Lecture 2, Module 3: Baseball Batting Statistics

CSCI S-5a: Programming in R

Let's clear the environment:

```r
rm( list = ls() )
```

Module Overview and Learning Objectives

In this module, we'll learn how to use R to calculate baseball batting statistics.

- In Section 1, we'll discuss why we're spending time studying this topic.
- In Section 2, we'll learn about the game of baseball.
- In Section 3, we'll see a variety of statistics that are used to quantify batting performance.
- In Section 4, we'll see how to format baseball performance statistics properly.
- In Section 5, we'll calculate the performance statistics for the great Babe Ruth.

When you've finished this module, you should be able to:

- Explain why we're spending so much time on baseball statistics.
- Explain the basic facts about the game of baseball.
- Calculate the batting average, on-base percentage, and slugging average for a player.
- Correctly format baseball performance statistics.

We won't meet any new built-in R functions in this module.

Section 1: Why are we doing this?

> **Main Idea:** *Let's understand why we are spending time on this topic.*

In this section, we'll discuss why we're spending time studying this topic.

Why are we wasting our time on baseball – isn't this supposed to be a course on R?

One problem that I have in designing this course is that I want to show you how to use R to perform complex multi-step calculations.

This is an important skill to learn, because everybody has to perform some sort of calculation at some point.

However, everybody is working in a different domain, so (with one exception) it's not really possible for me to find a problem that's applicable to everyone.

I like to use baseball statistics because it's a little bit of a challenge to work your way through all the definitions, but it's not terribly laborious.

The idea here is that you will observe the techniques that we use for baseball statistics, and then apply them to your own field of work.

This is the real appeal of studying baseball statistics – it provides a fantastic laboratory for learning many different R techniques.

In the end, you don't have to care about baseball, just as long as you use your imagination and creativity to take these methods and transfer them to your own projects.

As I've taught this course over the years, I've found that there are many techniques that can be illustrated by working with baseball statistics, and we'll return to this topic throughout the course.

In this lecture, baseball statistics will provide us with a case study for how to use variables in an extended computation.

One other very attractive feature of working with baseball statistics is that there is real-world data that is easily available.

So, you can obtain a player's basic statistics, calculate performance metrics, and then check to see that you get the correct result.

Finally, many students are interested in sports analytics, so hopefully for them this is a simple but real-world application.

Please note that we are **not** studying baseball statistics because I personally have any enthusiasm for the game.

In fact, I don't care about the game at all, but I get very excited about learning R using baseball statistics.

So that's why we are studying the topic of baseball batting statistics.

Now let's learn about the game of baseball!

Section 2: The Game of Baseball

> **Main Idea:** *We can learn some basic facts about baseball.*

In this section, we'll learn about the game of baseball.

Here are the essential facts about baseball that you'll need for CSCI S-5a.

In baseball, one player has a large stick.

The stick is called a "bat", and the player with the bat is called the "batter".

There's another player who throws a ball towards the batter.

The player who throws the ball is called a "pitcher", and each time he or she throws the ball it's called a "pitch".

In addition to the pitcher and the batter, there are a number of other players on the field with big funny-looking gloves, who are called "fielders".

There is a large five-sided object embedded in the ground called the "home plate", and there are also three square objects called "bases".

The batters on one team takes turns coming up to the home plate to try to hit the ball with the bat.

Each time a batter comes to home plate, it's called a "plate appearance".

There is a complex ritual in which the pitcher tries to throw the ball past the batter, and the batter tries to hit the ball with the bat.

We don't need to understand this process in CSCI S-5a; instead, all you need to know is that there are a number of different outcomes for a plate appearance.

The most common result of a plate appearance is that a batter is declared "out".

For instance, if a batter hits the ball and a fielder catches it in a big funny glove before it touches the ground, the batter is out.

There are many other ways for a batter to get out; all that we care about is that it is one possible outcome for a plate appearance.

Another possible outcome is that the batter hits the ball and the fielders don't catch it.

In this case he or she begins to run around touching the bases, and if possible runs back to home plate.

The player has to stop on one of the bases.

- If the batter ends up stopping on the first base, this is called a "single".
- If the batter ends up stopping on second base, this is called a "double".
- If the batter ends up stopping on third base, this is called a "triple".
- If a player hits the ball, touches all three bases, and finally touches home plate, this is called a "home run".

If the batter gets either a single, a double, a triple, or a home run, then that's called a "hit".

Note that a "hit" in baseball doesn't just mean that the batter was able to strike the ball with the bat; it means that the player gets a single, a double, a triple, or a home run, and doesn't get out.

In general, the batter is trying to get a hit, and wants to avoid getting out.

Another possible outcome for a plate appearance occurs when a pitcher throws too many bad pitches to the batter, in which case the batter automatically gets to go stand on first base, and this is called a "base on balls" or a "walk".

Also, if the pitcher hits the batter with the ball, the batter automatically gets to go stand on first base.

Finally, there is a thing called a "sacrifice fly"; I'm not going to get into the details, because this is a very rare event, but instead just want to note that this is something that we need to keep track of.

There are other possible outcomes for a plate appearance, but they are very exotic and I'm just going to call them "other".

Thus, for CSCI S-5a, there are nine possible outcomes for a plate appearance:

- An out
- A single
- A double
- A triple
- A home run
- A base on balls

- Hit by a pitch
- A sacrifice fly
- Other

So that's (almost) all that you need to know about baseball for CSCI S-5a.

Now let's learn how to calculate batting statistics.

Section 3: Batting Statistics

> **Main Idea:** *We can calculate statistics to measure batting performance.*

In this section, we'll see a variety of statistics that are used to quantify batting performance.

In general, batters want to get a hit, and want to avoid getting out.

Also, it's generally more desirable to get more bases, so a double is better than a single, a triple is better than a double, and a home run is best of all.

Some players are better at getting hits and not getting out, and some players tend to get more bases.

An enormous amount of time and energy has gone into trying to quantify this skill, and a variety of performance metrics have been developed.

In CSCI S-5a, we will usually be interested in the lifetime performance of a player, so all the numbers that we use are the totals over a player's career.

There are 8 basic observed quantities that we need to know in order to measure a player's batting strength:

- The total number of outs, denoted "O".
- The total number of singles, denoted "1B".
- The total number of doubles, denoted "2B".
- The total number of triples, denoted "3B".
- The total number of home runs, denoted "HR".
- The total number of bases on balls, denoted "BB".
- The total number of times a player was hit by a pitch, denoted "HBP".
- The total number of sacrifice flies, denoted "SF".

Once we know these 8 values, then we can define some other statistics.

The total number of hits, denoted "H", is the sum of the number of singles, doubles, triples, and home runs:

$$\mathrm{H} \;=\; 1\mathrm{B} + 2\mathrm{B} + 3\mathrm{B} + \mathrm{HR}$$

The number of "At-Bats", denoted "AB", is defined as the sum of the number of hits and the number of outs:

$$\mathrm{AB} \;=\; \mathrm{H} + \mathrm{O}$$

Notice that the at-bats don't include every possible plate appearance; for instance, a player might get a base on balls, and that's neither a hit or an out, so it's not counted as an "at-bat".

Finally, the number of *total bases*, denoted by “TB”, is defined as the sum of the number of singles, two times the number of doubles, three times the number of triples, and four times the number of home runs:

$$\text{TB} = \text{1B} + (2 \times \text{2B}) + (3 \times \text{3B}) + (4 \times \text{HR})$$

Now we can define the standard performance metrics for baseball batting.

All of the performance metrics involve calculating the ratio of some measure of the number of successes divided by some measure of the number of attempts, although there are different definitions of “successes” and “attempts”.

The traditional measure of batting skill is the *batting average*, denoted “BA”, is defined as the ratio of the number of hits divided by the number of at-bats:

$$\text{BA} \;=\; \frac{\text{H}}{\text{AB}}$$

The rationale for the batting average is that we want to measure how often a player can attempt to get a hit and actually succeed, so the numerator is the number of successes (i.e. the number of hits) and the denominator is the number of attempts (i.e. the number of at-bats).

Notice that the batting average excludes bases on balls and being hit by a pitch, even though in the end the batter still ends up standing on first base.

Many people think that this is unreasonable, because all that matters is whether or not the player ends up standing on first base, not how he or she got there.

Instead, we should measure often a player ends up getting a base out of all attempts, so that this includes bases on balls and being hit by a pitch as well as getting a hit.

Here a “success” is either a hit, a base on balls, or being hit by a pitch, all of which result in the player ending up standing on a base or getting a home run.

On the other hand, an “attempt” is defined to be a hit, an out, a base on balls, being hit by a pitch, or a sacrifice fly.

This measure is called the “On-Base Percentage”, and is denoted “OBP”:

$$\text{OBP} \;=\; \frac{\text{H} + \text{BB} + \text{HBP}}{\text{AB} + \text{BB} + \text{HBP} + \text{SF}}$$

Finally, we might also want to quantify how many bases a batter can get on average.

Here, we quantify “success” by the number of total bases, while the attempts are the number of at-bats.

Then the “Slugging Percentage”, denoted “SLG”, is defined as the total bases divided by the number of at-bats:

$$\text{SLG} \;=\; \frac{\text{TB}}{\text{AB}}$$

So that’s how we can calculate a variety of statistics that are used to quantify batting performance.

Now let’s learn how to properly report baseball statistics.

Section 4: Reporting Baseball Statistics

Main Idea: *We can report baseball statistics in the standard format.*

In this section, we'll see how to format baseball performance statistics properly.

In baseball, there is a standard way to report batting statistics.

Batting statistics are *always* reported as a decimal number, always with three decimal places to the right of the decimal point.

To achieve this, we have to use the `formatC()` function.

Let's say that a player gets 23 hits out of 87 at-bats.

Then we can format the batting average:

```r
# Example 1: Formatting a batting average

formatC(
    23 / 87,
    format = "f",
    digits = 3
)
```

```
## [1] "0.264"
```

Actually, this *still* isn't correct, because in practice the leading zero is suppressed, so it should really be ".264".

We'll explore how to properly format this value in Lecture 8, but if you want to explore this now you should read the documentation for the `substr()` function.

So that's how to format baseball performance statistics properly.

Now let's see an example of these ideas.

Section 5: Babe Ruth

> **Main Idea:** *We can calculate baseball batting statistics for Babe Ruth.*

In this section, we'll calculate the performance statistics for the great Babe Ruth.

First, here are Babe Ruth's career statistics:

Statistics	Value
Plate appearances	10,626
At-bats	8,399
Hits	2,873
Doubles	506
Triples	136
Home runs	714
Bases on balls	2,062
Hit by a pitch	43
Sacrifice flies	0

Let's store these values in variables:

```r
# Example 2: Babe Ruth batting statistics

babe.ruth.at.bats <- 8399

babe.ruth.hits <- 2873
babe.ruth.doubles <- 506
babe.ruth.triples <- 136
babe.ruth.home.runs <- 714

babe.ruth.bases.on.balls <- 2062
babe.ruth.hit.by.a.pitch <- 43
babe.ruth.sacrifice.flies <- 0
```

We can calculate the batting average:

```r
# Example 3: Babe Ruth batting average

babe.ruth.batting.average <-
    babe.ruth.hits / babe.ruth.at.bats

cat(
    "Babe Ruth batting average:",
    formatC(
        babe.ruth.batting.average,
        format = "f",
        digits = 3
    )
)
```

```
## Babe Ruth batting average: 0.342
```

Now let's calculate Babe Ruth's on-base percentage:

```r
# Example 4: Babe Ruth's on-base percentage

babe.ruth.on.base.percentage <-
    (babe.ruth.hits +
         babe.ruth.bases.on.balls +
         babe.ruth.hit.by.a.pitch) /
    (babe.ruth.at.bats +
         babe.ruth.bases.on.balls +
         babe.ruth.hit.by.a.pitch +
         babe.ruth.sacrifice.flies )

cat(
    "Babe Ruth on-base percentage:",
    formatC(
        babe.ruth.on.base.percentage,
        format = "f",
        digits = 3
    )
)
```

```
## Babe Ruth on-base percentage: 0.474
```

For the slugging percentage, we have a small problem.

Notice that the data that I provided did *not* include the number of singles.

Instead, it reported the total number of hits, along with the number of doubles, triples, and home runs.

We can re-arrange our formula to determine the number of singles:

$$\text{1B} = \text{H - 2B - 3B - HR}$$

Now we can compute the number of singles:

```r
# Example 5: Calculating Babe Ruth's singles

babe.ruth.singles <-
    babe.ruth.hits -
    babe.ruth.doubles -
    babe.ruth.triples -
    babe.ruth.home.runs
```

Now we can calculate the number of total bases:

```r
# Example 6: Calculating Babe Ruth's total bases

babe.ruth.total.bases <-
    1 * babe.ruth.singles +
    2 * babe.ruth.doubles +
    3 * babe.ruth.triples +
    4 * babe.ruth.home.runs
```

Finally, we can calculate Babe Ruth's slugging percentage:

```r
# Example 7: Calculating Babe Ruth's slugging percentage

babe.ruth.slugging.percentage <-
    babe.ruth.total.bases / babe.ruth.at.bats

cat(
    "Babe Ruth slugging percentage:",
    formatC(
        babe.ruth.slugging.percentage,
        format = "f",
        digits = 3
    )
)
```

```
## Babe Ruth slugging percentage: 0.690
```

So that's how to calculate baseball batting statistics for the legendary Babe Ruth.

Now let's review what we've learned in this module.

Exercise 2.1: Ted Williams Batting Statistics

Here are Ted Williams's career statistics:

Statistics	Value
Plate appearances	9,792
At-bats	7,706
Hits	2,654
Doubles	525
Triples	71
Home Runs	521
Bases on balls	2,021
Hit by a pitch	39
Sacrifice flies	20

Part (a)

Calculate Ted Williams' career batting average.

Part (b)

Calculate Ted Williams' career on-base percentage.

Part (c)

Calculate Ted Williams' career slugging average.

Module Review

In this module, we learned how to use R to calculate baseball performance metrics.

- In Section 1, we'll discuss why we're spending time studying this topic.
- In Section 2, we learned about the game of baseball.
- In Section 3, we saw a variety of statistics that are used to quantify batting performance.
- In Section 4, we saw how to format baseball performance statistics properly.
- In Section 5, we calculated the performance statistics for the great Babe Ruth.

Now that you've finished this module, you should be able to:

- Explain why we're spending so much time on baseball statistics.
- Explain the basic facts about the game of baseball.
- Calculate the batting average, on-base percentage, and slugging average for a player.
- Correctly format baseball performance statistics.

We didn't meet any new built-in R functions in this module.

Solutions to the Exercises

Exercise 2.1: Ted Williams Batting Statistics

Here are Ted Williams's career statistics:

Statistics	Value
Plate appearances	9,792
At-bats	7,706
Hits	2,654
Doubles	525
Triples	71
Home Runs	521
Bases on balls	2,021
Hit by a pitch	39
Sacrifice flies	20

Part (a)

Calculate Ted Williams' career batting average

Solution

Let's start by creating a set of variables to hold Williams' career batting statistics:

```
ted.williams.at.bats <- 7706

ted.williams.hits <- 2654

ted.williams.doubles <- 525

ted.williams.triples <- 71

ted.williams.home.runs <- 521

ted.williams.bases.on.balls <- 2021

ted.williams.hit.by.a.pitch <- 39

ted.williams.sacrifice.flies <- 20
```

Now let's calculate the batting average for Ted Williams:

```
ted.williams.batting.average <- ted.williams.hits / ted.williams.at.bats

cat( "Ted Williams batting average:",
     formatC(
         ted.williams.batting.average,
         format = "f",
         digits = 3
     )
)
```

```
## Ted Williams batting average: 0.344
```

Part (b)

Calculate Ted Williams' career on-base percentage

Solution

Next, let's calculate Ted Williams's on-base percentage:

```
ted.williams.on.base.percentage <-
    (ted.williams.hits +
         ted.williams.bases.on.balls +
         ted.williams.hit.by.a.pitch) /
    (ted.williams.at.bats +
         ted.williams.bases.on.balls +
         ted.williams.hit.by.a.pitch +
         ted.williams.sacrifice.flies)

cat( "Ted Williams on-base percentage:",
     formatC(
         ted.williams.on.base.percentage,
         format = "f",
         digits = 3
     )
)
```

```
## Ted Williams on-base percentage: 0.482
```

Part (c)

Calculate Ted Williams' career slugging percentage.

Solution

Remember that we have to first compute the number of singles before we can calculate the total bases:

```
ted.williams.singles <-
    ted.williams.hits -
    ted.williams.doubles -
    ted.williams.triples -
    ted.williams.home.runs

ted.williams.total.bases <-
    ted.williams.singles +
    2 * ted.williams.doubles +
    3 * ted.williams.triples +
    4 * ted.williams.home.runs
```

Finally, we can calculate Ted Williams's slugging percentage:

```
ted.williams.slugging.percentage <-
    ted.williams.total.bases / ted.williams.at.bats

cat( "Ted Williams slugging percentage:",
     formatC(
         ted.williams.slugging.percentage,
         format = "f",
         digits = 3
     )
)
```

```
## Ted Williams slugging percentage: 0.634
```