and a triple.

Quantitative variables are often classified as either *discrete* or *continuous*. Discrete variables are those whose set of possible values can be written as a list. For instance, suppose our subjects are NFL quarterbacks and we are analyzing performances in a given game. The variable "*number of interceptions*" takes values $0, 1, 2, \ldots$ and, hence, would be considered as discrete.

A continuous variable is one that takes any value in a range. For instance, consider the quarterback example and the variable "*completion percentage*" measured for each game. Completion percentage takes any value between 0 and 100 and hence is a continuous variable.

Notice that all categorical variables are discrete. Quantitative variables might be either discrete or continuous. In some cases, a variable might be modeled as either discrete or continuous. Consider the quarterback example and the variable completion percentage. A completion percentage of exactly 63.1% is technically impossible because it would require at least 1000 passing attempts; note that the fraction 631/1000 can't be simplified. However, the concept of completion percentage is clearly a continuous one, and it would be treated as a continuous variable. That is, a completion percentage in a game of 63.1% makes sense, even if it is technically impossible, while 2.4 interceptions in a game does not.

2.3 Frequency Distributions

The first step in analyzing a set of data is often some type of summarization. Consider the New York Yankees' 2011 season. If we are interested in the overall performance in the regular season, we could look at their game results, as presented in Table 2.1.

categorical variable

Table 2.1: Win and losses for the 2011 Yankees

W	W	L	W	L	W	L	W	L	W	W	L	W	W	L	W	W	W
L	L	W	W	L	W	W	W	L	L	L	W	L	W	W	L	L	L
L	L	L	W	W	W	L	W	W	L	W	W	L	L	W	W	W	W
L	W	W	L	L	L	W	W	W	L	W	W	W	L	W	W	W	W
L	L	W	W	W	W	W	W	W	L	L	W	L	L	W	W	L	L
W	W	W	L	W	L	W	L	W	W	W	L	L	W	W	W	W	W
W	W	W	L	L	L	W	W	L	W	W	W	L	W	W	L	W	L
L	W	L	L	W	W	W	L	W	W	W	W	W	W	L	L	L	L
W	W	W	L	L	W	L	W	W	W	W	L	W	W	L	L	L	L

Notice that W denotes a win and L denotes a loss. The results are in order that the games were played, across the rows, so that the first two games were wins, followed by a loss, for example. However, for many purposes this list contains too much detail; thus, we might summarize it by noting that in 2011 the Yankees won 97 games and lost 65 as shown in Table 2.2.

Frequency table

Table 2.2: Yankees Win-Loss Record in 2011

Outcome	Number	Percentage
Win	97	59.9%
Loss	65	40.1%
Total	162	

97/162

65/162

This is a simple example of *frequency table*, and the information contained in the table is a *frequency distribution*. In statistical terminology, the number of wins is called the *frequency* of wins, and the percentage of wins is its *relatively frequency*.

The frequency distribution of a categorical variable is defined in a similar manner; simply count the number of occurrences of each possible value of the variable to obtain the frequencies. The relative frequencies are found by dividing these counts by the total number of observations; relative frequencies can be reported as either proportions or percentages. If the variable is an ordinal one, the categories are generally placed in increasing order. A simple example is given in Table 2.3, which contains the results of Ryan Braun's 187 hits in 2011.

Table 2.3: Brian's hits in 2011

Result	Count	Percentage
S	110	58.8%
D	38	20.3%
T	6	3.2%
HR	33	17.6%
Total	187	

relative frequencies

frequencies

The frequency table for a discrete quantitative variable is handled is the same manner. For a continuous quantitative variable, the construction of a frequency table is a little more complicated because we can't simply list the possible values of the variables. In this cases, we divide the range of variables into non-overlapping classes so that each observation falls into exactly one class. Then, the frequency and relative frequency of each class are determined as for a categorical variable. Table 2.4 contains a frequency table of Tom Brady's passing yards per game in games started from 2001 through the 2011 season.

Therefore, there is some subjectivity in determining the frequency distribution of a continuous variable because of the subjectivity in how the classes are choses. For example, in the Tom Brady's yards example, Table 2.5 gives another valid frequency table. The goal in choosing classes is to make the table useful for the analysis in mind.

From the statistics standpoint, the classes have to be of the same length; however, in some cases it might be preferable to make some classes longer than others. One situation in which this occurs is when there are particular reference values of interest for the variable. For instance, in the Tom Brady's passing yards example, we might be interested in the following classes: less than

200, 200-299, 300-399, and at least 400 because these are values that are commonly used when discussing passing performance as shown in Table 2.6.

Table 2.4: Tom Brady's passing yards in games stated 2001 to 2011

classes

Class	Count	Percentage
0-100	5	3.1%
101-200	37	23.3%
201-300	72	45.3%
301-400	42	26.4%
401-500	2	1.3%
501-600	1	0.6%
Total	159	

Table 2.5: Tom Brady's passing yards in games stated 2001 to 2011 (second version)

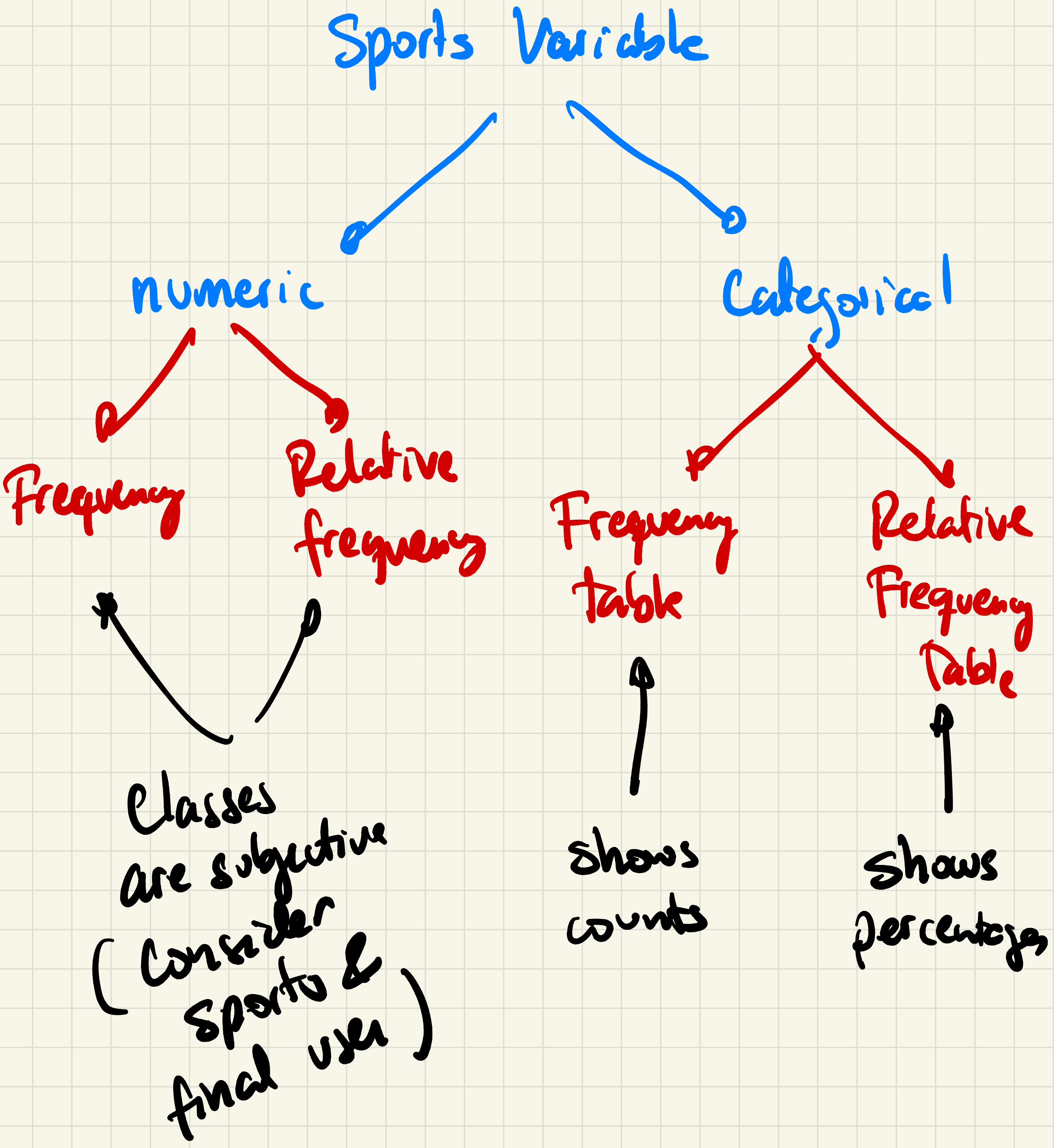
classes

Class	Count	Percentage
0-50	1	0.6%
51-100	4	2.5%
101-150	13	8.2%
151-200	24	15.1%
201-250	39	24.5%
251-300	33	20.8%
301-350	23	14.5%
351-400	19	11.9%
401-450	2	1.3%
451-500	0	0.0%
501-550	1	0.6%
Total	159	

Table 2.6: Tom Brady's passing yards in games stated 2001 to 2011 (third version)

Class	Count	Percentage
Less than 200	42	26.4%
200-299	72	45.3%
300-399	42	26.4%
400 or more	3	1.9%
Total	159	

If we are interested in the values of frequencies or relative frequencies, then the frequency table is more useful than a histogram because these values are clearly presented. However, the


Sports Variable
numeric
Categorical
Frequency
Relative frequency
Frequency Table
Relative Frequency Table
Classes are subjective (Consider sports & final user)
shows counts
shows percentages

histogram is useful for conveying the general pattern of frequencies, often referred to as the *shape* of the distribution.

2.4 Histogram

(numerical variables)

The shape can be thought of as the information contained in the histogram. In many respects, the *ideal* shape for a distribution is the familiar bell-shaped curve of the normal distribution; an example of such a histogram is given Figure 2.1.

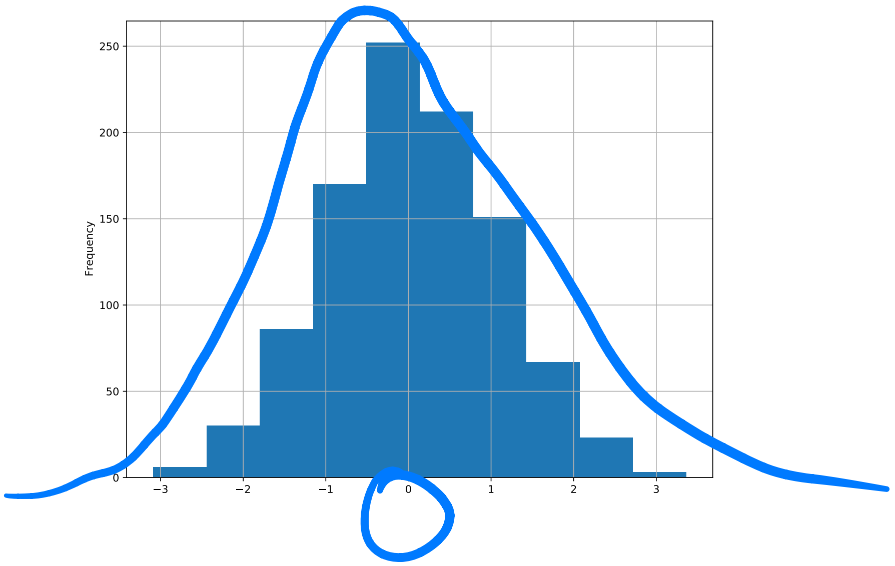


Figure 2.1: Shape of normal distribution

Of course, it is unrealistic to expect that any dataset will have a distribution that is exactly normal. However, the normal distribution can be used as a guide when assessing the distribution of a variable. An important aspect of shape is *symmetry*. Note that the normal distribution is exactly symmetric about its peak. However, not all distributions are symmetric. For instance, consider the Jammael Charles' rushing attempts in 2010 shown in Figure 2.2. Notice that the following box shows the R code that was used to generate Figure 2.2.

R code

```
## Histogram of Jammael Charles' rushing attempts in 2010
charles = read.csv(file = 'Dataset_2_2.csv')

hist(charles$Yards, col = 'gray', xlab = 'Yards', ylab = 'Frequency')
box()
```

The following box shows the Python code that was used to generate Figure 2.2.

Python code

```
import pandas as pd        import matplotlib.pyplot as plt

## 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()
```

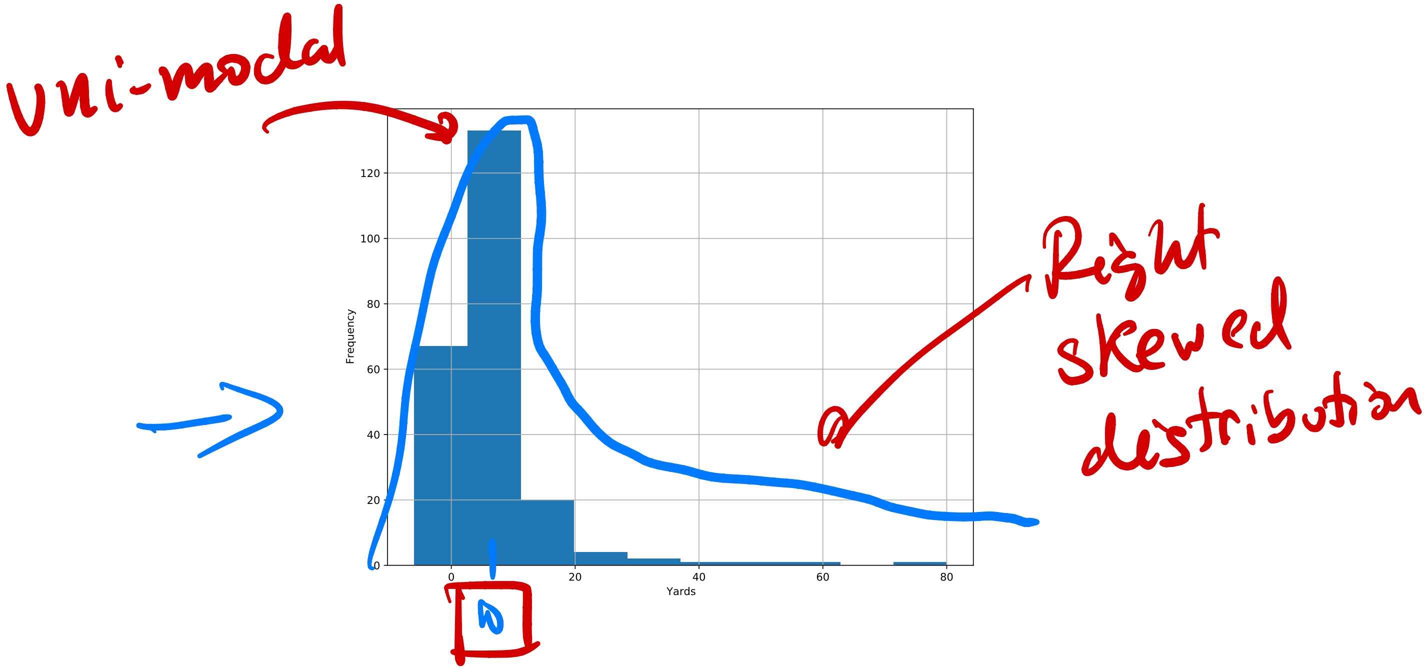


Figure 2.2: Charles' 2010 rushing yards by attempts

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.