16.90 Project 3

Spring 2014

For this project we are asked to investigate the effect of three different pitch types on the distribution of hits in a baseball game. In order to conduct this investigation, we are going to use a Monte Carlo simulation. A code that computes the flight of the ball given the bat and pitch parameters was provided. The inputs for the simulation are as listed in table 1.

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

Table 1: Possible inputs for baseball code

Each variable is treated as a random variable with a triangular distribution. We sample the triangular distribution using an inversion method as follows.

1. Choose u from $U[0,1]$
2. Compare to $F(x_{mpp}) = \dfrac{x_{mpp} - x_{min}}{x_{max} - x_{min}}$

   if $u < F(x_{mpp})$ then $x_{min} < x < x_{mpp}$
   $x_u = x_{min} + \sqrt{u(x_{max} - x_{min})(x_{mpp} - x_{min})}$
   else $x_{mpp} < x < x_{max}$
   $x_u = x_{max} - \sqrt{(1-u)(x_{max} - x_{min})(x_{max} - x_{mpp})}$

We use this process of inversion sampling to pick the inputs for the baseball dynamics code.

Thus, our method for the Monte Carlo simulation is as follows:

1. Use inversion sampling to obtain the input vector
2. Compute x, y, t using the baseball dynamics function
3. Classify the hit based on the following algorithm

   if $\max(y) \leq 4$
   hit is a ground ball
   elseif $\max(y) < 10$
   hit is a line drive
   elseif $\max(y) > 400$
   if $\min(y(399 < x \leq 400)) > 8$
   hit is a home run
   else
   hit is a fly ball
   else
   hit is a fly ball

Note, we check for $x_{max} > 400$ and not $x_{max} \geq 400$ since if $x_{max} = 400$, y would not be greater than 8, and the hit would be a fly ball. We repeat steps 1 through 3 for N iterations, each time keeping track of the inputs and the type of hit. It is desired that all of the probabilities have a tolerance of ± 0.01 with a confidence of 99%. To achieve this confidence, we must ensure that the following is true:

$$P\{-3\sigma_{\hat{p}} \leq \hat{p} - PA \leq 3\sigma_{\hat{p}}\} \approx 0.99$$

where $\sigma_{\hat{p}}$ is defined by noting that $P\{A\}$ can be described with a Bernoulli Random Variable and $\hat{p}$ has a normal distribution. This means that

$$\sigma_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{N}$$

At worst, $p(1-p)$ is 0.25 for $p = 0.5$. Thus, if we choose N such that the desired tolerance is met for $p = 0.5$, we are virtually guaranteed we will have the desired tolerance. We compute this N as follows:

$$\begin{aligned} -3\sigma_{\hat{p}} \leq \hat{p} - P\{A\} &\leq 3\sigma_{\hat{p}} \\ \Rightarrow |\hat{p} - p\{A\}| &\leq 3\sigma_{\hat{p}} = 0.01 \\ 3\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} &= 0.01 \\ 9(10,000)(0.5)(1-0.5) &= N \\ \Rightarrow N &= 22500 \end{aligned}$$

Using this N we are fairly confident that our desired tolerance will be met for all possible $\hat{p}$.

Once the simulation has been run, there are several pieces of information that are desired. First, we want to know $\hat{p}$, the probability of each hit type for each pitch. We also want to know the mean range for each pitch type, as well as the variance of the range and the standard error of the mean estimator. For the probability of each hit type $\hat{p}$, we use the definition of the mean to derive the estimator.

$$\hat{p} = \frac{\sum_{i=1}^{N} y_i}{N} \quad y_i = I(A) = \begin{cases} 1 & \text{if hit type condition met} \\ 0 & \text{if hit type condition is not met} \end{cases}$$

Here N is the number of trials in the simulation.

The mean range $\bar{x}$ is similarly defined. Here x is the maximum distance that the ball travels.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

x_i is the range, i.e. the maximum value of the output vector x from the baseball dynamics code. N is again the number of trials in the Monte Carlo simulation.

The variance is derived as follows.

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

This is the sample variance of the data. N is again the number of trials.

The standard error of the mean estimate is defined as follows.

$$\begin{aligned} \sigma_{\bar{x}}^2 &= V(\bar{x} - \mu_x) \\ &= V(\bar{x}) - V(\mu_x) \\ &= V(\frac{\sum_{i=1}^{N} x_i}{N}) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} V(x_i) \\ \sigma_{\bar{x}}^2 &= \frac{1}{N}\sigma_x^2 \end{aligned}$$

Since σ_x^2 is unknown, we use the unbiased estimate that was previously computed.

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})$$

Thus, the standard error of the mean estimate is given by

$$\sigma_{\bar{x}} = \sqrt{\frac{s_x^2}{N}}$$

Results

Running each pitch for $N = 22500$ trials, we obtain the results presented in the following tables.

Pitch Type	$\bar{x}$ [ft]	S_x^2 [ft^2]	$\sigma_{\bar{x}}$ [ft]
Curveball	144.9298	1.0107×10^4	0.6702
Sinking Fastball	115.5313	1.4194×10^4	0.7943
High Fastball	262.510	1.8669×10^4	0.9109

Pitch Type	$P\{$Ground Ball$\}$	$P\{$Line Drive$\}$	$P\{$Flyball$\}$	$P\{$Home Run$\}$
Curveball	0.2472	0.2316	0.5121	0.0092
Sinking Fastball	0.6593	0.184	0.1261	0.0306
High Fastball	0	0.0024	0.8269	0.1706

The results show that if you wanted to hit a ground ball, your best chance is off a sinking fastball which results in about a 65.93% chance of hitting a ground ball. Thus, if a ground ball is desired, the pitches should throw a sinking fastball. Similarly, if a home run is desired, a high fastball is best. This type of pitch results in a 17.06% change of a home run. However, this pitch also has the highest chance at a fly ball, 82.69%. Line drives are most often hit off curve balls, with a probability of 0.2316.

Histograms for all inputs and ranges are plotted below. The results show that the frequency of each range for the three pitch types. The histograms of the inputs show exactly what we would expect. Each input has a triangular distribution centered about the most probably value for that input. Values of the input away form x_{mpp} linearly decrease to the x_{min} and x_{max} values.

The distribution of these inputs also confirms the probabilities of the different hits that were observed during the simulation. For example, the high fastball inputs show that the majority of the values for θ are near 40°. This will send the ball in an upward trajectory, greatly increasing the chances that the hit will be a home run. This can be seen in the

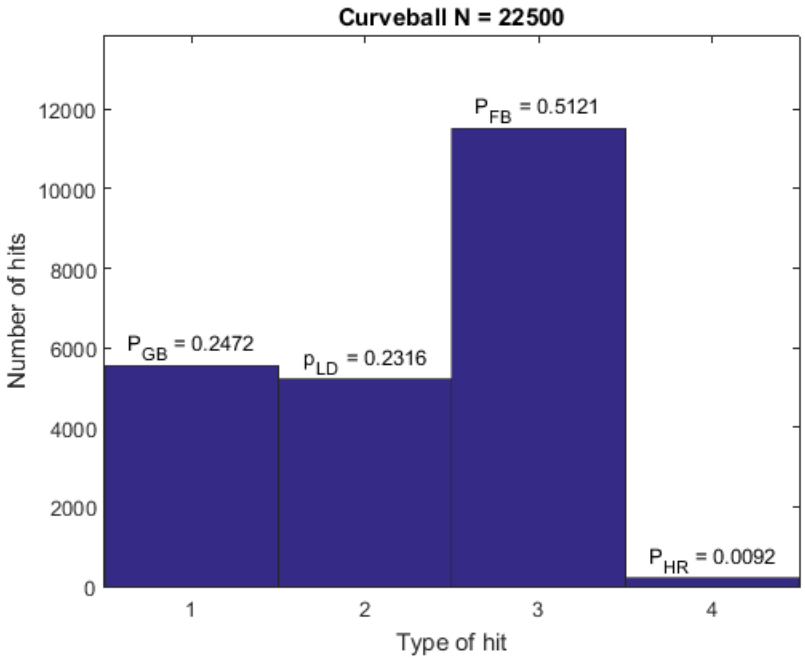
histogram of the frequency of hit types for the high fastball pitch. The hits are heavily concentrated in the home run and fly ball bias.

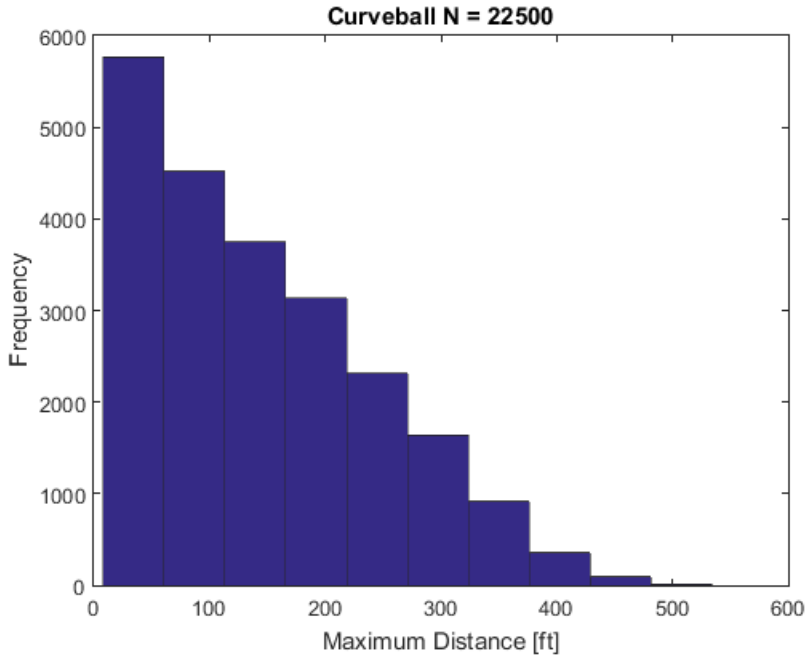
This same correlation between inputs and results can be seen for the sinking fastball. The values of θ are centered on 0, resulting in the ball leaving the bat straight. This increases the number of line drives and ground balls, as can be again seen in the histogram of hit type frequency. The end result is that while the shape of the histograms of the inputs do not change much between pitch types, the center of the distribution varies greatly, resulting in the varied hits for each pitch.

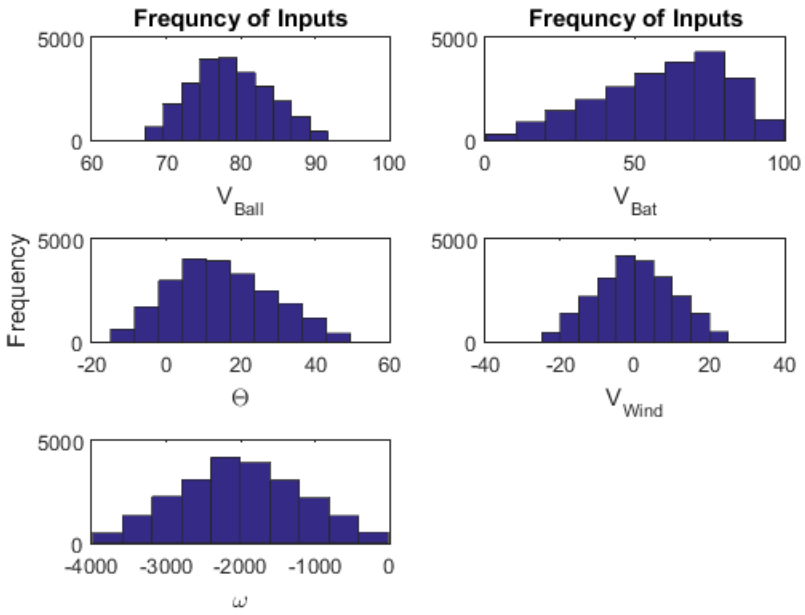
Also included are the plots of the variance of the range and the probability, as well as the confidence interval of $3\sigma_{\hat{p}}$ versus N. These plots show that for all pitch types the desired confidence interval was met. The green line denotes the "conservative" condition, i.e. when $\sigma_{\hat{p}} = 0.25$ corresponding to $\hat{p} = 0.5$. The light blue line denotes a tolerance of ± 0.01. Since by the end of the x-axis at the N used for the simulation all of the tolerance curves end inside the blue lines, we know that our desired tolerance for $\hat{p}$ was met for all of the pitches. This plot also shows that the number of pitches required to achieve the tolerance of ± 0.01 is less than our conservative estimate of N, which is not surprising since there were almost no hits with probability close to 0.5 for any of the three pitch types.

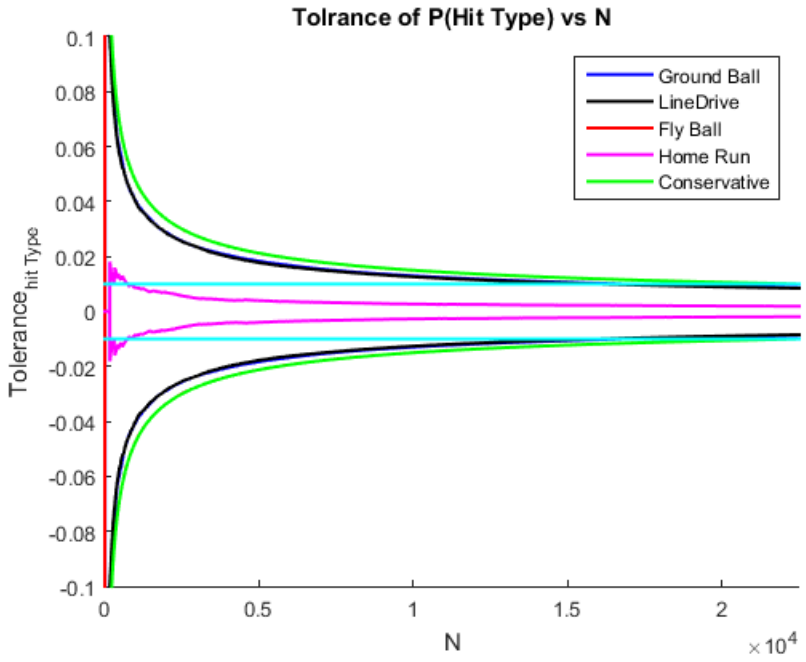
The plots of the variance of the range correspond to the value of s_x^2 as a function of N as N increases. The key information from this chart is that the slope of the variance tends to 0 as N increases. This tells us that our estimation of σ_x^2 with $s_x^2 = \dfrac{1}{N-1}\sum(x_i - \bar{x})$ was a good one. Interestingly, the sinking fastball had the most variability in the variance as N increased. This can be explained by looking at the histogram of the ranges and the probabilities of each hit type for the sinking fastball. This pitch's results have a large number of low range balls with a long tail. This means that the mean is susceptible to new data for low values of N. A larger number of iterations was necessary to achieve stability in the data. However, the variance curve does level off, so there is little need to worry about our estimate for the variance of the range. Similarly, the plots of the variance of $\hat{p}$ for each hit type show that they converge to 0 as N increases. This again shows that our N is large enough to achieve the desired accuracy and confidence in our estimators.

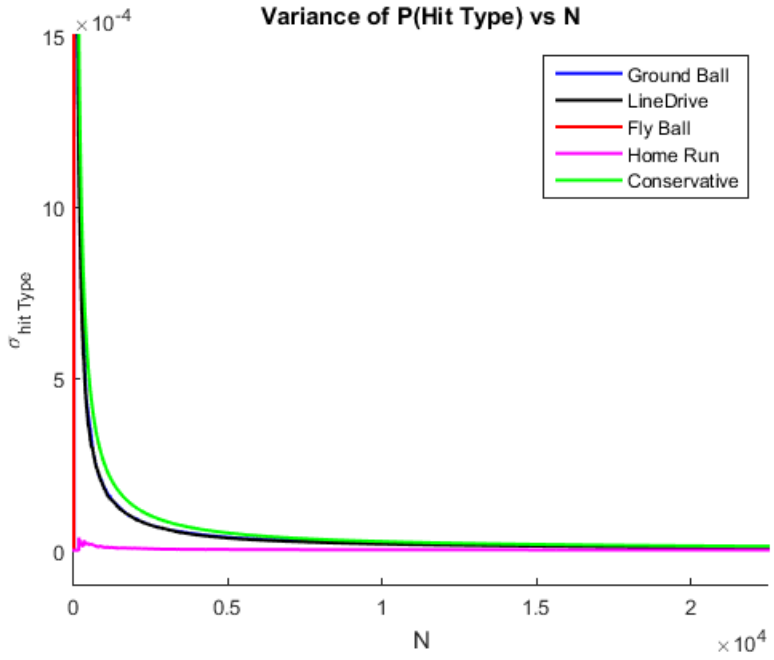
Curveball

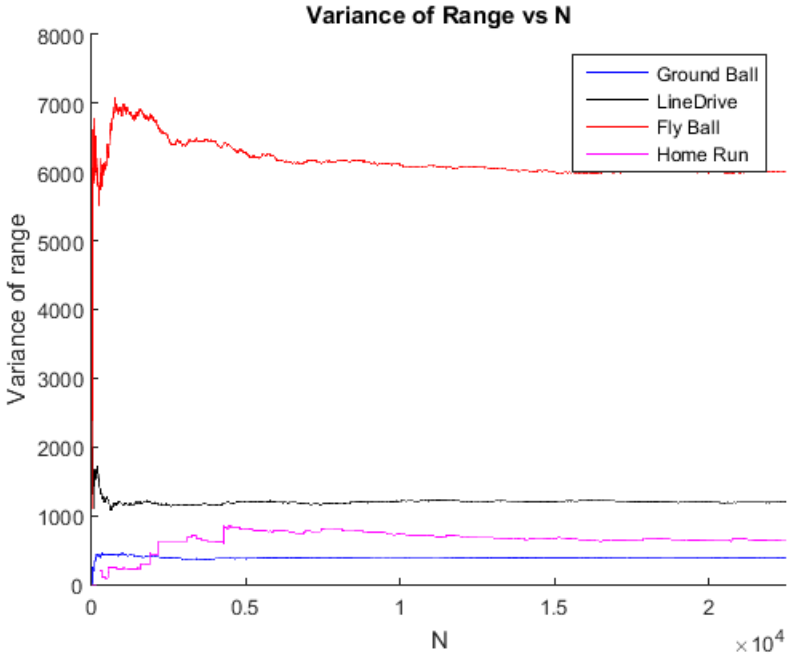

Curveball N = 22500
Number of hits
Type of hit
P_GB = 0.2472
P_LD = 0.2316
P_FB = 0.5121
P_HR = 0.0092
0
2000
4000
6000
8000
10000
12000
1
2
3
4


Curveball N = 22500
Frequency
Maximum Distance [ft]
0
1000
2000
3000
4000
5000
6000
0
100
200
300
400
500
600

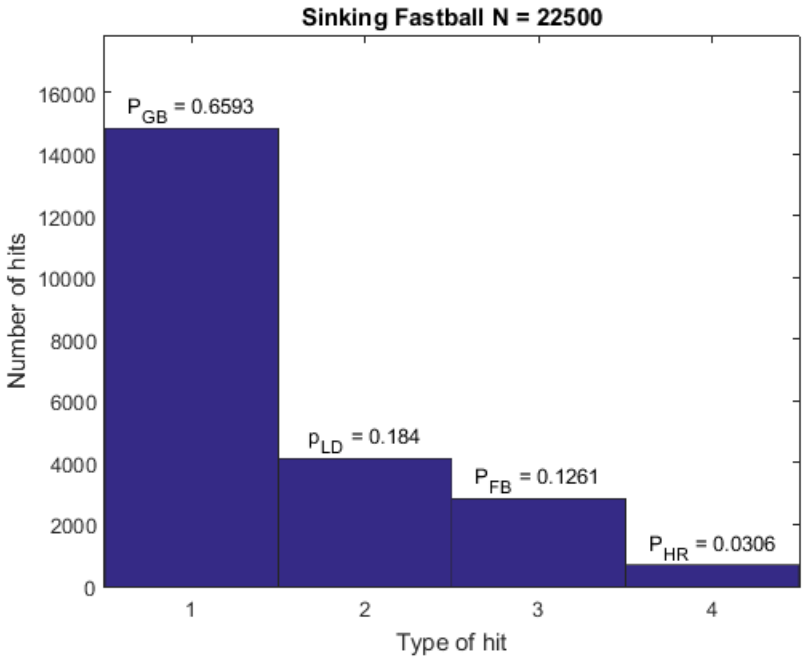

Frequncy of Inputs
Frequncy of Inputs
Frequency
5000
0
60
70
80
90
100
V_{Ball}
0
50
100
V_{Bat}
-20
0
20
40
60
Θ
-40
-20
0
20
40
V_{Wind}
-4000
-3000
-2000
-1000
0
ω

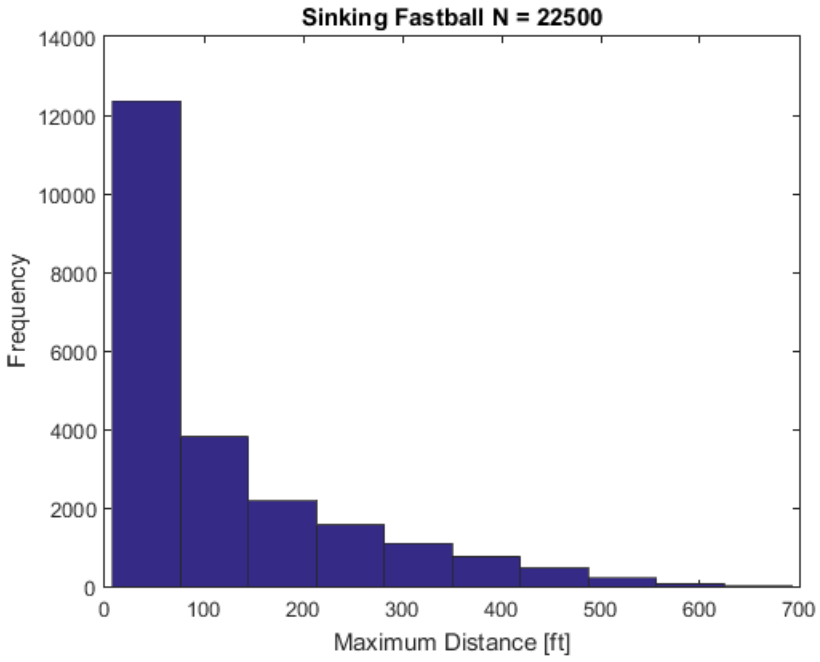

Tolrance of P(Hit Type) vs N
Ground Ball
LineDrive
Fly Ball
Home Run
Conservative
$Tolerance_{hit\ Type}$
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
0
0.5
1
1.5
2
N
$\times 10^4$

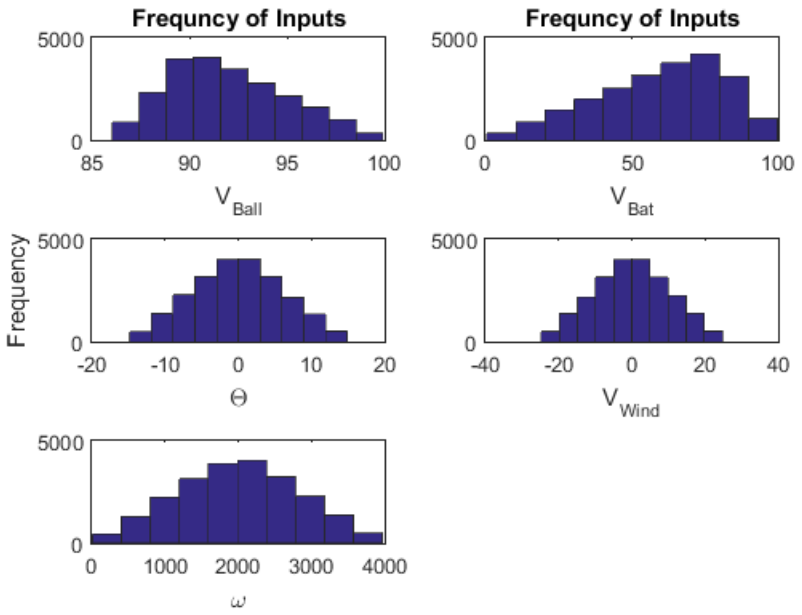

Variance of P(Hit Type) vs N
×10^-4
15
10
5
0
σ hit Type
Ground Ball
LineDrive
Fly Ball
Home Run
Conservative
0
0.5
1
1.5
2
N
×10^4

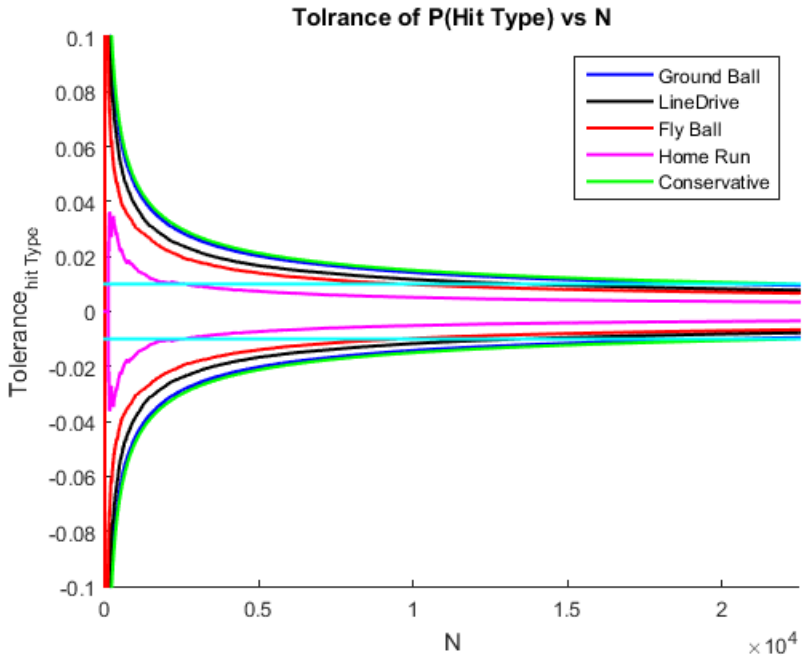

Variance of Range vs N
8000
7000
6000
5000
4000
3000
2000
1000
0
Variance of range
Ground Ball
LineDrive
Fly Ball
Home Run
0
0.5
1
1.5
2
N
×10^4

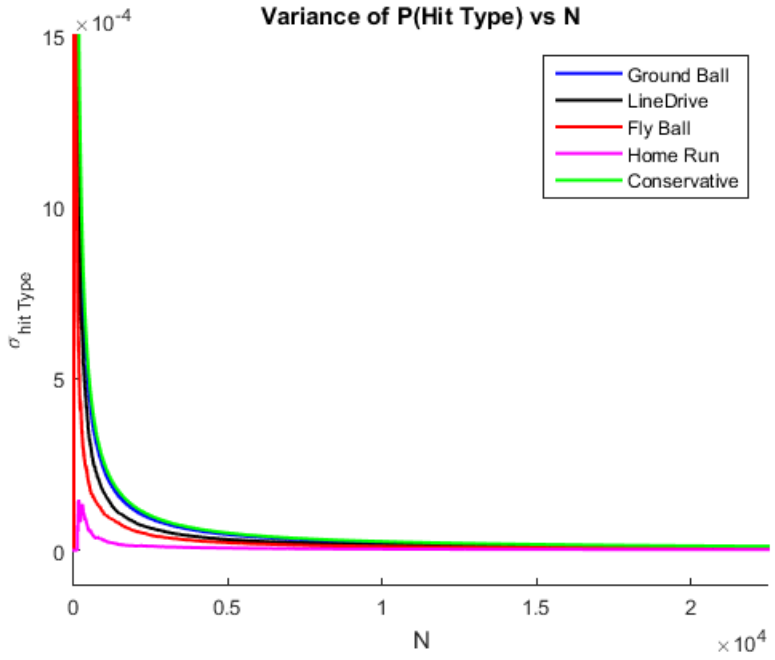
Sinking Fastball

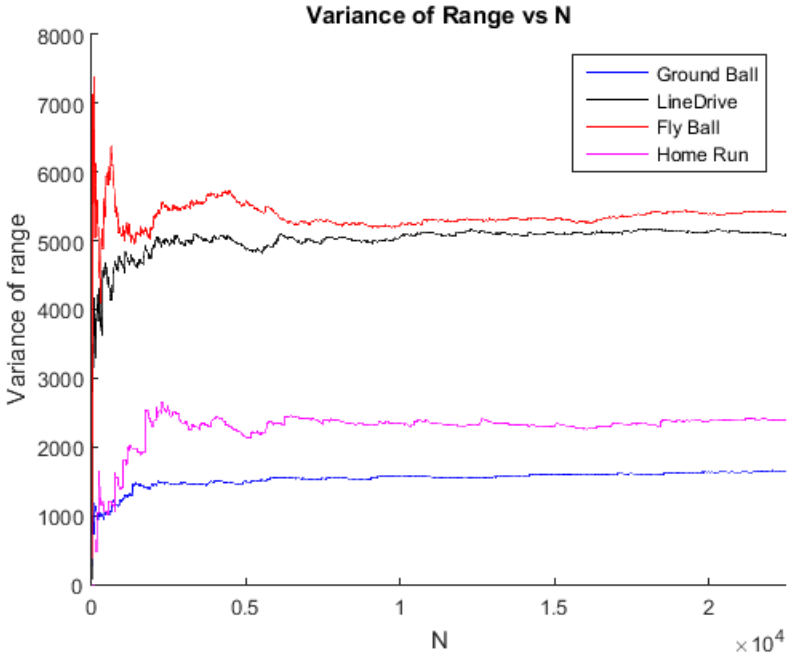




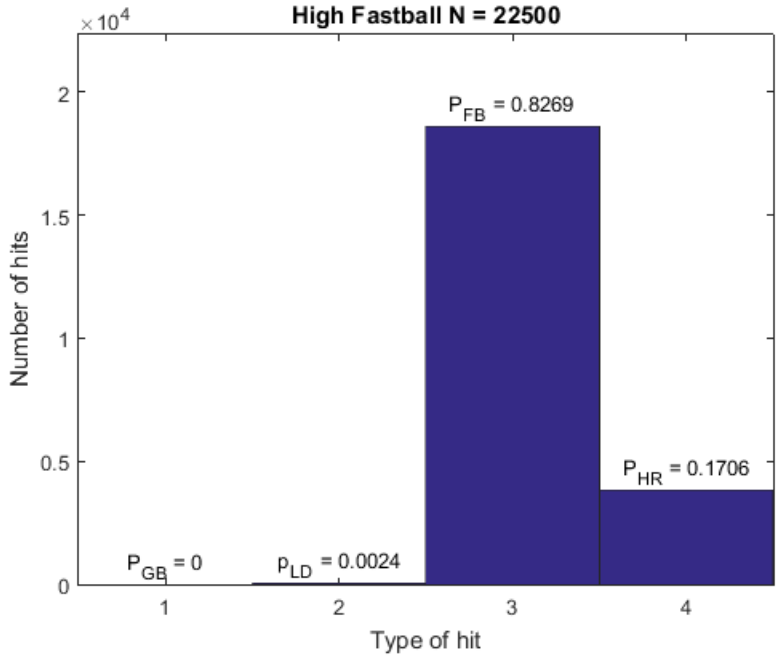

Frequncy of Inputs
Frequncy of Inputs
Frequency
5000
0
85
90
95
100
V_{Ball}
0
50
100
V_{Bat}
-20
-10
0
10
20
Θ
-40
-20
0
20
40
V_{Wind}
1000
2000
3000
4000
ω

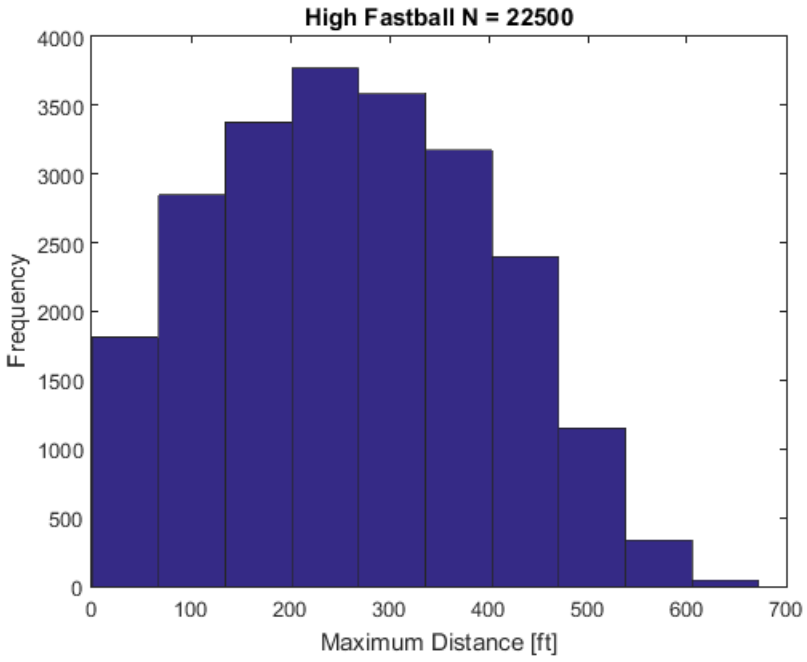

Tolrance of P(Hit Type) vs N
Ground Ball
LineDrive
Fly Ball
Home Run
Conservative
$Tolerance_{hit\ Type}$
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
0.5
1
1.5
2
N
$\times 10^4$

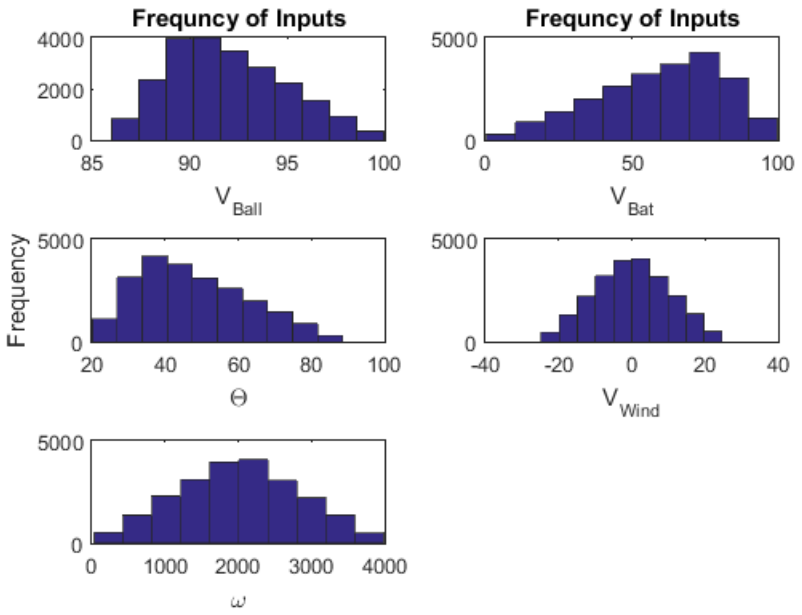

Variance of P(Hit Type) vs N
×10^-4
15
10
5
0
σ hit Type
Ground Ball
LineDrive
Fly Ball
Home Run
Conservative
0
0.5
1
1.5
2
N
×10^4

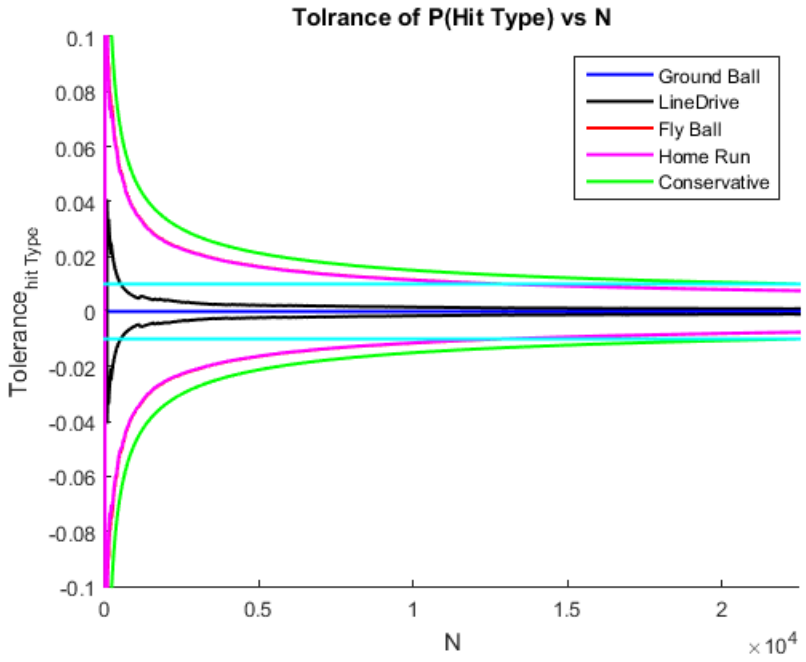

Variance of Range vs N
8000
7000
6000
5000
4000
3000
2000
1000
0
Variance of range
Ground Ball
LineDrive
Fly Ball
Home Run
0
0.5
1
1.5
2
N
×10^4

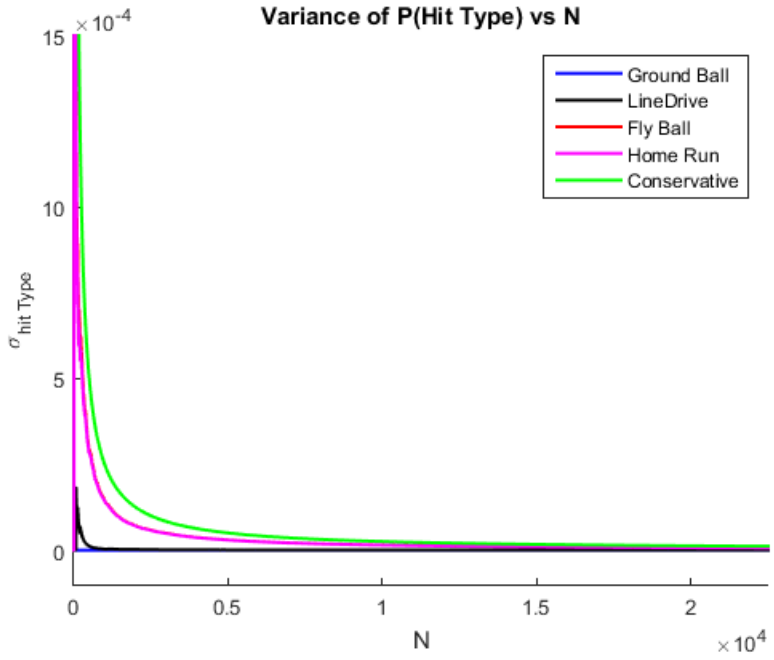
High Fastball

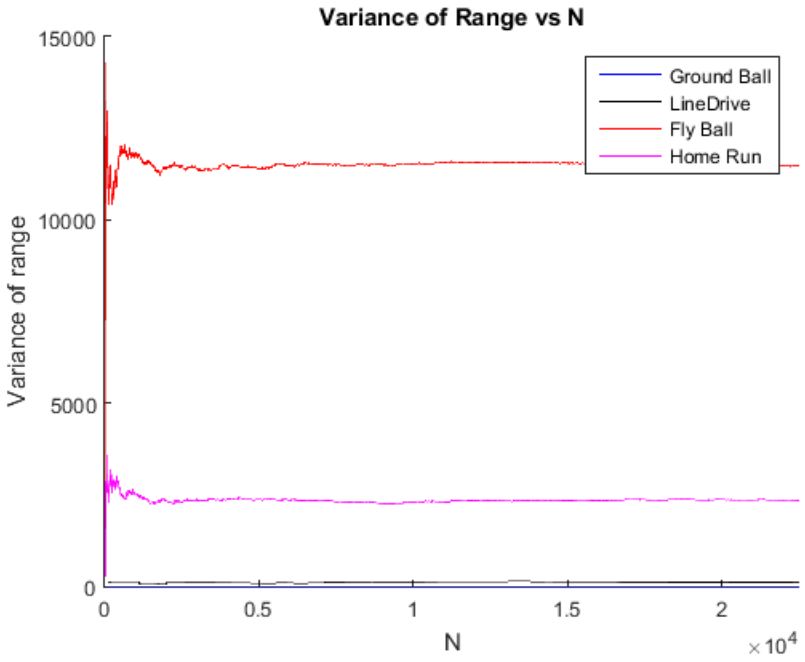





Frequncy of Inputs
Frequncy of Inputs
Frequency
V Ball
V Bat
Θ
V Wind
ω


Tolrance of P(Hit Type) vs N
Ground Ball
LineDrive
Fly Ball
Home Run
Conservative
Tolerance hit Type
N
×10^4


Variance of P(Hit Type) vs N
×10^-4
15
10
5
0
σ hit Type
0
0.5
1
1.5
2
N
×10^4
Ground Ball
LineDrive
Fly Ball
Home Run
Conservative


Variance of Range vs N
15000
10000
5000
0
Variance of range
0
0.5
1
1.5
2
N
×10^4
Ground Ball
LineDrive
Fly Ball
Home Run

16.90 Computational Methods in Aerospace Engineering

Spring 2014