16.90: Project #3
Probabilistic Simulation of a Baseball Batter
Solution

1 Background

In this project you are the statistician for the Boston Red Sox. The batting and pitching coaches want to better understand the results of three different pitches—a high fastball, a sinking fastball, and a "12-6" curveball. Using a Monte Carlo analysis, you will determine the probability that these pitches will result in a ground ball, line drive, fly ball, and home run.

1.1 Baseball Dynamics

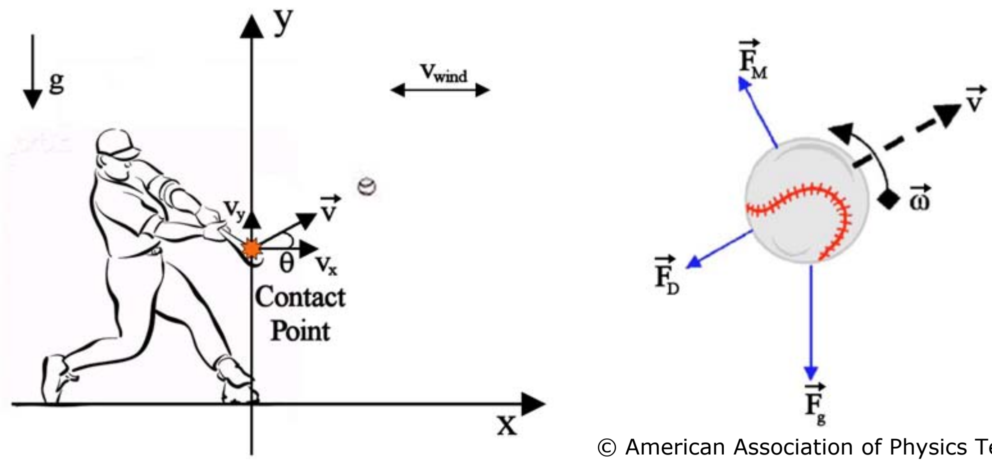

Figure 1: Schematic of the baseball dynamics after the ball has been hit, adapted from [1].

Provided for you is the function **bb_dyn.m** that calculates the trajectory of the baseball after it leaves the bat [1]. The inputs into this function are the velocity of the ball (v_{ball}, mph), the velocity of the bat (v_{bat}, mph), the angle the ball leaves the bat (θ, deg), the wind velocity (v_{wind}, mph), and the spin of the ball (ω, rpm). The total velocity of the baseball is a function of time and is defined as

$$v_{tot}(t) = \sqrt{v_x(t)^2 + v_y(t)^2}, \tag{1}$$

where v_x and v_y are the x and y-components of the velocity as shown in Figure 1. The total velocity of the ball as it leaves the bat, $v_{tot}(t=0)$, is determined by the equation

$$v_{tot}(0) = e_a v_{ball} + (1 + e_a) v_{bat}, \tag{2}$$

where e_a is the collision efficiency and is a function of the energy dissipation and recoil of the bat. We assume for this analysis that each pitch is hit by the batter. The trajectory of the baseball during flight is affected by the force due to the drag on the baseball ($\vec{F}_D$), the Magnus force due to the spin of the ball ($\vec{F}_M$), and the force due to gravity ($\vec{F}_g$) as shown in Figure 1. Note that the Magnus force acts in the $\vec{\boldsymbol{\omega}} \times \vec{\mathbf{v}}$ direction. During this analysis we assume that the trajectory of the ball is in the xy-plane and that there is no sidespin on the ball. Thus the spin, $\vec{\boldsymbol{\omega}}$, only has a z-component, which we call ω. The degrees of freedom of the

[1]Acknowledgements to Professor Stirling for writing the baseball dynamics analysis code.

baseball are the distance from the batter (x) and the height off the ground (y). The acceleration of the ball can then be written as

$$\left[\begin{array}{c} a_x \\ a_y \end{array}\right] = \frac{1}{m}\left[\begin{array}{c} -\frac{1}{2}\rho A C_d v_{tot}(v_x - v_{wind}) - \frac{1}{2}\rho A R C_m \omega v_y \\ -\frac{1}{2}\rho A C_d v_{tot} v_y + \frac{1}{2}\rho A R C_m \omega (v_x - v_{wind}) - mg \end{array}\right], \tag{3}$$

where v_x is the x-component of the velocity, v_y is the y-component of the velocity, v_{tot} is the magnitude of the velocity, g is gravity, C_d is the drag coefficient, C_m is the Magnus coefficient, A is the cross-sectional area of the ball, R is radius of the ball and ρ is the density of the air.

The code computes the trajectory of the ball as a function of time. The function returns the x-distance (ft) travelled by the ball from the bat, the y-distance (ft) travelled by the ball from the bat, and the time (sec), all for each timestep simulated.

Please see the comments in the Matlab code **bb_dyn.m**, which describe in detail the form of the function inputs and outputs.

1.2 Pitch Details and Batted Ball Categories

The fastball is a common type of pitch in baseball. Fastballs generally have backspin (i.e., a positive ω), which causes an upward force on the ball. A curveball is a breaking pitch with topspin (i.e., a negative ω) causing a downward force on the ball, which gives an increased drop in the motion of the ball. A "12-6" curveball has a downward action as it goes toward the plate.

There are four main categories used to describe batted balls—ground balls, line drives, fly balls, and home runs. For the purposes of this assignment, we will use the following definitions.

- Ground ball: A ball that rolls on the ground. The maximum height of the ball after it is hit never exceeds 4 ft.
- Line drive: This is a hard hit low-flying batted ball. The maximum height is less than 10 ft, but greater than 4 ft.
- Fly ball: A ball that is hit into the air, usually fairly high. For this project, consider a fly ball any ball that is hit such that the maximum height is greater than 10 ft, but is not a home run.
- Home run: A ball that cannot be caught in the field of play. We will assume that the centerfield fence is 400 feet away and 8 feet high.

1.3 Input Variability

We will assume that each of the inputs to the baseball dynamics code (v_{wind}, v_{bat}, v_{ball}, θ, and ω) can be modeled with a triangular distribution. The distribution parameters for each variable are given in Table 1. The ranges provided for the types of pitches correspond to the pitch definitions. For example, a high fastball will typically have a higher angle of the ball leaving the bat with a positive spin on the ball and a sinking fastball will typically be hit at a lower angle off the bat.

Table 1: Input variabilities are modeled with triangular distributions, where x_{min} is the minimum value, x_{mpp} is the most probable value, and x_{max} is the maximum value.

Variables	x_{min}	x_{mpp}	x_{max}
v_{wind} (mph)	-25	0	25
v_{bat} (mph)	0	78	100
High Fastballs			
v_{ball} (mph)	86	90	100
θ (deg)	20	35	89
ω (rpm)	0	2000	4000
Sinking Fastballs			
v_{ball} (mph)	86	90	100
θ (deg)	-15	0	15
ω (rpm)	0	2000	4000
Curveballs			
v_{ball} (mph)	67	77	92
θ (deg)	-15	10	50
ω (rpm)	-4000	-2000	0

2 Project Requirements

2.1 Estimation of the Batted Ball Results

Implement a Monte Carlo analysis to determine the probability of ground balls, line drives, fly balls, and home runs for each pitch type. Specifically,

1. Develop well-commented Matlab scripts to implement the Monte Carlo analysis for this problem.
   - You will need to insert code into the function **trirnd.m** to draw a random sample from a triangular distribution.
   - You will need to write a code to conduct the Monte Carlo simulation and perform the required analyses.
   - The function **bb_dyn.m** will be called inside your Monte Carlo loop but does not need to be modified.

2. Estimate the probability of of a ground ball, line drive, fly ball, and home run for each of the three pitch types. For each probability estimate (you will have 12 estimates) choose a sample size so that your estimate is at least ± 0.01 at 99% confidence.

   Solution: For each pitch type we run a Monte Carlo simulation. For each draw of the random inputs, we record the resulting hit type (ground ball, line drive, fly ball, or home run). The probability of each hit type is then estimated by computing the number of occurrences of that hit divided by the total number of samples. To determine the necessary Monte Carlo sample size, we need to consider the statistical properties of the probability estimator. Denote the probability we are trying to estimate by p, and our estimator by $\hat{p}$. As discussed in class, $\hat{p}$ is a random variable that has a normal distribution. The expected (mean) value of the probability estimator is given by the probability itself, $\mathbb{E}[\hat{p}] = p$ (i.e., it is an unbiased estimator). The variance of the probability estimator is given by $Var[\hat{p}] = p(1-p)/N$, where N is the number of samples in our Monte Carlo simulation.

   To obtain a probability estimate within ± 0.01 with 99% confidence, we require that

   $$3\sqrt{\frac{p(1-p)}{N}} < 0.01,$$

   where the 3 comes from the specification of the 99% confidence level for a normal distribution (to be more precise, we could use 2.58, but rounding to 3 will be a conservative estimate).

Now, we do not know p, but we can see that the variance of the estimator will be maximum in the case that $p = 0.5$. If we consider this worst case, then we will obtain a conservative value of N that we can use for all probability estimates. Using $p = 0.5$ and rearranging the expression above, we find that $N = 22,500$. If we use this value for all our simulations, we will be guaranteed that all probability estimates satisfy the stated accuracy requirements. A more sophisticated approach would be to monitor convergence on the fly, using our estimate $\hat{p}$ in place of p to calculate the termination criteria. In this latter case, we need to make sure that we first complete a minimum number of samples (say $N = 500$) so the probability estimates are reasonable before they are used in the termination calculation. We would also end up using a different number of samples for the Monte Carlo simulation for each pitch type.

Table 2 gives sample results for the probability estimates for each combination pitch and hit type.

Table 2: Estimated probability of the batted ball results using a Monte Carlo simulation for each pitch type.

Pitch Type	Ground Balls	Line Drives	Fly Balls	Home Runs
High Fastball	0	0.002	0.82	0.18
Sinking Fastball	0.66	0.18	0.12	0.04
Curveball	0.25	0.23	0.51	0.01

3. Plot a histogram of the computed maximum range of the baseball for each pitch type. (You should have one histogram per pitch type.) Also include a histogram of the inputs for the simulation. Do the inputs match what you expect? Discuss how the histograms vary for the different pitches.

   Solution: These plots are shown in Figures 2–7. The inputs are in a triangular distribution as expected with the appropriate ranges. We see that the distributions of the ball's range is quite different among the three different pitch types. High fastball pitches clearly result in higher ranges on average, although there is also large variability. Sinking fastballs lead mostly to the ball being hit a short distance, while curveballs lead mostly to short and mid-distance hits. Some high fastball pitches result in the ball travelling a negative distance (foul balls). This corresponds to situations where the ball leaves the bat at a high angle (i.e., being hit up into the air) and the wind is in the negative x-direction.

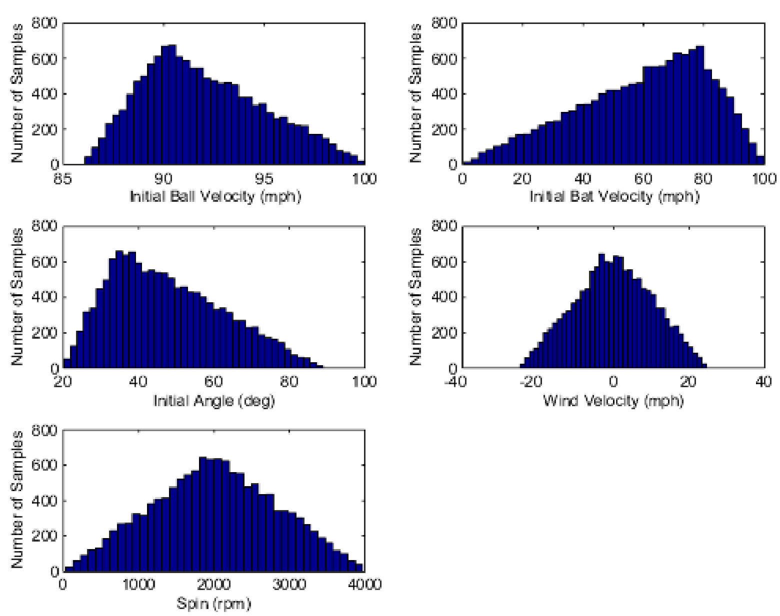

Figure 2: Histogram of the inputs for the high fastball.

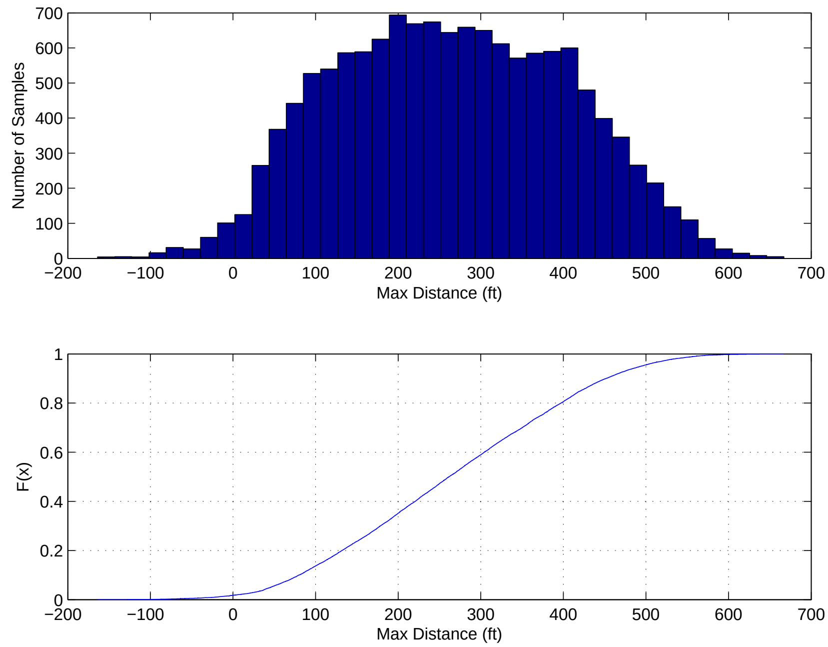

Figure 3: Histogram and CDF of the distance traveled by the ball for the high fastball.

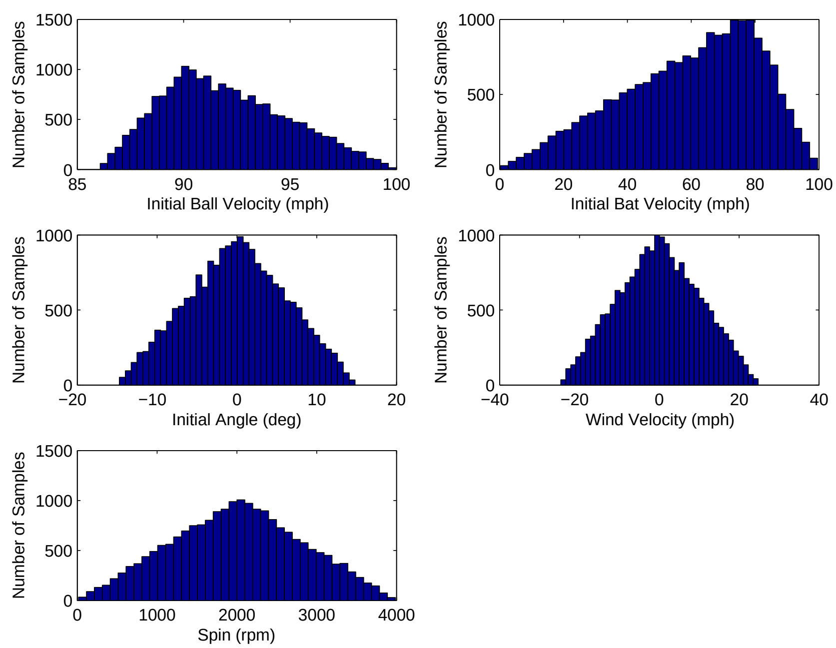

Figure 4: Histogram of the inputs for the sinking fastball.

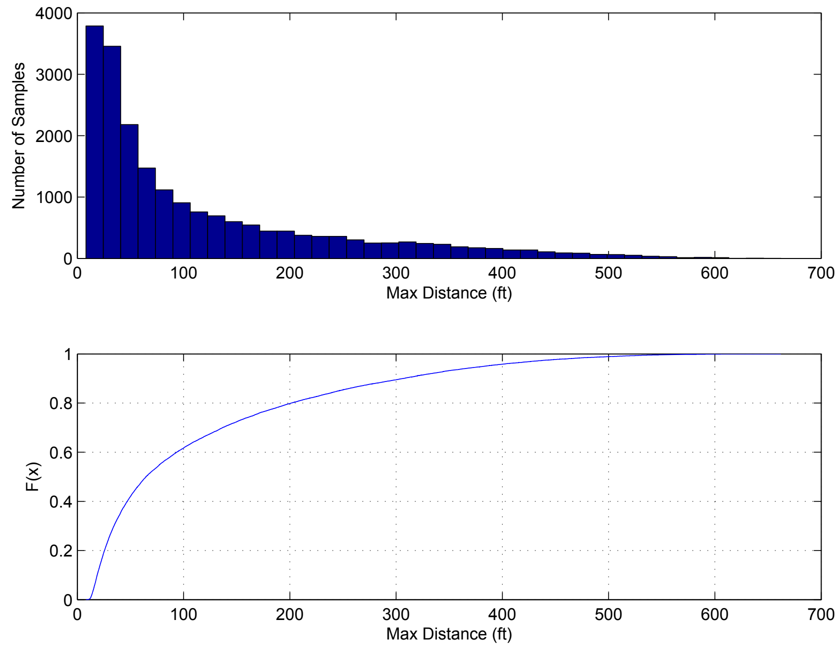

Figure 5: Histogram and CDF of the distance traveled by the ball for the sinking fastball.

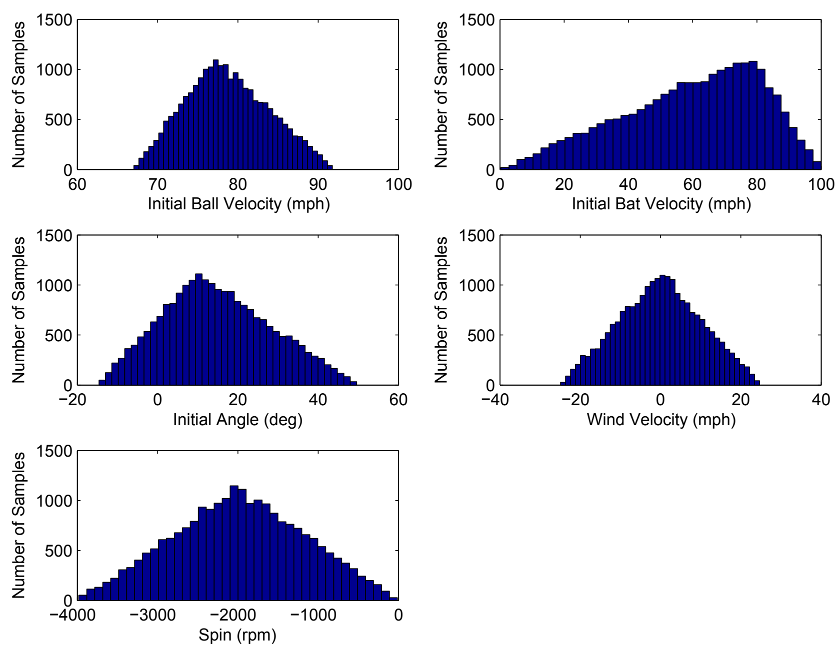

Figure 6: Histogram of the inputs for the curveball.

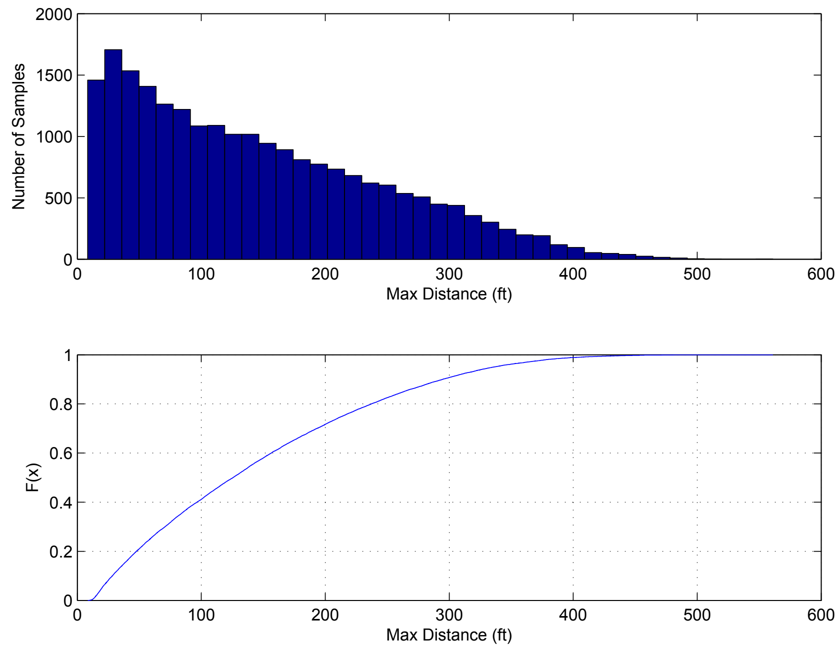

Figure 7: Histogram and CDF of the distance traveled by the ball for the curveball.

4. If you wanted to have a high percentage of a ground ball to try and create a double-play, which pitch would your recommend to your pitcher? If you were the batting coach, which ball would you recommend to the batters in order to hit a home run?

   Solution: If you want to have the highest percentage of a ground ball, the pitcher should throw a sinking fastball according to Table 2. The batting coach should recommend trying to hit a high fastball for the greatest probability of yielding a home run.

2.2 Estimation of the Mean Distance the Baseball Travels

The probability of the batted ball results is useful; however, one of the baseball managers would prefer to have the results in the form of an expected (mean) value of the distance the ball traveled in the x-direction. Using your initial results, determine the mean and variance of the x-distance traveled by the ball for each pitch type. Determine the standard error of the mean estimate.

Solution: The estimated mean distance traveled by the ball for each pitch type is given in Table 3, along with the corresponding estimator standard errors. The table also shows the estimated variance in the distance traveled by the ball for each pitch type. Note that your standard error could be different, depending on how you selected N.

Table 3: Estimated mean distance traveled by the ball, the corresponding estimator standard error, and the estimated variance of the distance traveled by the ball .

Pitch Type	Mean Distance (ft)	Variance (ft^2)	Standard Error (ft)
High Fastball	264.5	19662	1.214
Sinking Fastball	117.0	14384	0.840
Curveball	145.1	10261	0.675

3 References

[1] Alan M. Nathan. The effect of spin on the flight of a baseball. *American Journal of Physics* 76(2): 119–124 (2008).