Chapter 2

Describing and Summarizing Sports Data

2.1 Introduction

Analytical methods use data together with mathematical and statistical modeling to gain information and make decisions. A vast amount of data is collected about sports, and this data collection can only be expected to grow in the future.

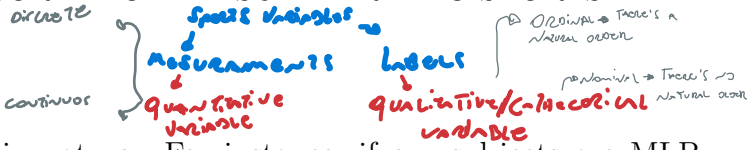
Although extracting information from data generally involves sophisticated statistical techniques, before describing such methods it is important to have a good understanding of the basic properties of data. In describing data, it is useful to think in terms of *subjects* and *variables*. A subject is an object on which data are collected; for sports data, the subjects are often players, but they might also be teams, games, seasons, or coaches. In some analyses, there is some subjectivity in how subjects are defined. For instance, suppose that we are studying the performance of NFL quarterbacks in the 2011 and 2021 seasons. We might treat each specific quarterback, such as Tom Brady, as a subject, with two sets of performance measures per quarterback, one for each year. On the other hand, we might treat the 2011 and 2021 versions of Brady as two different subjects. The choice depends on the goal of the analysis.

A variable is a characteristic of a subject that can be measured. For instance, if the subjects are MLB players in the 2011 season, games played, at bats, hits, and assists are all examples of variables. The set of all variables for the subjects under consideration makes up the data to be analyzed. The appropriate type of analysis depends on the properties of the data.

2.2 Types of Data Encountered in Sports

An important property of a variable is the set of values that can take, called its *measurement scale*. A measurement scale is often a set of numbers. For instance, if four subjects are NFL running backs in the 2011 season and our variable is the number of carries in that season, then the measurement scale is the set of integers $0, 1, 2, \ldots$. Note that, although there is a practical limit to the number of carries that a running back might have in a season, it is not usually necessary to specify that maximum in advance. A variable with a measurement scale consisting of numbers

is said to be *quantitative*.



However, not all variables are numerical in nature. For instance, if our subjects are MLB players in the 2011 season and our variable is the way in which the player bats, then the set of all possible values is left (L), right (R), and switch hitter (S). Such a variable is said to be *qualitative* or *categorical*.

Although categorical variables are often important and useful, there are limitations to the type of statistical methods that can be used on such variables. For instance, in the batting example, described in the previous paragraph, it does not make sense to refer to an "*average batting style*"; however, it would make sense to say that right-handed batter are most common or that 13% of the players are switch hitters.

The measurement scale of a categorical variable might be ordered or unordered. For instance, in the batting example, there is no natural ordering of R, L, and S. Such a measurement scale is said to be *nominal*, and the variable is said to be *nominal variable*. In other cases, the possible values of categorical variable might have a natural ordering. For instance, suppose that the subject is a batter in baseball and the variable is the outcome of a bat, with possible values single (S), double (D), triple (T), home run (HR) and no hit (NH). Note that it is possible to divide NH into more specific outcomes (such as strikeout, fly out, etc.), resulting in a slightly different variable. The five possible values S, D, T, HR, and NH are ordered, in the sense that NH is the least preferable, followed by S, D, T, HR in order. Although this is a nominal variable, it has a little more structure than the variable measuring batting style. A nominal variable in which the possible values are ordered is said to be *ordinal*, and it is said to have a *ordinal scale*.

Notice that an ordinal variable can be converted to a quantitative variable by assigning numerical values to each possible value. For instance, in a total bases calculation, we have the following mapping:

$$\begin{aligned} \text{NH} &\to 0 \\ \text{S} &\to 1 \\ \text{D} &\to 2 \\ \text{T} &\to 3 \\ \text{HR} &\to 4 \end{aligned}$$

Although there is some obvious logic to these choices, it is important to keep in mind that they are, to a large extent, arbitrary. Once such a numerical assignment is made, the resulting quantitative variable can be analyzed using the same methods used for other quantitative variables.

One way to think about appropriateness of such numerical assignment is to consider distances between levels corresponding to the assignment. For instance, using the total bases variable, the difference between a single and a double is the same as the difference between a triple and a home run, and the difference between a single a home run is the same as the difference between an out

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.

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.

Table 2.2: Yankees Win-Loss Record in 2011

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

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

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)

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

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

2.4 Histogram

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.

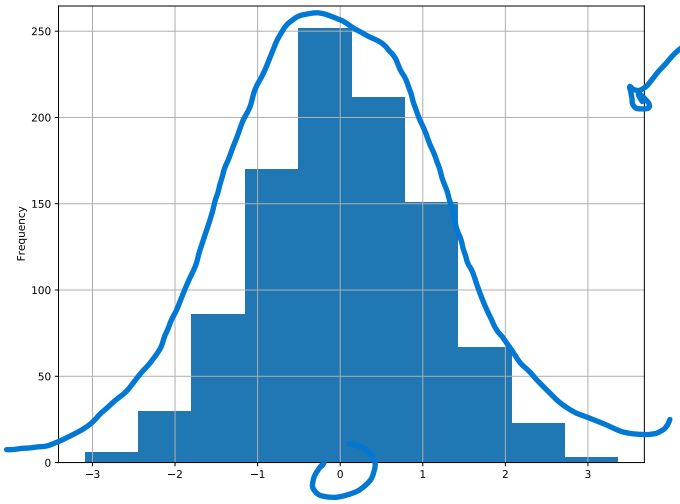


Figure 2.1: Shape of normal distribution

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

## 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()
```

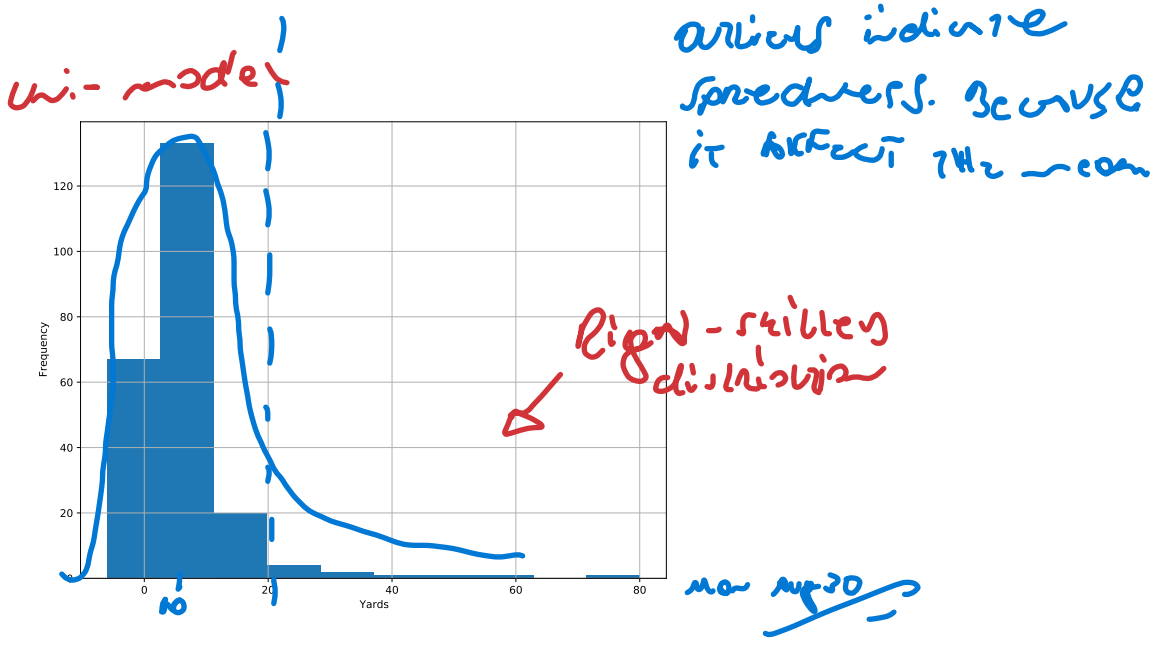


Figure 2.2: Charles' 2010 rushing yards by attempts

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()
```

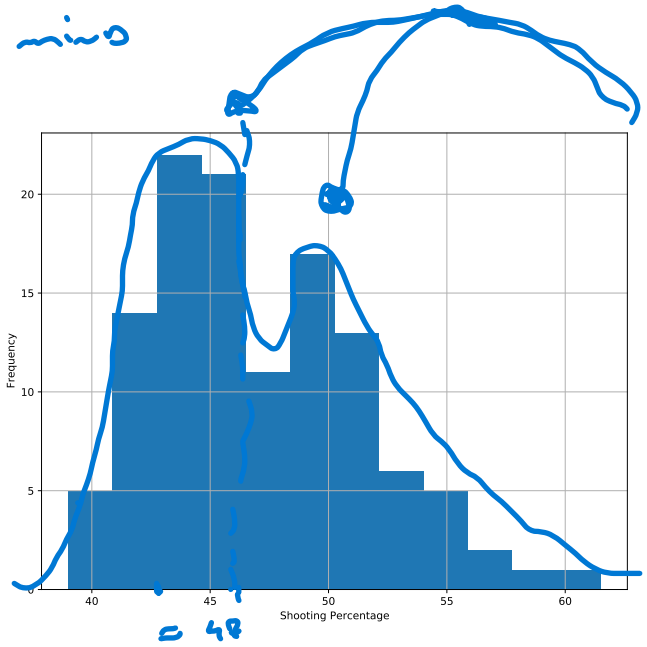


Figure 2.3: Shooting percentage

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()
```

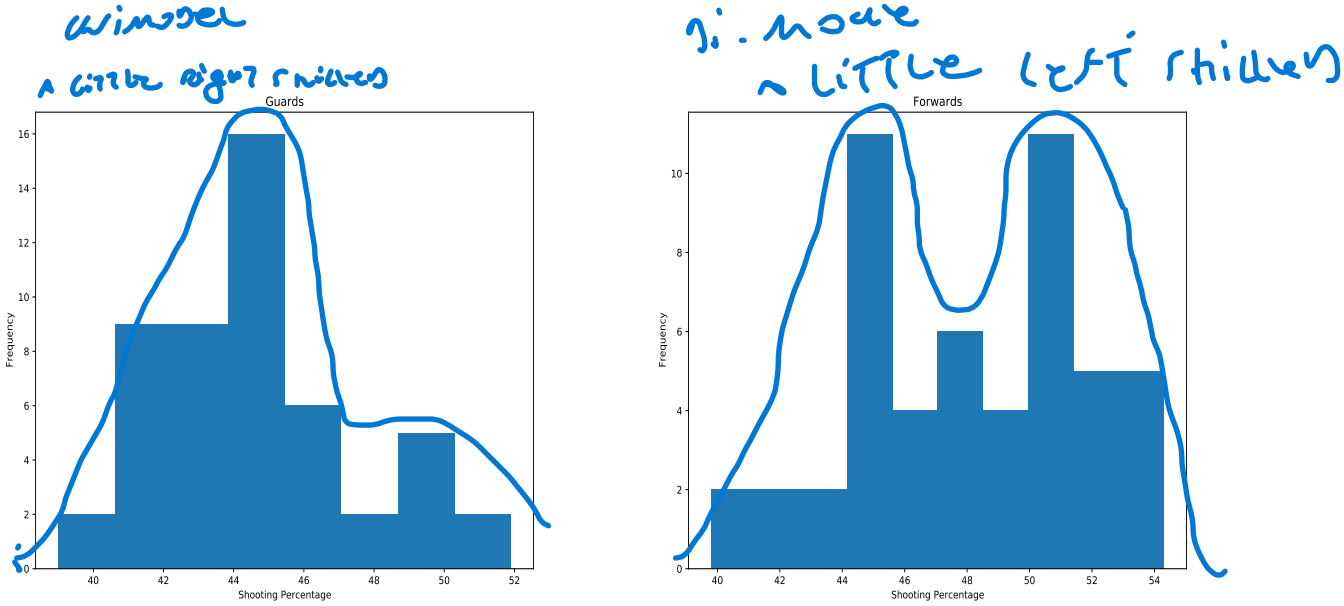


Figure 2.4: Left panel: shooting percentage of guards. Right panel: shooting percentage of forwards.

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
```

Python code

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

## Computing the mean of passing yards
brady['PY'].mean()
```

The median is often a useful alternative to the mean. The median is found by putting the data values in order and then finding the middle value. Thus, the median has the interpretation that half of the data values are above the median and half of the data values are less than the median. For the Brady's passing yard dataset, the median value is 249. The following box shows the R code that was used to compute the median of the Tom Brady's passing yard 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
median(brady$PY)
```

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

Python code

```
import pandas as pd

## Reading Tom Brady's passing yards data (2001-2011)
brady = pd.read_csv('Dataset_2_1.csv')

## Computing the mean of passing yards
brady['PY'].median()
```

The mean and median are two different ways to summarize a dataset, and which is better in a given situation depends on the goals of the analysis. If the distribution is approximately symmetric, then the mean and the median are approximately equal. For skewed distribution, however, the mean and the median are different because very large (or very small) observations have a greater effect on the mean than on the median.

2.6 Measuring Variation

Variation is an important part of all sports. The players in a league have varying skill sets, and their game performance vary. If two teams play each other several times, the outcomes will vary.

Understanding, and accounting for, variation is a central goal of analytical methods. In this section, we will discuss standard measures that are commonly used to quantify variation. In later chapters, we will discuss other measures that are used to quantify variation in sports.

The variation of numerical variables refers to how the observed values of the variable differ from one another. One approach to measuring variation is to choose a reference value for the data and consider how the values differ from the the reference value. A commonly used choice for the reference value is the mean of the data.

The *standard deviation* of a dataset is, roughly speaking, the average distance from an observation to the mean of the data values. To compute the standard deviation, we first average the squared distances of the data values from the mean value; this is called the *variance* of the dataset. Although variances are sometimes used directly, they have the drawback that the units are the square of the units the observations themselves. For instance, if the measurements is in yards, the units of the variance will be yards squared. Therefore, we typically use the square root of the variance, which is called the standard deviation. The standard deviation may be viewed as the "typical" distance of a data value to the mean of the dataset. Note that the units of the standard deviation are the same as the units of the variable under consideration.

in sports, standard deviation is preferred over variance

For example, consider the shooting percentage dataset of 2010-2011 NBA by position. The standard deviation of forwards is 3.66 while the standard deviation of guards is 2.88. The following box shows the R code that was used to compute the variance and standard of the shooting percentage.

R code

```
## Reading csv file
shooting = read.csv(file = 'Dataset_2_3.csv')

## Computing the variance of shooting percentage of forwards
var(shooting$SPCT[shooting$Pos %in% c('SF', 'PF')])

## Computing the variance of shooting percentage of guards
var(shooting$SPCT[shooting$Pos == 'G'])

## Computing the standard deviation of shooting percentage of forwards
sd(shooting$SPCT[shooting$Pos %in% c('SF', 'PF')])

## Computing the standard deviation of shooting percentage of guards
sd(shooting$SPCT[shooting$Pos == 'G'])
```

The following box shows the Python code that was used to compute the variance and standard of the shooting percentage.

Python code

```
import pandas as pd
```

Python code

```
import numpy as np

## Reading the csv file
shooting = pd.read_csv('Dataset_2_3.csv')

## Computing the variance of shooting percentage of forwards
var_forward = np.var(shooting['SPCT'][np.isin(shooting['Pos'], ['PF', 'SF'])])

## Computing the variance of shooting percentage of guards
var_guard = np.var(shooting['SPCT'][shooting['Pos'] == 'G'])

## Computing the standard deviation of shooting percentage of forwards
sd_forward = np.std(shooting['SPCT'][np.isin(shooting['Pos'], ['PF', 'SF'])], ddof = 1)

## Computing the standard deviation of shooting percentage of guards
sd_guard = np.std(shooting['SPCT'][shooting['Pos'] == 'G'], ddof = 1)
```

One drawback of the standard deviation is that it is difficult to interpret directly. One reason for the difficulty in interpreting the standard deviation is that it is sensitive to the measurement units used; specially, if the data values are all multiplied by a constant k, then the standard deviation of the new dataset is k times the original standard deviation.

To eliminate the dependence on the units used, it is sometimes useful to express the standard deviation as a proportion, or percentage, of the average data values. The *coefficient of variation* (CV) is the standard deviation divided by the mean. Note that the coefficient of variation is a dimensionless quantity; therefore, it does not depend on the measurement scale. If all data values are multiplied by a constant k, both the standard deviation and the average of the new data values are multiplied by k and k cancels when taking the ratio. The coefficient of variation is generally only used for variables in which the values can't be negative or zero. Table 2.7 shows the the coefficients of variation for shooting percentage of guards and forwards in 2010-2011 NBA season.

Table 2.7: Coefficient of Variation for shooting percentage in 2010-2011 NBA season

Position	CV(%)
Guard	6.5%
Forward	7.6%

Based on these results, forwards have the most variability in shooting percentage. The following box shows the R code that was used to compute the coefficient of variation for guards and forwards in 2010-2011 NBA season.

R code

```
## Computing the CV of shooting percentages of forwards
mean_forward = mean(shooting$SPCT[shooting$Pos %in% c('SF', 'PF')])
sd_forward = sd(shooting$SPCT[shooting$Pos %in% c('SF', 'PF')])
CV_forward = sd_forward / mean_forward

## Computing the CV of shooting percentages of guards
mean_guard = mean(shooting$SPCT[shooting$Pos == 'G'])
sd_guard = sd(shooting$SPCT[shooting$Pos == 'G'])
CV_guard = sd_guard / mean_guard
```

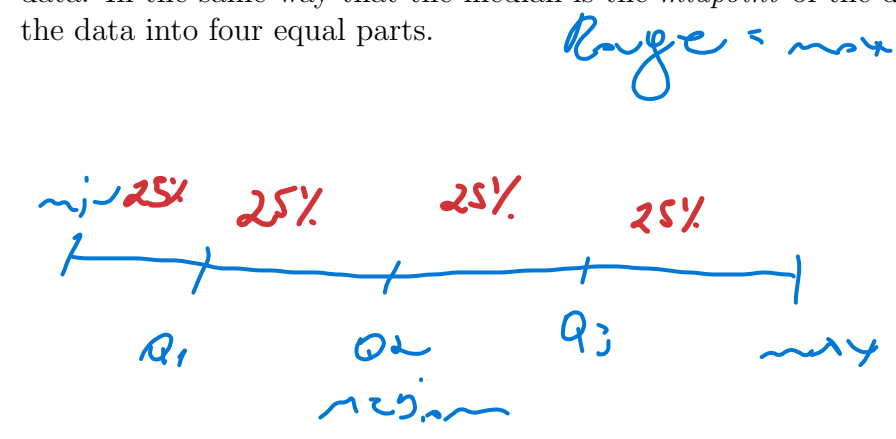
The following box shows the Python code that was used to compute the coefficient of variation for guards and forwards in 2010-2011 NBA season.

Python code

```
## Computing the CV of shooting percentages of forwards
mean_forward = np.mean(shooting['SPCT'][np.isin(shooting['Pos'], ['PF', 'SF'])])
sd_forward = np.std(shooting['SPCT'][np.isin(shooting['Pos'], ['PF', 'SF'])], ddof = 1)
CV_forward = sd_forward / mean_forward

## Computing the CV of shooting percentages of guards
mean_guard = np.mean(shooting['SPCT'][shooting['Pos'] == 'G'])
sd_guard = np.std(shooting['SPCT'][shooting['Pos'] == 'G'], ddof = 1)
CV_guard = sd_guard / mean_guard
```

Another approach to measuring the variation of a variable is to look at the spread of the values of the dataset. One such measure is the *range*, which is defined as the difference between the maximum value minus the minimum value in the dataset. However, the range is too sensitive to extreme values. An alternative approach is to base a measure of variation on the *quartiles* of the data. In the same way that the median is the *midpoint* of the dataset, the three quartiles divide the data into four equal parts.



Interquartile range : Q3 - Q1 (IQR)

The *inter-quartile range* (IQR) of the dataset is the upper quartile minus the lower quartile. Hence, the IQR give the length of the interval containing the middle half of the data. Note that the IQR offers at least two advantages over the standard deviation. One is that it has a more direct interpretation that is often useful in understanding the variability in a variable. Another is that it less sensitive to extreme values than is the standard deviation, in the same sense that the median is less sensitive to extreme values than is the mean. Note, however, that, like the standard deviation, the IQR is sensitive to the measurement scale used, and multiplying each data value by a constant k leads to the multiplication of the IQR by k. Table 2.8 shows the standard deviation and IQR of the shooting percentages of forwards and guards in 2010-2011 NBA season.

Table 2.8: IQR and standard deviations shooting percentage in 2010-2011 NBA season

Position	Q_1	Q_3	IQR	SD
Guard	42.65	46.15	3.5	2.88
Forward	45	50.85	5.85	3.66

The following box show the R code that was used to compute Q_1, Q_3, and IQR of the shooting percentages.

R code

```
## Reading csv file
shooting = read.csv(file = 'Dataset_2_3.csv')

## Computing the IQR of forwards shooting percentages
forward = shooting$SPCT[shooting$Pos %in% c('SF', 'PF')]
Q3_forward = quantile(forward, 0.75)
Q1_forward = quantile(forward, 0.25)
IQR_forward = as.numeric(Q3_forward) - as.numeric(Q1_forward)

## Computing the IQR of guards shooting percentages
guard = shooting$SPCT[shooting$Pos == 'G']
Q3_guard = quantile(guard, 0.75)
Q1_guard = quantile(guard, 0.25)
IQR_guard = as.numeric(Q3_guard) - as.numeric(Q1_guard)
```

The following box show the Python code that was used to compute Q_1, Q_3, and IQR of the shooting percentages.

Python code

```
import pandas as pd
import numpy as np
```

Python code

```
## Reading the csv file
shooting = pd.read_csv('Dataset_2_3.csv')

## Computing the IQR of forwards shooting percentages
forward = shooting['SPCT'][np.isin(shooting['Pos'], ['PF', 'SF'])]
Q3_forward = np.percentile(forward, 75)
Q1_forward = np.percentile(forward, 25)
IQR_forward = Q3_forward - Q1_forward

## Computing the IQR of guards shooting percentages
guard = shooting['SPCT'][shooting['Pos'] == 'G']
Q3_guard = np.percentile(guard, 75)
Q1_guard = np.percentile(guard, 25)
IQR_guard = Q3_guard - Q1_guard
```

Note that, in Table 2.8, the IQR is larger than the standard deviation (for forwards and guards). This is not surprising because the IQR and the standard deviation measure variation in different ways, and the relationship between IQR and standard deviation depends on the shape of the underlying distribution. For instance, for variables that follow a normal distribution, the standard deviation is approximately three-fourths of the IQR.

2.7 Sources of Variation

There can be many sources of the variation in a sequence of observations, and it is often interesting to consider the relative contributions of these different sources. This section considers this issue in the context of the variation in points (or runs) scored by teams over the course of a season.

Specifically, consider the variation in runs scored per game in the 2011 MLB season. Using data on all 162 games played for each of the 30 MLB teams, the standard deviation of runs scored is 3.01 runs. These data are taken the game logs on Baseball-Reference.com.

Note that this variation in runs scored is caused by two factors: the variation in the run-scoring ability of the various teams, called the *between-team* variation, and the variation in each team's game-to-game run scoring, called the *within-team* scoring. Note that these two factors reflect different aspects of variation: between-team variation represents the differences between the 30 MLB teams, and the within-team variation represents the fact that the runs scored by a specific team vary considerably throughout the season.

Both types of variation can be measured by standard deviations, calculated from the datasets under consideration. For instance, to measure between-team variation, we can use the standard deviation of the team-by-team average runs scored for the season; this value is 0.508.

To measure the within-team variation, we can calculate the standard deviation of runs scored for each of the 30 teams and average the results. The average standard deviation of the 30 teams

is 2.98 runs. Therefore, we have three measures of variation of runs scored in MLB games: the overall variation, the between-team variation, and the within-team variation. The overall variation is measures by S_0, the standard deviation of the set of all 4860 runs scored values in all games played in 2011 is $S_0 = 3.01$. The between-team variation is measured by S_B, the standard deviation of the average runs scored for the 30 MLB teams is $S_B = 0.508$. The within-team variation is measured by S_w, obtained by computing the standard deviation of the runs scored for each team and then averaging these 30 values is $S_w = 2.98$.

Not surprisingly, these three measures are related. Specifically,

$$S_0^2 \approx S_B^2 + S_w^2$$

This approximation is valid whenever both the number of teams under consideration and the number of observations for each team are relatively large, say greater than 10. Partition the overall variation in this way allows us to consider the proportion of variation in runs scored per game that is due to the fact that teams have different offensive capabilities. Specifically, since $S_0^2 = 9.06$ and $S_B^2 = 0.26$, and

$$\frac{S_B^2}{S_0^2} = \frac{0.26}{9.06} = 0.029$$

From the above result, we can conclude that approximately 3% of the variation in scoring in MLB games is caused by the variation between teams. The other 97% is because of the variation within each team, which can be attributed to a number of factors, including the differing abilities of opposing pitchers and the natural variation in run scoring in baseball.

2.8 Measuring Variation in Qualitative Variables

The concept of variability is most commonly applied to quantitative variables. However, in some cases, we are interested in measuring the variability in qualitative variables. For instance, consider the types of pitches thrown by Clayton Kershaw during the 2012 baseball season. Using the PITCHf/x data as reported on fangraphs.com, 62.5% of Kershaw's pitches were fastballs, 22.6% were sliders, 11.2% were curveballs, 3.4% were changeups, and 0.3% did not fall into one of the recorded categories and are labeled as "other." For comparison, consider the pitch distribution of Cole Hamels who threw 51.3% fastballs, 8.9% curveballs, 30.3% changeups, 9.1% cutters, and 0.4% other. We might be interested in determining which of these pitchers had more variability in his pitch selection, or more generally, we might be interested in how variability in pitch type is related to success on the mound.

Note that the variable analyzed here, "pitch type," is qualitative; therefore, measures of variability based on the standard deviation do not apply. Instead, we focus on how the variable values are distributed across the various categories. Consider the following categories for pitchers: fastball (FA), slider (SL), curveball (CU), changeup (CH), cutter (FC), sinker (SI) and other (O). For a given pitcher, let

$$p_{FA},\ p_{SL},\ p_{CU},\ p_{CH},\ p_{FC},\ p_{SI},\ p_O$$

denote the proportions of pitches in each of the categories, respectively. For example, for Kershaw,

$$p_{FA} = 0.625,\ p_{SL} = 0.226,\ p_{CU} = 0.112,\ p_{CH} = 0.034,\ p_O = 0.003$$

Note that p_{SI} and p_{FC} are 0. A measure of variability of pitch type is a function of these seven proportions. There are a few basic properties such a measure should satisfy: It should be non-negative and equal to 0 only if all the pitches are of one type, and it should take its maximum value if each pitch type is equally likely because that is the most variation we can have in the pitch distribution. One measure of variability satisfying these requirements is the *entropy*. The entropy is given by

$$-(p_{FA}\ln(p_{FA}) + p_{SL}\ln(p_{SL}) + p_{CU}\ln(p_{CU}) + p_{CH}\ln(p_{CH}) + p_{FC}\ln(p_{FC}) + p_{SI}\ln(p_{SI}) + p_O\ln(p_O))$$

where ln denotes the natural log function, and $0\ln(0)$ is interpreted as 0. The entropy can be interpreted as the "predictability" of the pitch type based on the observed proportions; it is used in many ares of sciences. Because $\ln(1) = 0$, if one proportion is 1 while the others are all 0, the entropy is 0; otherwise, it is positive. It may be shown that the maximum value of the entropy is achieved of all the proportions are equal: in the pitch example, this maximum value is $\ln(7)$. The standardized entropy can be calculated by dividing the entropy by this maximum value. The standardized entropy then lies in the interval 0 to 1. For Kershaw, the entropy of his pitch distribution is

$$-(0.625\ln(0.625) + 0.226\ln(0.226) + 0.112\ln(0.112) + 0.034\ln(0.034) + 0.003\ln(0.003)) = 1.0075$$

Because $\ln(7) = 1.9459$, the standardized entropy of Kershaw's pitch distribution is $\frac{1.0075}{1.9459} = 0.518$. For comparison, the standardized entropy of Hamel's pitch distribution is 0.596, which indicates that there is more variability in Hamel's pitch selection than there is in Kershaw's.