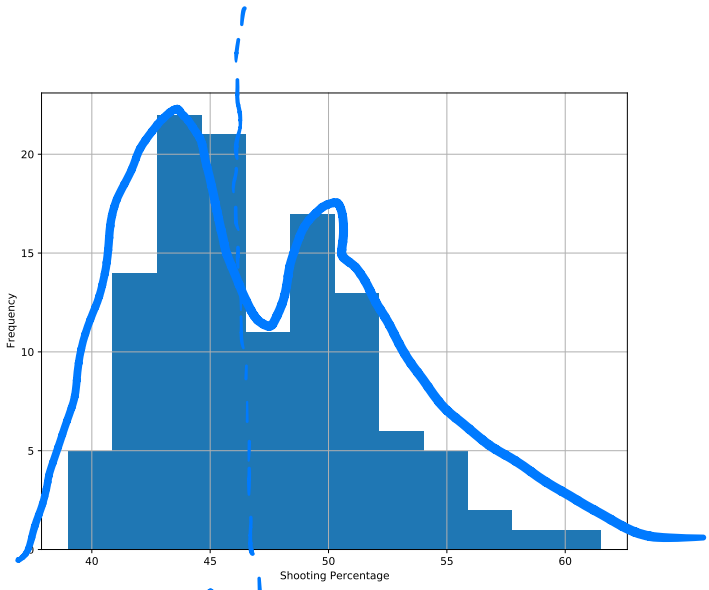

Figure 2.3: Shooting percentage

In Figure 2.3, we see the histogram has one peak around 44% and a second, smaller, peak around 50%. A bimodal distribution often occurs when two subgroups, with different distributions for the variable under consideration, are combined. In the NBA shooting percentage, we might expect different distributions of shooting percentage for guards and for forwards. Figure 2.4 shows the distributions of shooting percentage of guards and forwards. The following box shows the R code that was used to generate Figure 2.4.

R code

```
## Histogram of shooting percentage
shooting = read.csv(file = 'Dataset_2_3.csv')

## Histogram of guards
hist(shooting$SPCT[shooting$Pos == 'G'], col = 'gray',
     xlab = 'Shooting Percentage', ylab = 'Frequency')
box()

## Histogram of forwards
hist(shooting$SPCT[shooting$Pos %in% c('SF', 'PF')], col = 'gray',
     xlab = 'Shooting Percentage', ylab = 'Frequency')
box()
```

The following box shows the Python code that was used to generate Figure 2.4.

Python code

```
import pandas as pd

## Histogram of shooting percentage
shooting = pd.read_csv('Dataset_2_3.csv')

## Creating a new figure
plt.figure(figsize = (10, 8))

plt.hist(shooting['SPCT'][shooting['Pos'] == 'G'], histtype = 'bar', bins = 8)
plt.title('Guards')
plt.xlabel('Shooting Percentage')
plt.ylabel('Frequency')
plt.grid()

plt.hist(shooting['SPCT'][np.isin(shooting['Pos'], ['PF', 'SF'])],
         histtype = 'bar', bins = 10)
plt.title('Forwards')
plt.xlabel('Shooting Percentage')
plt.ylabel('Frequency')
plt.grid()
```

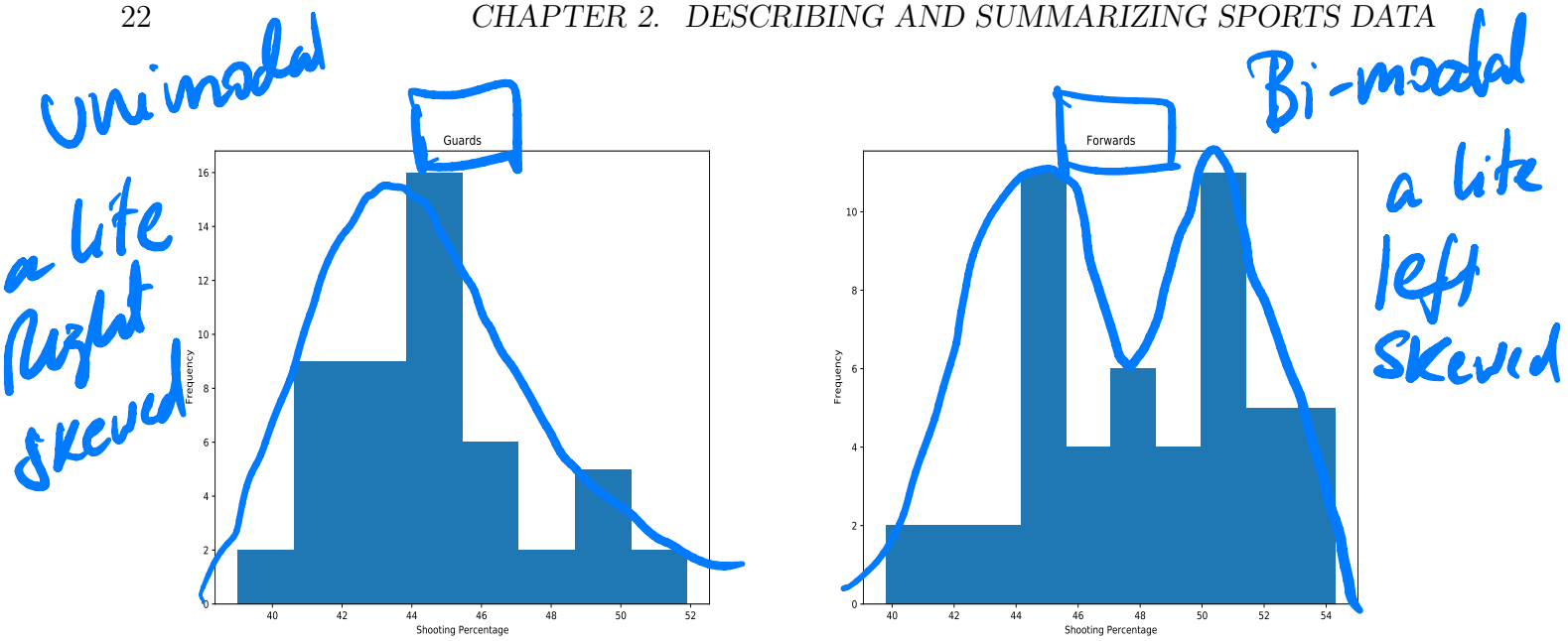

Figure 2.4: Left panel: shooting percentage of guards. Right panel: shooting percentage of forwards.

2.5 Mean and Median

Although a frequency distribution provides some summarization of a dataset, in some cases a single number summary is useful. For quantitative data, the *mean* and the *median* are the most commonly used summaries of this type. The mean of a numeric dataset is simply the average, that is, the total sum divide by the number of observations. For example, consider the data on Tom Brady's passing yards in his 159 starts (2001-2011), the mean is 251.1 yards per game. Averages are commonly used in traditional sports statistics. The following box shows the R code that was used compute the average yards per game of Tom Brady's dataset.

R code

```
## Reading Tom Brady's passing yards data (2001-2011)
brady = read.csv(file = 'Dataset_2_1.csv')

## Computing the mean of passing yards
mean(brady$PY)
```

The following box shows the Python code that was used compute the average yards per game of Tom Brady's dataset.

Python code

```
import pandas as pd
```