Bayes' Theorem In Baseball

MAT 380 080
November 17, 2014
Christopher Amici

Professional baseball has been around since 1869, when the Cincinnati Red Stockings were founded, and people probably have been trying to predict baseball games since than. In *The Celebrant* by Eric Rolfe Greenberg, people gambling on baseball, down to wheter the pitch will be a ball or a strike, is a major component of the book. Using probability one can predict those balls and strikes.

Predicting the balls and strikes is more complicated than it sounds. If it was just about the pitcher throwing it in the strike zone or not it would be a 50% chance of him throwing a ball and 50% chance of him throwing a strike. There is a batter, however, and an umpire. The batter can swing at a pitch that is a ball and miss, making it a strike. Maybe the umpire made a mistake and said it was a ball when it was a strike. Those situations and more need to be taken into account. Thanks to Thomas Bayes, we have some help with that.

Thomas Bayes was ordained a Nonconformist minister at Tunbridge Wells in 1694. In 1719, Bayes studied logic and theology at the University of Edinburgh. There however is no evidence he studied

mathematics at the University of Edinburgh but there is evidence that he had an opportunity to. In 1752 Bayes left the ministry and by 1764 had his probability theory on *Essay Towards Solving a Problem in the Doctrine of Chances* published in the *Philosophical Transactions of the Royal Society of London*. This probability theory has become known as Bayes' Theorem and it involves conditional probability. Bayes' Theorem is what will be used to find the probability a pitcher will throw a strike under certain given condition.

Before getting into Bayes' Theorem, there are a few things that need to be defined.

1. Background Defintions, Lemmas, Theorems, and Corollaries

Let S denote the sample space, E, E_i, F, etc. events and the notation $Pr(\cdot)$ the probability of whatever.

Definition: An event, E, is trivial *iff* $Pr(E) = 0$ or $Pr(E) = 1$

The Space Axiom: S is the space $\Rightarrow Pr(S) = 1$.

The Event Axiom: E is an event $\Rightarrow 0 \leq Pr(E) \leq 1$.

The Trivial Event: An event, E, is trivial *iff* $Pr(E) = 0$ or $Pr(E) = 1$

The Collection of Mutually Exclusive Events Axiom: Let I be an index set. If the collection $\{E_i\}_{i \in I}$ being mutually exclusive, then $\Pr(\bigcup_{i \in I} E_i) = \sum_{i \in I} \Pr(E_i)$.

Lemma: Let S be a well defined sample space; E and F are events.

(1) E^c is an event.

(2) $E \cap F$ is an event.

(3) $E \cup F$ is an event.

(4) $E - F$ is an event.

The Null Corollary: Let S be a well defined sample space whilst E $= \emptyset$. It is the case that $Pr(E) = 0$.

The Complement Corollary: If E is an event, then $Pr(E^c) = 1 - Pr(E)$.

The Subset Corollary: If E and F are events such that $E \subseteq F$, then $Pr(E) \leq Pr(F)$.

The Spanning Corollary: Let I be an index set. The collection $\{E_i\}_{i \in I}$ being mutually exclusive and exhaustive, then $\Pr(\bigcup_{i \in I} E_i) = \sum_{i \in I} \Pr(E_i) = 1$.

The Same Event Corollary: E and F are events such that $E = F$, then $Pr(E) = Pr(F)$.

The Union Theorem: E and F are events, then $Pr(E \cup F) = Pr(E) + Pr(F) - Pr(E \cap F)$.

Conditional Probability: The probability of some event A given an event B already occurred is defined as

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

2. Bayes' Theorem

In Probability Theory we always begin with a sample space (a universe in Set Theory); events (well defined sets); and outcomes (elements of sets).

In order to work with Bayes' Theoerm we have to define some more

terms. So, let S be a sample space space and let there exist the collection Ω be a collection of non-empty events such that the collection is finite meaning $|\Omega| = n$ where $n \in \mathbb{N}$.

Without loss of generality let $\Omega = \{E_1, E_2, ..., E_{n-1}, E_n\}$. Ω partitions the sample space S if the events are pairwise disjoint and the genrealized union of Ω is S meaning $E_m \cap E_p = \emptyset$ when $m \neq p$ and $\bigcup_{k=1}^{n}(E_k) = \bigcup \Omega = S$

Theorem: Let S be the sample space, let $\{E_1, E_2, ..., E_{n-1}, E_n\}$ be a finite collection of non-empty events and let F be a non-trivial event. It is the case that the probability that the event E_m occurs first given that the event F happens second is the quotient of the probability E_m occurs first times the probability that F occurs second given E_m occurs first over the sum of the probabilities of each of the events E_j occurs first times the probability that the event F happens second given that the event E_j happens first. In mathematicval notation that is:

$$Pr\left(E_m|F\right) = \frac{Pr\left(E_m\right) \cdot Pr\left(F|E_m\right)}{\sum_{i=1}^{n}(Pr\left(E_i\right) \cdot Pr\left(F|E_i\right))}$$

Let the notation $Pr(E_{pFirst})$ mean the probability the event E_p happens first; let the notation $Pr(F_{Second})$ be the be the probability that the event F happens; let the notation $Pr(E_{pFirst}|F_{Second})$ be the probability of the event E_p occurs first given the event F happens second; and, let the notation $Pr(F_{Second}|E_{pFirst})$ be the probability of the event F occurs second given the event E_p happens first. $E_{1First}, ..., E_{nFirst}$ are all the other events that could of happened first given that F_{Second} happened second.

$$Pr\left(E_{pFirst}|F_{Second}\right) = \frac{Pr\left(E_{pFirst}\right) Pr\left(F_{Second}|E_{pFirst}\right)}{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{1First}\right) + ... + Pr\left(E_{nFirst}\right) Pr\left(F_{Second}|E_{nFirst}\right)}$$

Where E_{pFirst} is the event, that happened first, that you want to find the probability of given that F_{Second} happened second. $E_{1First}, ..., E_{nFirst}$ are all the other events that could of happened first given that $FSecond$ happened second.

Proof:

$n = 2$:

Then, $\Omega = \{E_1, E_2\}$, $E_1 \cap E_2 = \emptyset$, $E_1 \cup E_2 = S$, $E_1 \neq \emptyset$, $E_2 \neq \emptyset$, and $Pr\left(E_1\right) \neq 0$ and $Pr\left(E_2\right) \neq 0$.

Since the expression can be used to find the probability of E_1 or E_2, let E_1 be the $p = 1$.

Consider the expression:

$$\frac{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{1First}\right)}{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{1First}\right) + Pr\left(E_{2First}\right) Pr\left(F_{Second}|E_{2First}\right)}$$

$$= \frac{Pr\left(E_{1First}\right) \frac{Pr(F_{Second} \cap E_{1First})}{Pr(E_{1First})}}{Pr\left(E_{1First}\right) \frac{Pr(F_{Second} \cap E_{1First})}{Pr(E_{1First})} + Pr\left(E_{2First}\right) \frac{Pr(F_{Second} \cap E_{2First})}{Pr(E_{2First})}}$$

$$= \frac{Pr\left(F_{Second} \cap E_{1First}\right)}{Pr\left(F_{Second} \cap E_{1First}\right) + Pr\left(F_{Second} \cap E_{2First}\right)}$$

$$= \frac{Pr\left(E_{1First} \cap F_{Second}\right)}{Pr\left(E_{1First} \cap F_{Second}\right) + Pr\left(E_{2First} \cap F_{Second}\right)}$$

$$= \frac{Pr\left(F_{second}\right) Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(F_{Second}\right) Pr\left(E_{1First}|F_{Second}\right) + Pr\left(F_{Second}\right) Pr\left(E_{2First}|F_{Second}\right)}$$

$$= \frac{Pr\left(F_{second}\right) Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(F_{Second}\right)\left(Pr\left(E_{1First}|F_{Second}\right) + Pr\left(E_{2First}|F_{Second}\right)\right)}$$

$$= \frac{Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(E_{1First}|F_{Second}\right) + Pr\left(E_{2First}|F_{Second}\right)}$$

$$= Pr\left(E_{1First}|F_{Second}\right)$$

Therefore, the equation is true for $n = 2$.

$n = 3$:

Then, $\Omega = \{E_1, E_2, E_3\},\ E_1 \cap E_2 \cap E_3 = \emptyset,\ E_1 \cup E_2 \cup E_3 = S,$ $E_1 \neq \emptyset, E_2 \neq \emptyset, E_3 \neq \emptyset$, and $Pr(E_1) \neq 0, Pr(E_2) \neq 0$, and $Pr(E_3) \neq 0$.

Since the expression can be used to solve the probability of E_1, E_2, or E_3 let E_1 be the $p = 1$.

Consider the expression:

$$\frac{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{1First}\right)}{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{1First}\right) + Pr\left(E_{2First}\right) Pr\left(F_{Second}|E_{2First}\right) + Pr\left(E_{3First}\right) Pr\left(F_{Second}|E_{3}\right.}$$

$$= \frac{Pr\left(E_{1First}\right) \frac{Pr(F_{Second} \cap E_{1First})}{Pr(E_{1First})}}{Pr\left(E_{1First}\right) \frac{Pr(F_{Second} \cap E_{1First})}{Pr(E_{1First})} + Pr\left(E_{2First}\right) \frac{Pr(F_{Second} \cap E_{2First})}{Pr(E_{2First})} + Pr\left(E_{3First}\right) \frac{Pr(F_{Second} \cap E_{3First})}{Pr(E_{3First})}}$$

$$= \frac{Pr\left(F_{Second} \cap E_{1First}\right)}{Pr\left(F_{Second} \cap E_{1First}\right) + Pr\left(F_{Second} \cap E_{2First}\right) + Pr\left(F_{Second} \cap E_{3First}\right)}$$

$$= \frac{Pr\left(E_{1First} \cap F_{Second}\right)}{Pr\left(E_{1First} \cap F_{Second}\right) + Pr\left(E_{2First} \cap F_{Second}\right) + Pr\left(E_{3First} \cap F_{Second}\right)}$$

$$= \frac{Pr\left(F_{second}\right) Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(F_{Second}\right) Pr\left(E_{1First}|F_{Second}\right) + Pr\left(F_{Second}\right) Pr\left(E_{2First}|F_{Second}\right) + Pr\left(F_{Second}\right) Pr\left(E_{3First}|\right.}$$

$$= \frac{Pr\left(F_{second}\right) Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(F_{Second}\right)\left(Pr\left(E_{1First}|F_{Second}\right) + Pr\left(E_{2First}|F_{Second}\right) + Pr\left(E_{3First}|F_{Second}\right)\right)}$$

$$= \frac{Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(E_{1First}|F_{Second}\right) + Pr\left(E_{2First}|F_{Second}\right) + Pr\left(E_{3First}|F_{Second}\right)}$$

$$= Pr\left(E_{1First}|F_{Second}\right)$$

Therefore, the equation is true for $n = 3$.

$n \geq 3$:

Then, $\Omega = \{E_1, ..., E_n\}$, $E_1 \cap ... \cap E_n = \emptyset$, $E_1 \cup ... \cup E_n = S$, $E_1 \neq \emptyset, ..., E_n \neq \emptyset$, and $Pr(E_1) \neq \emptyset, ..., Pr(E_n) \neq \emptyset$.

Since the expression can be used to find the probability of E_1,...E_n, let E_1 be the $p = 1$.

Consider the expression:

$$\frac{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{pFirst}\right)}{Pr\left(E_{1First}\right) Pr\left(F_{Second}|E_{1First}\right) + ... + Pr\left(E_{nFirst}\right) Pr\left(F_{Second}|E_{nFirst}\right)}$$

$$= \frac{Pr\left(E_{1First}\right) \frac{Pr(F_{Second} \cap E_{1First})}{Pr(E_{1First})}}{Pr\left(E_{1First}\right) \frac{Pr(F_{Second} \cap E_{1First})}{Pr(E_{1First})} + ... + Pr\left(E_{nFirst}\right) \frac{Pr(F_{Second} \cap E_{2First})}{Pr(E_{2First})}}$$

$$= \frac{Pr\left(F_{Second} \cap E_{1First}\right)}{Pr\left(F_{Second} \cap E_{1First}\right) + ... + Pr\left(F_{Second} \cap E_{nFirst}\right)}$$

$$= \frac{Pr\left(E_{1First} \cap F_{Second}\right)}{Pr\left(E_{1First} \cap F_{Second}\right) + ... + Pr\left(E_{nFirst} \cap F_{Second}\right)}$$

$$= \frac{Pr\left(F_{second}\right) Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(F_{Second}\right) Pr\left(E_{1First}|F_{Second}\right) + ... + Pr\left(F_{Second}\right) Pr\left(E_{2First}|F_{Second}\right)}$$

$$= \frac{Pr\left(F_{second}\right) Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(F_{Second}\right) \left(Pr\left(E_{1First}|F_{Second}\right) + Pr\left(E_{nFirst}|F_{Second}\right)\right)}$$

$$= \frac{Pr\left(E_{1First}|F_{Second}\right)}{Pr\left(E_{1First}|F_{Second}\right) + ... + Pr\left(E_{nFirst}|F_{Second}\right)}$$

$$= Pr\left(E_{1First}|F_{Second}\right)$$

Therefore, the equation is true for $n \geq 3$. □

3. Application

Bayes' Theorem is a useful tool that is used in baseball by the Oakland Athletics, Boston Red Soxs, and Chicago Cubs. It is used to see how baseball players do in certain situations. It also anaylsis there statistics in previous years to predict how they will do in the future. So, in the ball or strike example, suppose that the event we want to find the probability of is the pitcher throws a first pitch strike. There are many things that need to be taken into account: if the pitcher tries to throw a strike, if the batter swings and misses, if the batter swings and gets a hit, if the batter gets hit by the pitch, if the umpire misses the call, if the pitcher plants his foot down wrong and falls to the ground in agonizing pain, etc.

The probability that a pitcher throws a strike is going to be the percentage he threw a strike, the previous game, that year or even for his whole career. For example let's find the probability a pitcher will throw a strike given that it is the first pitch. There are many things that can be taken into account, which is the problem with using Bayes' theorem. The events that can happen do not exhaust. Let S be a strike, B be a ball, and F be past first pitches. The variables are then placed into the equation, so

$$Pr(S|F) = \frac{Pr(S)Pr(F|S)}{Pr(S)Pr(F|S) + Pr(B)Pr(F|B)}$$

$Pr(S)$ is the probability that he throws a strike, $Pr(F|S)$ is the probability that it he threw a strike on the first pitch in the past, $Pr(B)$ is

the probability he throws a ball, and $Pr(F|B)$ is the probability that he threw a ball on the first pitch in the past.

This equation can be applied to the performance of Cole Hamels, starting pitcher for the Philadelphia Phillies. Since the season is over, this equation is going to be used to find the probability that Cole Hamels throws a first pitch strike next season. Thanks to baseball-reference.com, the percentage of strikes thrown, balls thrown, and strikes thrown on the first pitch from the 2014 season can be used in the equation.Unforunatly, when it comes to pitch count, a ball hit in play is considered a strike so the numbers taken from baseball-reference.com include that in the percentage of strikes thrown.

Now Cole Hamels' career numbers can be put into the equation.

$$Pr(S|F) = \frac{(.662)(.614)}{(.662)(.614) + (.338)(.386)} \approx .757$$

So, based on Cole Hamels' performance in the 2014 season, there is about a 75.7 percent chance of him throwing a strike on the first pitch in the 2015 season. Notice that the probability of him throwing a strike on the first pitch is 14.3 percent higher than the probability he threw a strike on the first pitch last year. That being said, Cole Hamels will probably not throw a strike on the first pitch 75.7 percent of the time next season, bet if someone were to bet that he will throw a strike on the first pitch they would probably be right.

In 2006, Bobby Abreu was traded from the Phillies to the Yankees. Many Phillies fans were happy about that because, although Abreu had a good batting average with runners in scoring position,

fans claimed he would walk more in that situation, rather than get a hit. Using Bayes' theorem, the probability that Bobby Abreu would get a hit in 2006 given that a runner is in scoring postion can be found. Let H be a hit, W be a walk, O be an out, and R be runners in scoring position. The variables are then placed into the equation, so

$$Pr(H|R) = \frac{Pr(H)Pr(R|H)}{Pr(H)Pr(R|H) + Pr(W)Pr(R|W) + Pr(O)Pr(R|O)}$$

$Pr(H)$ is the probability he gets a hit, $Pr(R|H)$ is the probability he got a hit with runners in scoring position in the past, $Pr(W)$ is the probability he walks, $Pr(R|W)$ is the probability he walked with runners in scoring position in the past, $Pr(O)$ is the probability he gets out, and $Pr(R|O)$ is the probability he got out with runners in scoring position in the past. The out probability also includes sacrifice flies and outs that resulted in a runner scoring, so in some cases an out can be a good thing. So now the probabilities from 2006 before he got traded can be inserted into the equation,

$$Pr(H|R) = \frac{(.215)(.231)}{(.215)(.231) + (.208)(.2223) + (.578)(.545)} \approx .121$$

So in 2006 with the Phillies, Bobby Abreu had approximately a 12.1 percent chance of getting a hit with runners in scoring position. The equation for the probability he gets a walk when runners are in scoring position is

$$Pr(W|R) = \frac{Pr(W)Pr(R|W)}{Pr(W)Pr(R|W) + Pr(H)Pr(R|H) + Pr(O)Pr(R|O)}$$

Now the probabilities can be inserted into this equation,

$$Pr(W|R) = \frac{(.208)(.223)}{(.208)(.223) + (.215)(.231) + (.578)(.551)} \approx .113$$

So there was approximately a 11.3 percent chance of Bobby Abreu walking with runners in scoring position in 2006. The fans were wrong, Abreu had a better chance of getting a hit than walking with runners in scoring position, but let's compare this to the best player today to see why fans would think this.

That best position player in major league baseball today is considered to be Mike Trout, so using the same equation we can find the probability he would get a hit with runners in scoring position in the 2014 season.

$$Pr(H|R) = \frac{(.245)(.229)}{(.245)(.229) + (.118)(.159) + (.637)(.611)} \approx .121$$

The probability of Mike Trout getting a hit with runners in scoring postion is approximately the same as Bobby Abreu's in 2006, however Mike Trout might have less of a chance of walking with runners in scoring position.

$$Pr(W|R) = \frac{(.118)(.159)}{(.118)(.159) + (.245)(.229) + (.673)(.611)} \approx .040$$

This past season there was approximately a 4 percent chance of Mike Trout walking with runners in scoring position. This means that Mike Trout recorded more outs than Bobby Abreu, but the outs that Mike Trout recorded could have also resulted in runs. For example a sacrfice fly or a fielder's choice would be considered an out but can still advance

the runners. So Phillies fans were not entirely wrong about Bobby Abreu. He did not walk more than he got a hit with runners in scoring position, but compared to Mike Trout, Bobby Abreu walked more than he should have.

The problem with Bayes' Theorem however is that it cannot be easily applied. There are many other factors that were not taken into account, as stated earlier, like the batter or umpire. Also, defining a complete sample space is almost impossible in real life. For example, in the strike example, many things can happen other than a strike. The pitch can be a strike, ball, the batter can be hit by the pitch, a bird can fly over the plate, etc. This leads to so many events that the events do not exhaust, there are so many events that the entire sample space is not taken into account, which is why many mathematicians do not like applying Bayes' Theorem to real life events. The events also do not come out perfectly because of human error. Bayes' Theorem however is a helpful guide that should not be heavily relied on.

4. Glossary of Baseball Terms

At Bat: An offensive player is up to bat.

Ball: A pitch which does not enter the strike zone and is not struck at by the batter.

Batter: The offensive player who is currently positioned in the batter's box.

Bunt: A legally batted ball, not swung at but intentionally met with the bat and tapped within the infield.

Fair Ball: A legally batted ball that settles on or over fair territory.

Fair Territory: That part of the playing field within and including the first base and third base lines, from home plate to the playing field fence and perpendicularly upwards.

Fielder: One of the nine defensive players, including pitcher, catcher, first baseman, second baseman, third baseman, shortstop, left fielder, center fielder and right fielder.

Fielder's Choice: The act of a fielder who handles a fair grounder and, instead of throwing to first base to put out the batter runner, throws to another base in an attempt to put out a preceding runner.

Fly Ball: A bat results in a high-flying ball.

Foul Ball: A batted ball that lands on foul territory between home plate and first base or third base, bounds past first or third base on or over third territory, first touches foul territory beyond first or third base, or touches a player, umpire or any object not part of the playing field while over foul territory.

Foul Territory: That part of the playing field outside the first and third base lines extended to the outfield fence and perpendicularly upwards.

Out: A declaration by the umpire that a player who is trying for a base is not entitled to that base.

Pitch: The ball delivered by the pitcher to the batter.

Pitcher: The fielder designated to pitch the ball to the batter.

Run Batter In: Also known as "RBI", a record of points earned by a player for assisting his teammates in scoring points while up to bat.

Run: The score made by an offensive player who has rounded the bases and returned to home plate.

Runner: An offensive player who is advancing toward, touching or returning to any base.

Sacrifice Fly: A fly ball out and a runner scores a point.

Scoring Position: Runner is on second or third base.

Strike: A legal pitch when so called by the umpire, which:

(1) Is struck at by the batter and missed

(2) Is not struck at, if the ball passes through the strike zone

(3) Is fouled by the batter when he has less than two strikes

(4) Is bunted foul

(5) Touches the batter as he strikes at it

(6) Touches the batter in flight in the strike zone

(7) After being batted, travels directly from the bat to the catcher's hands and is legally caught by the catcher (foul tip)

Strike Zone: An area directly over home plate, from the bottom of the batter's kneecaps to the midpoint between the top of the batter's shoulders and the top of the batter's uniform pants.

Umpire: The official who judges the legality of individual plays and who otherwise enforces the rules of the game.

Walk: Also called "base on balls"; after four pitches are delivered the batter is allowed advance to first base, forcing other runners on base to advance as well.

Sources

Conversations with Dr. McLoughlin and lecture notes on conditional probability

Conversation with Mr. Dan Herlin

Baseball-reference.com, http://www.baseball-reference.com/ 11/10/2014

The MacTutor History of Mathematics archive http://www-history.mcs.st-andrews.ac.uk/ 11/13/2014

Sports Management Degrees http://www.sports-management-degrees.com/baseball/ 9/10/2014

PBS: Baseball for Beginners http://www.pbs.org/kenburns/baseball/beginners/glossary.html 11/21/2014

Epic Sports http://baseball.epicsports.com/baseball-glossary.html 11/21/2014