Chapter 4

Statistical Methods

4.1 Introduction

In an ideal world, we would have unlimited amount of data, and all relevant questions could be answered with certainty. Is Tom Brady better than Peyton Manning? Have them played hundreds and hundreds of games with the same teammates against similar opponents and analyze the results. Of course, in the real world, this is not possible, and we have to base our analyses on the available data.

Statistical methods play at least two roles in these situations. First, they provide methods for extracting the maximum amount of information from a dataset. Second, they give us a way to quantify the uncertainty that results from having to base these conclusion on such limited data. The goal of this chapter is to give an overview of statistical reasoning and the type of statistical methods that are useful in analyzing sports data.

4.2 Using the Margin of Error to Quantify the Variation in Sports Statistics

One goal of analytic methods is to use data to draw conclusions about the process generating data. For instance, suppose we want to study the performance of an NFL running back. Let Y denote the yards gained on a particular carry and let $Y_1, Y_2, \ldots, Y_n$ denote the results of the running back's carries in a given season.

Then, Y can be modeled as a random variable; its distribution is a way to describe the running back's "true ability level," which plays a central role in his performance. This true ability level is unknown, but its properties are reflected in the observed data. A simple model for relating the observed result to this hypothetical true ability level is to assume that $Y_1, Y_2, \ldots, Y_n$ are independent random variables each with distribution of Y. More complicated models might take into account the fact that different carries occur in different situations (e.g., down, distance to first down) and against different defenses. However, for now, the focus is on the simple model.

Under the simple model, the probability distribution of Y completely characterize the running back's ability on running plays. Therefore, if we know this distribution, we can use it to help predict future results, to compare this player to other running backs, to aid in play calling, and so on. However, because the distribution is unknown, we must use the information in $Y_1, Y_2, \ldots, Y_n$ to learn about the probability distribution of Y. The practice of using sample data to determine the properties of the underlying probability distributions is knows as statistical estimation or simply as estimation.

In many cases, estimation can be based on the analogy principle. To estimate a characteristic of the distribution of Y, we use the corresponding characteristic of the data $Y_1, Y_2, \ldots, Y_n$. For instance, to estimate the mean of Y, we use the sample mean of $Y_1, Y_2, \ldots, Y_n$; to estimate the standard deviation of Y, we use the sample standard deviation of $Y_1, Y_2, \ldots, Y_n$, and so on.

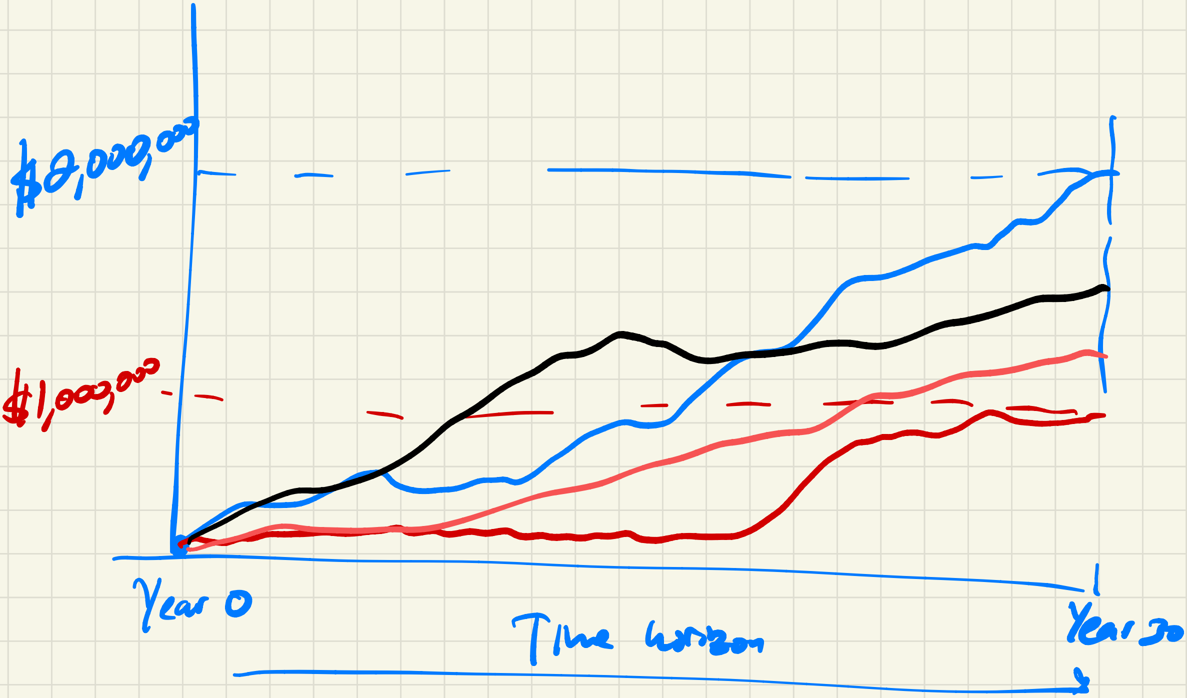
Suppose the running back in question is Jamaal Charles in the 2010 NFL season. We can estimate properties of the distribution of Y. For instance, the sample mean of the observed Y values is 6.38 yards per carry; therefore, our estimate of the mean of Y is 6.38 yards per carry. Similarly, the sample median of the observed Y values is 4 yards, yielding an estimate of the median of Y.

When interpreting results of this type, it is important to recognize the inherent variability in sports results. For instance, if we were somehow able to replay the 2010 NFL season, we would not expect Charles to have a yards-per-carry value of exactly 6.38 again, although we would expect it to be "close to 6.38." Furthermore, because of this variability, we know that the observed statistics do not exactly measure a player's true ability; for instance, we do not expect Charles's "true yards per carry" for 2010 to be exactly 6.38 yards, although, again, we expect that it will be close to 6.38 yards. Here we can think of his true yards per carry as the yards per carry that would be obtained of we were able to observe a large number of rushing attempts by Charles; alternatively, we can think of it as the mean of a random variable Y representing the results of a Charles rushing attempts.

Of course, simply saying that if we were to replay the 2010 NFL season, Charles's yards per carry would be close to 6.38 yards or his true yards per carry is close to 6.38 yards is not enough; we need a quantitative measure of what *close to* means. In both cases, this is given by a statistical measure known as the *margin of error*.

To understand what the margin of error is measuring, consider the following experiment: suppose we simulate Charles's performance during the 2010 NFL season by randomly choosing the results of 230 rushing attempts; recall that in the actual 2010 NFL season, Charles had 230 rushing attempts. An important consideration in carrying out this experiment is the probability distribution to use for the simulation. The best information we have about the distribution of the yards gained on Charles's rushing attempts is the actual results. Therefore, to simulate the result of his first rushing attempt, we randomly select one result from his 230 actual results; to simulate the result of his second rushing attempt, we select another result from his 230 actual results, and so on. Doing this 230 times gives us a simulated season for Charles; when I carried such a process, I observed a simulated yards per carry of 6.54 yards.

The difference between the simulated value of 6.54 and the actual value of 6.38 gives us important information regarding the natural variability in Charles's yards per carry. However, we could obtain more information by simulating the 2010 season several times; I did this, obtaining yards-per-carry values of 6.37, 6.29, 6.15, 5.71, 7.21 and so on. The margin error can be viewed


$2,000,000
$1,000,000
Year 0
Time horizon
Year 30

as a measure of the variability of a large number of such simulated values; specifically, it is two times the standard deviation of a long sequence of simulated value. In the example of Charles's yards per carry, it is 1.16 yards.

We can interpret the margin of error in one of two ways. The most straightforward interpretation is based on the idea of repeating the "experiment" (e.g., a season, a game, or a career) and calculating a range of values that we expect to include the statistic of interest. For instance, in Charles's rushing example, if we were to repeat 2010 NFL season, we would expect Charles's yards-per-carry value to fall in the range

$$\overset{\text{estimation}}{6.38} \pm \overset{\text{margin of error}}{1.16} = (5.22,\ 7.54)$$

Therefore, the interval (5.22, 7.54) summarizes the variability in Charles's yards-per-carry value. A second interpretation, based on the underlying "true" characteristic of a player or team, is a little more subtle, but it is often more useful. Recall that a statistical estimate is our "best guess" of the characteristic of a probability distribution. However, we know that the estimate is not equal to that characteristic. For example, in the example for Charles, his observed average yards per carry is 6.38 yards; as noted, we expect this to be close to the true mean value of Y, but it is unlikely that it will be exactly equal to it. The observed value of 6.38 is an estimate of this true mean value.

The margin of error can also be interpreted in terms of how close we expect this hypothetical true mean value to be to the observed value of 6.38. The interval

$$\overset{\bar{X}}{6.38} \pm \overset{\text{M.E.}}{1.16} = (5.22,\ 7.54)$$

gives a range of values such that we are "reasonable certain" that Charles's true average yards per carry lies in that range. At this point, it is reasonable to ask why we are interested in the true mean yards per carry or in a range of results that might occur of the season is repeated. In fact, for some purposes we are not particularly concerned with these hypothetical values. This might be the case, for example, if we are just trying to summarize a player's or team's performance.

However, in other cases, we are primarily interested in understanding the process that generated the data to compare players or teams or to better understand what might happen in the future. In these cases, the underlying true values that are of interest and the variability of estimates are an important consideration both in understanding how future results might relate to the available data and in determining how much confidence we should place in the conclusions of the analysis. Either interpretation of the margin error gives us important information about the variability of estimates and their relationship to the hypothetical true value.

For example, suppose we are analyzing Kevin Durant's scoring for the 2011-2012 NBA season. Let X denote Durant's points scored for a given game, and suppose we model X as a random variable with some probability distribution. Although different aspects of this distribution might be of interest, depending on the context, here we focus on the mean of X, which we denote by μ. The available data are the points scored for the 66 games Durant played, and using the analogy principle, our estimate of μ is simply the average of the 66 values, 28, Durant's points per game (PPG) for the 2011-2012 regular season.

Although 28 exactly represents what actually occurred in the 2011-2012 season, it is only an estimate of μ. Durant's "true PPG," that is, the PP~~E~~G he would achieve in a hypothetical long sequence of games. The uncertainty in this estimate is described by the margin error. Here the margin of error is 1.7; how this was determined is discussed in the following section. Therefore, our estimate of μ, Durant's true PPG for 2011-2012 season, could be reported as

$$\text{average} \rightarrow 28 \pm 1.7 \leftarrow \text{margin of error}$$

or as the interval (26.3, 29.7). That is, if we were somehow able to observe a long sequence of Durant's games, under the conditions of the 2011-2012 season, our best guess for his average PPG in this sequence is 28, and we are reasonably certain that it would fall between 26.3 and 29.7. Alternatively, we can use the interpretation of the margin of error in terms of repeating the season. That is, if we were able to repeat the 2011-2012 NBA season, we would expect Durant's PPG value to lie in the range 26.3 to 29.7.

There are two basic approaches to calculating the margin of error. The first is to use statistical formulas that are designed for this purpose. This approach is discussed in Section 4.3. The second approach is based on the interpretation of the margin of error in terms of hypothetical repetitions of the experiment. That is, we can compute the margin of error using compute simulation to obtain these hypothetical repetitions; this approach is discussed in Section 4.4.

4.3 Calculating the Margin of Error of Averages and Related Statistics

Although we will not generally calculate the margin of error by hand, using statistical formulas, it is useful to consider the margin-of-error formula in a few simple cases to better understand the issues that drive the accuracy of an estimate.

Consider the Durant scoring example from the previous example. Let X denote Durant's points scored for a given game, and suppose we are interested in μ, the mean of X. Our estimate of μ is the sample mean of Durant's point scored for the 66 games he played. The formula for the margin error in this case, in which we are estimating the mean of a random variable using a sample mean, is

$$\frac{2S}{\sqrt{n}} \tag{4.1}$$

where S denotes the sample standard deviation of his game-by-game points scored and n denotes the number of data values used in the sample average. For example, $n = 66$ and S can be calculated by examining his game-by-game statistics, which show that $S = 6.9$. Thus, the margin of error is