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.

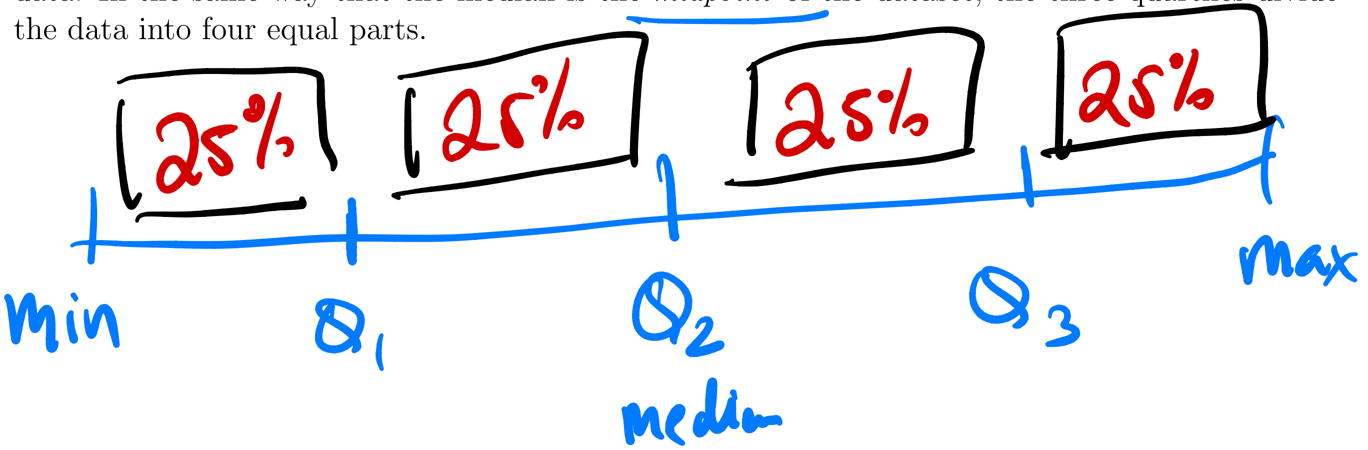
Python code

```
## 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
```

Range = max - min

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 (IQR) = $Q_3 - Q_1$

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)    75%
Q1_forward = quantile(forward, 0.25)    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)    75%
Q1_forward = np.percentile(forward, 25)    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

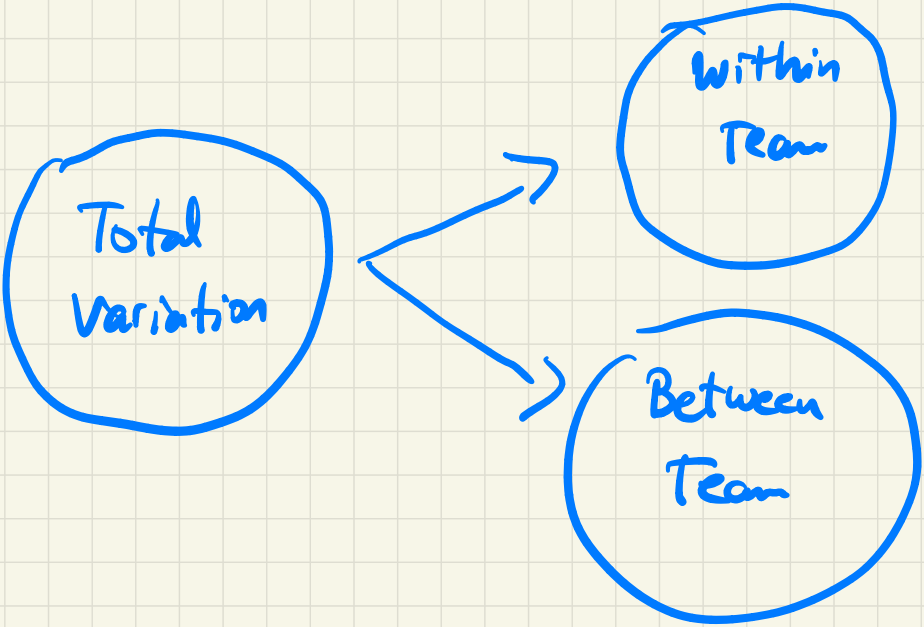
There can be many sources of the variation in a sequence of observations, and it is often interesting to consider the relative contributions of these different sources. This section considers this issue in the context of the variation in points (or runs) scored by teams over the course of a season.

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


Total Variation
Within Team
Between Team

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,

approx

$$\boxed{S_0^2 \approx S_B^2 + S_w^2}$$

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

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

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

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

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

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

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

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

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.

$$P_{FA} = P_{SL} = P_{CU} = P_{CH} = P_{FC} = P_{SI} = P_{O} = \frac{1}{7}$$

entropy

$$\hookrightarrow \frac{1}{7}\ln\left(\frac{1}{7}\right) = \frac{1}{7}\ln(7^{-1}) = -\frac{1}{7}\ln(7)$$

$$\text{entropy} = -\left(\underbrace{\frac{1}{7}\ln\left(\frac{1}{7}\right)} + \cdots + \underbrace{\frac{1}{7}\ln\left(\frac{1}{7}\right)}\right)$$

$$= -\cancel{7}\left(\frac{1}{\cancel{7}}\right)\ln\left(\frac{1}{7}\right)$$

$$= -\ln\left(\frac{1}{7}\right)$$

$$= -\ln(7)^{\boxed{-1}}$$

$$= -(-1)\ln(7)$$

$$= \ln(7)$$